Mental health without mirrors

A non-descriptivist approach to mental health and the intervention with people with delusions

Tesis Doctoral presentada por

Miguel Núñez de Prado Gordillo

Para optar al Grado de Doctor con Mención Internacional en el Programa de Doctorado en Psicología Clínica y de la Salud

Directores

María Xesús Froxán Parga

Facultad de Psicología, Universidad Autónoma de Madrid

Manuel de Pinedo García

Facultad de Filosofía, Universidad de Granada

Universidad Autónoma de Madrid

Facultad de Psicología

Departamento de Psicología Biológica y de la Salud

Madrid, 2022

A Amalia.

Y a mis hermanos, Jorge y Juan.

Contents

Agradecimientos / Acknowledgements

Aquí quiero abogar, como hace la gente que admiro, por un individualismo de las miserias y un colectivismo de los éxitos: los vicios de esta tesis son únicamente míos; sus virtudes, colectivas. En mi caso, de hecho, especialmente colectivas -en parte por eso está escrita en primera persona del plural. La posibilidad misma de llevarla a cabo, el origen de los intereses que en ella se reflejan, sus contenidos, las formas de vida y las maneras de vivir que constituyen su contexto; todo ello se lo debo a un sinnúmero de personas. Personas de muy diversos grupos, que me han regalado su cariño y su tiempo en distintos momentos de este proceso. A todas vosotras, tanto las aquí incluidas como las que he tenido que dejar fuera, gracias de corazón.

Por no dejar un hueco de la tesis sin hacer referencia a la de Manu Almagro, he estructurado los agradecimientos como lo hace él, siguiendo un orden semi cronológico que refleja mejor el hilo de contingencias que han desembocado en este trabajo.

Quiero empezar por tanto dándole las gracias a mi familia. En primer lugar, porque me han dado las condiciones materiales para poder siquiera pensar en dedicarme a esto y quienes me han brindado apoyo moral y económico siempre que lo he necesitado. Pero fundamentalmente porque, aunque quizás no lo sepan -ni yo lo sabía realmente hasta hace poco- son mi núcleo. Hay familias que son jardines barrocos, todo líneas rectas, simetría y orden. Y hay familias, como la mía, que son jardines ingleses; caóticas, exuberantes y, con todo, mucho más acogedoras. Mi madre, Maite, encarna para mí la justicia, la entereza y el amor por los demás. Ella me ha enseñado la importancia de los valores y los cuidados, y que la mejor manera de darlos pasa muchas veces por cuidarse uno mismo. También le debo el perfeccionismo, la constancia y la forma intensa y obsesiva de disfrutar con lo que nos gusta. Mi padre, Miguel, es para mí un modelo de flexibilidad, tolerancia y creatividad. Él despertó en mí el interés por lo conceptual, el sentido de lo místico y la apertura a la experiencia y a otras maneras de vivir. Su multiperspectivismo y su forma de resistir y atravesar las

costumbres que se le han impuesto son lo que más admiro de él. Mis hermanos Jorge y Juan son, además de los reyes de mi casa -en eterna lucha por el trono del sofá- el núcleo de mi núcleo. Conozco pocas personas tan bonitas como ellos. No puedo en tan poco espacio hacer justicia a su bondad, ingenio y valentía. En ellos me veo como en nadie; al mismo tiempo, representan todo lo que aspiro a ser. Os quiero mucho. Gracias por estar siempre ahí, incluso cuando no estáis.

Mis amigos y amigas de la sierra, Alvarito, Bobby, Borzyk, Campi, Castañs, Charly, Clara, David Man, Ele, Eskizo, Joako, Libin, Mantecón, Mery, Migui, Milos, Monti, Puxy, Quesada, Rocha, Víctor, Werty y tantas más que me dejo fuera, han sido mi segunda familia durante mucho tiempo. Les debo media vida de carcajadas, anécdotas, desmemorias, intoxicaciones, locuras varias y mucho amor y comprensión. Siendo personas tan variopintas, todas ellas saben vivir de una forma que no he encontrado en ningún otro sitio; si tuviese que definirla de alguna manera, sería por oposición al tipo de actitudes que a veces encuentra uno en la academia (narices arrugadas, ademanes afectados y miradas por encima del hombro). Por el contrario, la suya es el tipo de actitud que quiero llevar siempre conmigo. Alvarito, Joako, Mantecón, Migui, Víctor, gracias por todas las noches dulces y agridulces, y por haber sido mi casa cuando más falta me hacía. Bobby, gracias por tu cariño, tu humor ácido y tu núcleo dulce; pensar en ellos me ha dado fuerzas cuando me han faltado estos últimos meses. Borzyk, Monti, Puxy, gracias por ayudarme a revisarme siempre desde el cariño. Campi, Castañs, Clara, David, Mery, Ques, gracias por la música, los juegos, los planes locos, las discusiones. Elena, gracias por haberme acompañado parte de este camino. Libin, gracias a ti y a Paula por los buenos ratos en vuestro pequeño paraíso albaicinero. Milos, gracias por tu perspectiva tragicómica y por saber verme por dentro. A todas y todos: gracias por haber sido mi casa, mi lecho del río y mi norte, y gracias por haberme regalado vuestro tiempo. No sería quien soy sin todas vosotras y vosotros.

Quiero referirme especialmente a mis dos amigos de toda la vida, Jorge Monterde y Charly Trujillano. Jorge es, entre muchas otras cosas, el origen de mi interés por la filosofía y la psicología; el marco teórico de esta tesis es en parte una respuesta al tipo de inquietudes que teníamos en la adolescencia. Su habilidad para detectar, explotar y habitar las contradicciones y, sobre todo, su profunda humanidad, son mi referencia. Charly es mi lumbre. Ha estado conmigo siempre, en la cercanía y en la distancia, en el error y en el acierto, en la tormenta errática de las noches y en la calma chicha de los atardeceres. Su sentido del humor, su transparencia y su calidez son mi abrigo. Es un auténtico regalo haber podido disfrutar de vuestro amor y amistad durante tantos años.

También quiero dar las gracias a todas las personas con quienes compartí piso en Madrid durante estos últimos años; ya solo haber aguantado mi entropía merece un reconocimiento. Monti, gracias por haberte lanzado conmigo a la aventura de encontrar un piso habitable y por haber evitado que acabásemos en un zuloft de esos tan bien iluminados. Ana, gracias por intentar enseñarme catalán y por todas las noches de fiesta combativa. Especial mención a Carles Bixquert, Javi de la Rocha y Jaime de la Torre (Werty), el núcleo duro de Cañaveral 13. Vosotros habéis sido mi casa los últimos años -la base, al menos. Si realmente se gana tiempo de vida riendo, con vosotros quedan compensados de sobra todos los excesos. Sois tres personas increíbles y no puedo más que agradecer todo lo que me habéis dado: las rimas, los bucles, las caseras espías; el humo, los licores, las tardes en la Remonta; las carcajadas, las lágrimas, los abrazos; por todo ello, y por todo el apoyo y cariño que me habéis brindado estos años, mil gracias. Hicisteis que volver a casa fuese siempre un momento dulce.

Mi experiencia en el grado de psicología no hubiese sido tan especial como lo fue sin personas como Adri, Alonso, Carol, Cris, Hoyas, Javi, Javichuelas, Jordana, Luis, María, Nerea, Paloma y Sara. Guardo con mucho cariño todas las risas, bailes, disfraces, música y conversaciones con las que llenaron mi vida universitaria. Con ellas también aprendí de psicología gran parte de lo que no supe sacar del grado y se empezaron a perfilar mis intereses en la filosofía de la psicología; también fueron ellas quienes educaron primero mi perspectiva política y me enseñaron a ver mis privilegios y contradicciones. Gracias por vuestra paciencia, vuestro cariño y por todo lo que me habéis enseñado. Tengo que hacer mención especial a Palo y a Chuelas. Palo, nuestra amistad y nuestras conversaciones han sido un refugio y una fuente de aprendizaje constante para mí, y tu apoyo ha sido vital para transitar la vida académica y superar la etapa predoctoral. La academia es un espacio más seguro y más bonito gracias a personas como tú. Chuelas me ha regalado su amor y comprensión desde los primeros días de la carrera hasta hoy. Su amistad es de las cosas que más aprecio; un espacio atemporal y precioso hecho de arena del Atlántico y de canciones de Ben Howards. Mil gracias por tu afecto y tus consejos en esta última etapa de la tesis, Javi; sin ellos estaría todavía escribiendo.

Quiero dar las gracias también a todas las personas que conocí estando de Erasmus: Anastasia, Andrés, Bibi, Eze, Manu, Manfre, Sandra, Thann, Valentina y tantas otras personas que me dejo fuera. Mil gracias por todos los buenos momentos que me hicisteis pasar. Eze y Manfre, cuyos cuidados y cariño han estado conmigo todos estos años, merecen una mención especial. Gracias, entre otras muchas cosas, por vuestra música, que puso banda sonora a algunos de mis recuerdos más preciados.

El máster en lógica y filosofía de la ciencia fue para mí una experiencia transformadora, en todos los sentidos. Luego me referiré a algunas de las personas que contribuyeron a que así fuera; por ahora, quiero dar gracias a todo el profesorado y el alumnado del curso 2015/2016, por haberme enseñado (y aguantado) tanto. Llegué con varios trastornos -el cientifismo y el eliminativismo, entre otros; todos vosotros y vosotras contribuisteis a la terapia. Mención especial a mis compañeros y compañeras de promoción, y en particular a Alba, Dani, Llanos, Manu, Ricardo y Sara. Os agradezco de corazón las clases extra que recibí entre la 1 y las 5 de la mañana entre penúltima y penúltima; sin ellas, no hubiese podido sacar el máster adelante.

Durante el doctorado he tenido la suerte de disfrutar de un contrato predoctoral a cargo de la Universidad Autónoma de Madrid (FPI-UAM 2017), asociado al proyecto de investigación "Estudio funcional de la interacción clínica en pacientes con diagnóstico de enfermedad mental" (PSI2016-76551-R). Ello me ha permitido dedicarme al trabajo investigador a tiempo completo y no tener que depender de otras fuentes de ingresos, y por ello le estaré siempre agradecido a la Autónoma. En este sentido, también quiero agradecer a los dos departamentos a los que he estado adscrito: el Departamento de Psicología Social y Metodología y el Departamento de Psicología Biológica y de la Salud. Mención especial a Hilda Gambara, quien se ofreció a ser mi directora de tesis al principio de la misma y avaló mi solicitud. Sin ella, no hubiese podido dedicarme a esto y por ello le estoy enormemente agradecido. Quiero también dar las gracias a Andrés Mejía, con quien tuve la suerte de compartir las clases, y a Nacho Montero, el principal responsable de que decidiese estudiar psicología en la Autónoma en primer lugar, y cuya guía y ayuda han sido fundamentales en varios momentos de este proceso. También quiero dar las gracias a la gente del aula PDIF, especialmente a Palo y José Ángel, por haberme dado un espacio de trabajo tan cálido durante los primeros años del doctorado. Por último, quiero dar las gracias a todo el personal de administración por su ayuda en distintos momentos de este proceso, su dedicación y su paciencia con mis numerosos despistes.

Una de las cosas que me permitió hacer el contrato predoctoral es dar clases en la universidad, que ha sido una de las experiencias más gratas durante este periodo. Al alumnado, os pido disculpas si a veces mis pizarras parecían un Pollock. También siento haberme quedado con varios de vuestros bolígrafos y haber perdido el sentido del tiempo en más de una ocasión, olvidándome de los descansos. Os agradezco en el alma el interés por las clases y que os tomarais el tiempo para rellenar las encuestas; me disteis algunos de los momentos más bonitos durante el doctorado y me ayudaron a reconciliarme con la carrera académica. Quiero agradecer especialmente a Cris Rodríguez su interés y dedicación en las clases. Me

llena de alegría que hayas decidido perseguir la carrera investigadora; hace falta mucha más gente como tú en la academia.

También quiero agradecer a toda la gente de la Asamblea Dignidad Predoctoral de la UAM, en la que participé activamente durante los primeros años del doctorado. Gracias a su entusiasmo y al de organizaciones como la Federación de Jóvenes Investigadores se está consiguiendo dignificar la carrera académica y dotar a los investigadores e investigadoras jóvenes de una trayectoria más estable y menos alienante. Os agradezco en el alma vuestro esfuerzo y espero poder contribuir más al mismo de ahora en adelante.

Entre marzo y julio de 2019 pude disfrutar de una estancia doctoral en la Facultad de Filosofía de la Universidad de Granada bajo la supervisión de Manuel de Pinedo García. Sobre Manolo y el grupo de investigación de acogida hablaré más adelante; simplemente decir que la perspectiva antidescriptivista defendida en esta tesis se la debo a dicho grupo y a los maravillosos meses que pasé allí formándome e investigando. Estoy enormemente agradecido a todos aquellos y aquellas que la hicieron posible.

También quiero dar las gracias a mis compañeras de piso durante esos meses, Celia, Marta y Sonia -y Mar, que realmente también vivía con nosotras. Fuisteis un hogar para mí desde el minuto uno de ir a ver el piso. Guardo como un tesoro las conversaciones en el balcón y la azotea, vuestros intentos de alimentar a esta alimaña, las guerrillas anticucarachiles y las sesiones de terapia en torno a la mesa del salón. Me hicisteis muy feliz.

Between September and December 2020, I had the opportunity to do a three-month research visit at the Institut Jean Nicod (CNRS-ENS-EHESS) in Paris under the supervision of Elisabeth Pacherie, thanks to an Erasmus+ KA103 Higher Education grant and a complementary grant by Santander Bank. I am deeply indebted to Prof. Pacherie for accepting to be my supervisor in the first place, amidst a worldwide pandemic, and for reading and discussing the initial drafts of the second part of this dissertation with me. I'm also grateful to her and the Agency research team for inviting me to attend their seminar and discuss my proposal in it. I'm not sure if Prof. Pacherie knows how much of a difference her argumentative style made for me, but I now take her as a role model for how one should discuss other people's ideas, especially those whom one disagrees with. I would also like to thank Nathalie Evin for all her help with the various administrative procedures. Amir, thanks for letting me your room in Bourg-la-Reine; among other things, because it allowed me to meet Guillaume, the weirdest and funniest landlord I've had so far. Axel and Romain, thank you very much for your warmth and all those insightful coffee breaks. I wish the circumstances had been different, and that I could have had more time to get to know you all.

Durante mi estancia en París tuve también la oportunidad de volver a encontrarme con Nuño Amador. No nos veíamos (que recordáramos al menos) desde hacía unos diez años; en los primeros cinco minutos de sentarnos juntos consiguió que me sintiese como en casa. Nuño es una de las personas más bonitas y cálidas que conozco, y me alegro en el alma de habernos reencontrado. A él y al resto de la tropa parisina -especialmente a Abel, Álvaro, Ana, Ander, Helena, Víctor y Youcef- les estoy increíblemente agradecido por el tiempo que me hicieron pasar en París. Tengo unas ganas locas de volver a veros.

Entre julio y septiembre de 2021 tuve la oportunidad de realizar una estancia (online) en la Escuela de Psicología de la Universidad de Valparaíso, bajo la supervisión de Pablo Andrés López Silva. La estancia estaba en principio pensada para desarrollarse de forma presencial, pero circunstancias relacionadas con la pandemia lo impidieron. Sin embargo, la atención, interés y ayuda constantes de Pablo, su comprensión y sus consejos hicieron de esta una de las estancias más productivas y agradables que he tenido. Le estoy enormemente agradecido por todas las reuniones y discusiones que tuvimos, por haberme permitido trabajar con él mano a mano, por haberme invitado a dar una charla en su seminario y por haber estado pendiente todos estos últimos meses de tesis, cuando hacía ya tiempo que había acabado la estancia. Ojalá tener la oportunidad de visitar Valparaíso y agradecerte todo tu trabajo en persona.

No puedo dejar de mencionar a Laura, mi psicóloga, que estuvo conmigo cuando más lo necesitaba. Laura me ayudó a poner mi casa en orden, a retejer un hilo narrativo propio, a poner a raya otras voces. Me dio estabilidad en mis seísmos y supo encauzar mis maremotos. Me dio herramientas con las que trabajar y trabajarme, y me ayudó a ver y a resolver muchas de mis contradicciones. Me enseñó a respetarme más y a respetar más a los demás. Y lo hizo todo sin pedir nunca nada a cambio. Gracias de corazón, Laura.

Sin duda, a quien más tengo que agradecer que esta tesis haya llegado a buen puerto es a mis dos directores, María Xesús Froxán Parga y Manuel de Pinedo García, así como a las personas que forman parte de sus dos grupos de investigación: Acoveo, en la Universidad Autónoma de Madrid, y Filosofía y Análisis (aka Granada Gang), en la Universidad de Granada.

A María Xesús la conocí al final del grado en psicología. Ella me devolvió el interés por la psicología, avivó mis inquietudes filosóficas y me alentó a perseguir la carrera investigadora. Sabiendo que mis intereses estaban más conectados con la filosofía de la psicología, me animó a formarme e investigar en esa línea; fue por recomendación suya que decidí hacer el máster de lógica y filosofía de la ciencia. De ahí volví con un cambio drástico de perspectiva, diciendo cosas con las que creo que no mucha gente dedicada al análisis de la conducta

estaría de acuerdo; María Xesús no solo no puso freno a mis inquietudes, sino que les hizo un hueco en su línea de investigación, me exhortó a continuar investigando, aceptó dirigirme la tesis y me apoyó en mis muchos intentos de conseguir un contrato predoctoral. Gracias a ella me he podido dedicar todos estos años a un trabajo que me apasiona, en el que he podido prosperar y mejorar continuamente gracias a su supervisión y su moldeamiento y modelado constantes. Ha sido un privilegio tener de directora a una referente de la psicología clínica en España, un titán capaz de conciliar el llevar un grupo de investigación cada vez más productivo y numeroso, hacer una labor docente intachable, dirigir un centro clínico, transformar vidas con su terapia y todavía encontrar tiempo para correr maratones. Eres para mí un modelo de fuerza, dedicación y generosidad, María Xesús, y te agradezco de corazón todo lo que me has dado.

También tengo que agradecerle el haberme hecho un hueco en su grupo de investigación, el grupo Acoveo. En él he conocido a personas increíbles que, además de haberme ayudado y apoyado en incontables ocasiones, han enriquecido mi vida a más no poder. En Acoveo decimos que las tesis son tesis de grupo; esta no es una excepción. Carol, Caru, Concha, Cris, Dani, Elena, Gladis, Inés, María, Nat, Ris, Tommy y Víctor, gracias por vuestro aliento y vuestro trabajo. Esta tesis es también vuestra. A Natalia Andrés le agradezco todo el apoyo que me dio en algunos de los momentos más complicados de este proceso, así como su reforzamiento constante de mi variabilidad en el vestir. A Gladis Pereira le agradezco especialmente sus discusiones conmigo sobre la compatibilidad entre antidescriptivismo y conductismo; también le agradezco las noches de baile flamenco, sus palabras de ánimo y su cariño. Caru Basakatua, Jesús Alonso y Ricardo de Pascual y han sido mis tres luceros estos años. Caru, gracias por enseñarme tanto, por todo el tiempo que has invertido entrenando mis sensibilidades, por mostrarme una manera diversa de habitar la diversidad -la desviación atípica. Más allá de eso, gracias por tu amistad, por cuidarme, por preocuparte por mí de manera tan genuina. Casaros a ti y a Angela es una de las cosas más bizarras y bonitas que he hecho (por cierto, mil gracias por esas deliciosas bolas de queso). Ris, gracias también por educar mis disposiciones, por ayudarme a reconocer mi posición social y por darme así herramientas para ser mejor persona. Tus consejos y tu capacidad para responsabilizar sin culpar han sido clave en algunos de los momentos más duros de este proceso; tus florituras verbales y tu ácida dulzura han hecho mejores los mejores momentos. Jesús, has sido mi bastón y una fuente constante de aprendizaje estos cuatro años. Me has dado un espacio en el que pensar y repensar mis ideas, has puesto a prueba una y otra vez mis posiciones y me has permitido mejorarlas. Pero por encima de todo has sido un verdadero amigo. Tu trato franco y cálido hace que contigo me sienta en familia. Muchas gracias a les tres por vuestro

compañerismo y vuestra amistad. Y gracias también a Angela, Clara, Dani y Guille por apoyarnos y acompañarnos tantas noches.

Tal y como dije al inicio, esta es una tesis especialmente colectiva. En concreto, parte de su contenido se debe no solo a mi grupo de origen, sino también a mi grupo de adopción: el Granada Gang. A este grupo no solo le debo toda mi formación en filosofía, sino también una comunidad abierta, inclusiva y comprometida que ha transformado mi manera de vivir y de pensar en mi trabajo. En primer lugar, quiero darle las gracias a Manolo Pinedo, mi codirector. Ya en el máster aceptó ser mi tutor del TFM, a sabiendas del esfuerzo extra que iba a suponer lidiar con mi falta de base y reformar y reentrenar mi cabezonería cientifista. Su extraordinaria habilidad para la docencia consiguió abrirme las puertas de la percepción a lo normativo y me trajo de vuelta el espacio lógico de las razones. En el doctorado aceptó ser mi supervisor de la estancia y más delante de la tesis doctoral; su labor de supervisión desde entonces, en una época tan complicada, ha sido intachable y le estoy infinitamente agradecido por ello. Por encima de todo, admiro de él sus inquietudes morales y su filosofía -si son separables-, y su forma de enactuarlas: su sensibilidad a las transformaciones en el espacio social y político y su profundo humanismo hacen de su filosofía una filosofía viva, habitable, de carne y hueso; una filosofía capaz de derribar cualquier tribuna y al mismo tiempo reforzar la autoridad de las voces que menos gozan de ella. En una filosofía así uno quiere quedarse a vivir. Por eso y por tantas otras cosas, muchísimas gracias, Manolo.

Quiero dar también las gracias a los profesores y profesoras Esther Romero, Juan José Acero, María José Frápolli y Neftalí Villanueva. Aunque apenas hemos discutido juntos directamente sobre los contenidos de mi tesis, sus posiciones han influido considerablemente la misma, sea por sus artículos, sus comentarios en los seminarios o sea por boca de las personas a las que han entrenado y que me han entrenado a mí. A Neftalí quiero agradecerle también su suspicacia inicial respecto a mi orientación antipsiquiátrica, que ha contribuido enormemente a hacer de esta una disertación más respetuosa y sensible a las posibles razones detrás de posiciones que hasta entonces veía como irreconciliables con la mía. Mil gracias también a los compañeros y compañeras del grupo: Alba, Amalia, Andrés, Dani, David, Edu, Enrique, Francesco, José, Liñán, Llanos, Manolo Heras, Manu, Mirco, Nemesio, Palma, Pedro, Tori, Víctor y Xavi, así como a Alba Fuentes, Ana, Cris, Lorena, Mar, María, María José y Sara; entre todos y todas hacéis del Gang un sitio en el que cualquiera soñaría poder trabajar. Manolo, gracias por tu apoyo y tu guía todos estos años. Desde la ayuda incondicional que me brindaste en el máster, sin apenas conocerme, hasta el día de hoy, no ha habido vez que te vea y no me hayas arrancado un par de carcajadas, regalado una sesión breve de orientación laboral y recordado por qué te admiro tanto. Gracias también por tu trabajo

incansable en la FJI y por abrirnos paso a quienes venimos detrás. Víctor, un millón de gracias por todo el apoyo que me has brindado estos meses, por contar conmigo para tus proyectos y por haber dedicado tus pocos momentos de descanso a resolver mis dudas, darme ánimos y dibujar todo un horizonte de posibles líneas en las que trabajar juntos. Nuestras conversaciones han hecho que mis ganas de dedicarme a la investigación no hayan hecho sino crecer.

Mis amigos y amigas del máster y de la estancia, Alba Moreno, Amalia Haro, Ana Muros, Dani Galdeano, Edu Pérez, Llanos Navarro, Manu Almagro y Xavi Osorio, me han dado tanto a tantos niveles que no sé bien por dónde empezar. Esta tesis doctoral -y la persona que la ha escrito- están hechas en gran medida de momentos compartidos con vosotras, de vuestra atención y ayuda constantes, de vuestros seminarios diurnos y nocturnos, de nuestros viajes, conciertos y conversaciones. También os debo su banda sonora; este trabajo suena a Soto, a SpokSponha y a todas nuestras barras chorra. Por encima de todo, os debo una manera más crítica, más libre y más bonita de vivir. Alba, tus consejos y tu capacidad para señalar mis tropiezos desde el cariño me ayudaron afrontar algunos de los momentos más difíciles de estos últimos años y finalmente acabaron cambiándome la vida. No puedo estar más agradecido por ello. Ana, mil gracias por haber sabido siempre sacar tiempo y ganas -incluso en tus momentos más duros- para escucharme, apoyarme y ayudarme a repensar todo lo bueno que hay en mi vida y en mi trabajo; gracias también por tu buen gusto, al que le debo la mejor ampliación posible de mi repertorio de camisas. Edu, tu ayuda y tu trabajo han sido fundamentales para terminar de entender gran parte de las ideas aquí expuestas, y tu casa (y la de tus padres) han sido el escenario de algunos de los momentos más bonitos que he vivido estos años. Gracias por tantísimos recuerdos. Llanos, tus palabras de cariño y elogio todo este tiempo han sido muchas veces un freno a los síndromes y metasíndromes del impostor que nos acechan a todas. Te agradezco de corazón todas las ocasiones en las que has hecho las veces de espejo, haciendo así que me sintiera menos solo, menos perdido y más merecedor de mi propio cariño y respeto. Xavi, eres el tipo de investigador que uno desearía encontrar más a menudo en la academia; un modelo de naturalidad y cercanía que muestra cómo la excelencia en el trabajo no va necesariamente ligada a actitudes distantes y envaradas -si acaso, todo lo contrario. Mil gracias por haberme dado tan buenos ratos y haber llenado de carcajadas todos estos años, además de por la increíble portada de tesis que has diseñado. Quiero hacer una mención especial a Dani y a Manu, que han sido mis principales mentores no oficiales desde el máster. Las conversaciones con vosotros son como exposiciones de arte atemporales, a las que uno quiere volver siempre a explorar nuevos aspectos y de las que uno siempre sale más lúcido y más inspirado. Dani, escribiendo

esta tesis he vuelto muchas veces a nuestras tardes de verborrea, flamenco y sol y sombra. En ellas me formé y en ellas continúo formándome. Gracias por tanto, Dani, y gracias también a ti y a Cris por haber sido mi avanzadilla en Granada; me acogisteis tantas veces y con tanto cariño que al final se me olvidó que vivía en otro sitio. Manu, has sido todo un compañero de viaje. Son pocas las veces que apareces mencionado en esta tesis teniendo en cuenta lo mucho que debe esta a la tuya. Me falta aquí espacio y me va a faltar siempre tiempo para terminar de agradecerte todo lo que has hecho por mí, todo lo que me has enseñado y apoyado estos años, todos los momentos que has dedicado a arrojar luz sobre mis dudas, a darme cariño y ánimos para seguir adelante. Tu forma de actuar, de pensar y de sentir son un estímulo constante para mejorar, como investigador y como persona. Gracias por tu amistad, Manu.

Amalia, contigo se me acaban las palabras. Para empezar, te debo una familia en expansión; una que abarca ya, entre otras personas, a tus padres, Lola y Manolo, a quienes tengo que agradecer todo su cariño y apoyo estos meses, y a nuestro gato, Sócrates, a quien tengo que agradecer varias capas de abrigo extra y un renovado amor por los animales. Pero por encima de todo te debo una relación llena de significado, de complicidad, de humor, de ayuda mutua. Esta tesis está escrita en inglés, pero se lee en jiennense; tu amor, tus cuidados, tus palabras de aliento y tus conversaciones son sus principales coautores. Su escenario son nuestras noches de balcón, divagaciones y coreografías en el confinamiento; nuestras tardes en el río y en la placeta del mercado de Bratislava; cristales de colores en el sur, bóvedas de roca en el norte; y ahora por fin nuestra Homa. Es el tiempo que me has regalado, y por eso esta tesis va principalmente dedicada a ti. Tenerte en mi vida ha transformado mi forma de ver el mundo, a los demás y a mí mismo. Me has aguantado en los momentos más amargos y me has llevado de la mano a los más dulces. Me has hecho más seguro y a la vez más autocrítico, y me has dado un espacio en el que poder equivocarme y aprender de mis errores. Me has hecho crecer y desear seguir creciendo contigo. Sobre todo, me has dado un hogar de certezas en estos años nómadas; a través de sus ventanas, todos los mundos posibles son mundos mejores. Te quiero. Gracias por compartir tu vida conmigo.

Publications

Núñez de Prado-Gordillo, M., Alonso-Vega, J., de Pascual-Verdú, R., & Pereira, G. L. (forthcoming). Minding the mind and the "mental" in mental health care: an introduction to the philosophy of mental health and the therapeutic models. In *Aproximaciones al estudio del comportamiento y sus aplicaciones, vol. 3*. Universidad de Guadalajara.

de Pascual-Verdú, R., Núñez de Prado-Gordillo, M., Pereira, G. L., & Alonso-Vega, J. (forthcoming). El análisis de conducta como herramienta de transformación social: ideología y ciencia. In *Aproximaciones al estudio del comportamiento y sus aplicaciones, vol. 3*. Universidad de Guadalajara.

Núñez de Prado-Gordillo, M. (forthcoming). Psicología del sentido común y ciencia del comportamiento: una aproximación antidescriptivista a la relación entre mente y naturaleza. In R. González & M. Colombo (Eds.). *Análisis de la Conducta: Teoría y Aplicaciones Clínicas*. Psara Ediciones.

Núñez de Prado-Gordillo, M., Abalo Rodríguez, I., Estal Muñoz, V., & Froxán Parga, M. X. (2020). Cuestiones filosóficas en torno al análisis de la conducta. In M. X. Froxán Parga (Coord.) *Análisis funcional de la conducta humana: Concepto, metodología y aplicaciones* (pp. 51-79). Pirámide.

Alonso-Vega, J., Ávila-Herrero, I., Núñez de Prado-Gordillo, M., & Pereira Xavier, G. (2020). Análisis de la conducta y prácticas culturales. In M. X. Froxán Parga (Coord.) *Análisis funcional de la conducta humana: Concepto, metodología y aplicaciones* (pp. 179-201). Pirámide.

De Pascual Verdú, R. & Núñez de Prado-Gordillo, M. (2020). Aplicación del análisis funcional a diversos casos clínicos: Análisis funcional de conductas depresivas en una persona afectada de parálisis cerebral. In M. X. Froxán Parga (Coord.) *Análisis funcional de la conducta humana: Concepto, metodología y aplicaciones* (pp. 248-259). Pirámide.

Froxán-Parga, M.X., Ávila-Herrero, I., Trujillo-Sánchez, C., Serrador-Diez, C. & Núñez de Prado-Gordillo, M. (2019). Análisis de la correspondencia Decir-Hacer-Reportar en terapia: un estudio piloto. *Journal of Behavior, Health and Social Issues, 11*(2), 55-68. http://dx.doi.org/10.22201/fesi.20070780.2019.11.2.75671

Froxán-Parga, M. X., Núñez de Prado-Gordillo, M., Álvarez-Iglesias & Alonso-Vega, J. (2019). FBA-based interventions on adults' delusions, hallucinations and disorganized speech: a single case meta-analysis. *Behaviour Research and Therapy*, *120*, 103444. https://doi.org/10.1016/j.brat.2019.103444

Alonso-Vega, J., Núñez de Prado-Gordillo, M., Pereira, G. L., & Froxán-Parga, M. X. (2019). El tratamiento de Enfermedades Mentales Graves desde la investigación de procesos. *Conductual*, *7*, 44-65. https://www.conductual.com/articulos/El%20tratamiento%20de%20enfermedades%20mentales%20graves%20desde%20la%20investigacion%20de%20procesos.pdf

Froxán-Parga, M. X., Calero-Elvira, A., Pardo-Cebrián, R., & Núñez de Prado-Gordillo, M. (2018). Verbal Change and Cognitive Change: Conceptual and Methodological Analysis for the Study of Cognitive Restructuring Using the Socratic Dialog. *International Journal of Cognitive Therapy, 11*, 200-221. https://doi.org/10.1007/s41811-018-0019-8

Froxán-Parga, M. X., Núñez de Prado-Gordillo, M. & de Pascual, R. (2017). Cognitive techniques and language: A return to behavioral origins. *Psicothema, 29*, 352-357. https://doi.org/10.7334/psicothema2016.305

Abstract

Conceptual debates in the field of mental health since at least the second half of the 20th century have tended to revolve around two core issues: the problem of mind, primarily related to the ontological and explanatory status of those mental states and processes that are posited to explain psychopathological behaviors and experiences (e.g., irrational beliefs), and the problem of normativity, related to the role of norms and values in the determination of what counts as "pathological" or "disordered" (vs., for instance, mere social deviancy). In the last decade, following new waves of criticism against traditional nosological tools like the Diagnostic and Statistical Manual of Mental Disorders (DSM-5 and its upcoming revised edition), and the recent emergence of new models and research initiatives both within and without institutional psychiatry (e.g., the Research Domain Criteria initiative by the National Institute of Mental Heath, the enactive approach to psychiatry, etc.), these topics have gained again the traction they lost, to some extent, with the advent of the biopsychosocial model in the 1970's. Since explicit and implicit conceptions of mind and normativity can have a great impact on how mental health problems are conceptualized, assessed, and treated, providing a proper answer to these topics is fundamental for mental health research and practice.

The main goal of this dissertation will be to present a non-descriptivist approach to mental health that deals better with these two problems than both classical and current approaches, and to show its main benefits and implications for the intervention with people with delusions. It will be divided in two main parts. In Part I (Chapters 1, 2, 3, and 4), we'll first introduce the classical and contemporary therapeutic models of mental health problems. After that, we'll explain which are their philosophical underpinnings, and how these yield untenable answers to the problems of mind and normativity. The reason, we'll argue, lies in their common commitment to descriptivism, or the idea that folk-psychological interpretation (i.e., the practice of ascribing mental states to one another and assessing the truth value of such mental-state ascriptions) subserves a primarily descriptive purpose, i.e.,

some possible combination of objects, properties, events, or relations among them. This commitment prevents the development of better approaches to the mental in mental health, confining the range of possible answers to the mind-body problem (the ontological aspect of the problem of mind) between the two dead-ends of reductivism and eliminativism. Then, we introduce our non-descriptivist approach to the mental, based on a recent pragmatist reading of Wittgenstein's and Ryle's work. According to this approach, mental-state ascriptions are truth-evaluable, but they play a primarily evaluative and regulative, rather than descriptive function. Their truth or falsity is not determined then by a description of some state of affairs -neither internal nor external to the agent- but rather depends on the myriad social norms that regulate the practice of folk-psychological interpretation. This, we'll argue, yields a better account of the problems of mind and normativity, hence constituting a sounder conceptual framework for mental health research and practice.

In Part II (Chapters 5, 6, 7, and 8), we'll explain the main benefits of this framework for the intervention with people with delusions. In particular, we'll focus on a longstanding debate about their standard conceptualization as (irrational, or strange) beliefs and its implications for assessment and treatment. We'll first introduce this debate. As we'll see, while antidoxasticists reject that delusions can be properly conceptualized as beliefs, several defenses of doxasticism stress the scientific and ethico-political value of this conceptualization. Drawing from our non-descriptivist framework, we'll point out that doxasticism can and should be primarily defended because of its ethico-political merits, not because it provides a good roadmap for scientific research or clinical practice. On the one hand, we'll claim that this conceptualization is worth retaining because it may help prevent undue and potentially harmful deagentializing practices against people with delusions. On the other hand, we'll point out that traditional cognitivist models, which understand delusions as beliefs "gone wrong" resulting from internal information processing failures, can hinder the efficacy of psychological interventions with people with delusions. We'll then see how non-cognitivist, functional analytic models provide a sounder intervention framework. Jointly considered, the main virtues of these approaches lie in their emphasis on the need for conducting individualized pre-treatment functional assessments of target behaviors, the emphasis on the role of verbal rules in the development and maintenance of psychological problems, and the shift away from "problem reduction" models of recovery. Finally, we'll point out how our pragmatist kind of non-descriptivism can contribute to the progress of these approaches by pointing out some philosophical misconceptions in their theoretical frameworks, which could be limiting their efficacy and their ability to produce clinically significant changes.

Resumen

Desde al menos la segunda mitad del siglo XX, los debates conceptuales en el ámbito de la salud mental han girado en torno a dos cuestiones fundamentales: el problema de lo mental, relacionado con el estatus ontológico y explicativo de aquellos estados y procesos mentales que se postulan para explicar las conductas y experiencias psicopatológicas (por ejemplo, creencias irracionales), y el problema de la normatividad, relacionado con el papel de las normas y los valores en la determinación de lo que cuenta como "patológico" (frente a, por ejemplo, lo que constituye una mera forma de desviación social). En la última década, la nueva ola de críticas a las herramientas nosológicas tradicionales (por ejemplo, el DSM-5 y su próxima edición revisada), así como la reciente aparición de nuevos modelos e iniciativas de investigación dentro y fuera de la psiquiatría institucional (la iniciativa RDoC del Instituto de Salud Mental estadounidense, la aproximación enactivista a la psiquiatría, etc.), han generado un renovado interés por estas dos cuestiones, que parecía haberse perdido, en cierta medida, tras la aparición del modelo biopsicosocial a finales de los setenta. Explícitas o implícitas, las diversas concepciones de lo mental y de la normatividad pueden tener un gran impacto en la forma de conceptualizar, evaluar y tratar los problemas de salud mental; por ello, dar una respuesta adecuada a estas cuestiones es fundamental para la investigación y la práctica clínicas.

El objetivo principal de esta tesis doctoral será presentar una aproximación antidescriptivista al ámbito de la salud mental, capaz de abordar estos dos problemas de forma más satisfactoria que otros enfoques clásicos y contemporáneos, y mostrar sus principales beneficios e implicaciones para la intervención con personas con delirios. La tesis estará dividida en dos partes. En la primera parte (Capítulos 1, 2, 3 y 4), comenzaremos introduciendo los principales modelos terapéuticos, clásicos y contemporáneos, de los problemas de salud mental. Más adelante, explicaremos cuáles son sus fundamentos filosóficos, deteniéndonos en por qué estos dan respuestas inadecuadas a los problemas de la mente y la normatividad.

La razón, como veremos, radica en su compromiso común con el descriptivismo, o la idea de que la práctica interpretativa que caracteriza la psicología popular o del sentido común, basada en las atribuciones de estados mentales y la evaluación del valor de verdad las mismas, tiene un propósito principalmente descriptivo; esto es, el de representar alguna combinación posible de objetos, propiedades, eventos o relaciones entre ellos. Este compromiso impide el desarrollo de formas más adecuadas de dar cuenta de lo mental en salud mental, restringiendo la gama de posibles respuestas al problema mente-cuerpo (el aspecto ontológico del problema de lo mental) y abocándonos a dos callejones sin salida: el reduccionismo y el eliminativismo. A continuación, presentaremos nuestra concepción antidescriptivista de la mente, basada en una reciente lectura pragmatista de la obra de Wittgenstein y Ryle. Según este enfoque, las atribuciones de estados mentales son efectivamente veritativo-evaluables, pero su función es principalmente evaluativa y regulativa, no descriptiva. Su verdad o falsedad, por tanto, no viene dada por una descripción de posibles hechos -ni internos ni externos a la persona- sino que depende de las muy diversas normas sociales que regulan nuestras prácticas interpretativas. Esto, como veremos, constituye una mejor aproximación al problema de lo mental y al problema de la normatividad, constituyendo así un marco conceptual más sólido para la investigación en salud mental y la práctica clínica.

En la segunda parte (Capítulos 5, 6, 7 y 8), explicaremos cuáles son los principales beneficios de este enfoque para la intervención con personas con delirios. En particular, nos centraremos en el debate sobre su conceptualización estándar en términos de creencias (irracionales o extrañas) y sus implicaciones para la evaluación y el tratamiento. Primero presentaremos este debate. Como veremos, mientras que el antidoxasticismo rechaza que los delirios puedan ser correctamente conceptualizados como creencias, distintas defensas del doxasticismo destacan el valor científico y ético-político de esta conceptualización. Partiendo de nuestro marco antidescriptivista, señalaremos que el doxasticismo, si bien no constituye una adecuada hoja de ruta para la investigación científica o la práctica clínica, sí puede y debe ser defendido por sus virtudes en el ámbito ético-político. Por un lado, como veremos, esta conceptualización puede contribuir a prevenir prácticas indebidas y potencialmente dañinas de desagencialización de las personas con delirios. Por otro lado, señalaremos que los modelos cognitivistas tradicionales, que entienden los delirios como creencias "equivocadas", resultado de fallos en el procesamiento interno de la información, pueden socavar la eficacia de las intervenciones psicológicas con personas con delirios. A continuación, veremos cómo ciertos modelos no cognitivistas, los analítico-funcionales, proporcionan un marco de intervención más sólido. Tomados en conjunto, las principales virtudes de estos enfoques residen en su énfasis en la necesidad de llevar a cabo evaluaciones

funcionales individualizadas previas a la intervención sobre las conductas objetivo, el énfasis en el papel de las reglas verbales en el desarrollo y mantenimiento de los problemas psicológicos, y el alejamiento de los modelos de recuperación basados en la mera "reducción de problemas". Por último, señalaremos cómo nuestro enfoque antidescriptivista contribuye al avance de estos enfoques señalando el carácter erróneo de algunos de sus supuestos filosóficos, que podrían estar limitando su eficacia y su capacidad de producción de cambios clínicamente relevantes.

Introduction

The year is 2345. Mental health services have developed along the lines settled by 21st century biological psychiatry. The once long-awaited arrival of precise medicine methods of mental health diagnosis and treatment are now a reality. With regard to nosology, the 26th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-26) is the epitome of biological psychiatrists' long battle against neuro-skeptics and half-baked psy-alternatives: it clearly specifies in a value-free way what exactly mental (i.e., brain) disorders are; mental health problems are described in a dimensional rather than categorical manner, including prodromal syndromes as well; and classification is no longer grounded on mere descriptions of symptom clusters with low levels of reliability and validity, but on specific neural and genetic *biomarkers*. With regard to clinical assessment, high-fidelity neuroimaging techniques are now commonly employed to detect any kind of functional or anatomical deviance from statistically normal brain functioning and on-site genetic testing allows mental health professionals to tell you whether you won or lost the genetic lottery with regard to psychopathological traits and predispositions. Finally, regarding psychiatric treatment, highly sophisticated genetic engineering techniques, neural stimulation procedures, and target pharmacological interventions have been developed to tackle each specific biological abnormality; and the intervention market is ample enough to guarantee individualized treatment for each person.

About the same time, in a galaxy far, far away, a human-like civilization has been discovered in the planet of Karnatahclan. Karnatahclaniards and Earthians first made contact a couple of decades ago, and both civilizations now maintain stable relations through an ongoing series of diplomatic space travels. Karnatahclaniards share with (human) Earthians the vast majority of what we call here "cognitive functions" or capacities: they perceive the world around them in a similar way and have similar conscious or phenomenal experiences

(e.g., inner speech, visual imagery, etc.); they reason in a very similar fashion, since they share our deductive, inductive, and abductive reasoning methods and rules of inference; they have the capacity to learn from experience, store such information in memory, and deploy it when exercising their also comparable decision making capacities; and they have the ability to communicate in a shared language to exchange information, express one's attitudes towards different states of affairs, prescribe certain courses of action, and perform other kinds of speech acts.

Karnatahclaniards also gather in various kinds of social groups, and their civilization shares with ours many of its vices and virtues: they have a similarly intricated currency system, similar economic sectors and a similarly unstable and self-devouring mainstream productive model; they have similarly divided geopolitical regions, which are managed through different kinds of governmental institutions; and they have different cultures, which can be roughly classified in terms of varying social customs, life styles, and patterns of social behavior. However, Karnatahclaniard anatomy doesn't look a single bit like ours. For example, instead of flesh and bones, their bodies are mainly made out of a uniform gooey substance, which Karnatahclaniards can instantly solidify to adopt whatever shape they want. In fact, they're not even a carbon-based, but a silicon-based life form. As a result, their cognitive capacities are realized via completely different physical processes from ours; for example, their communication is mainly based on visual-olfactory stimuli, and words and sentences result from a particular combination of body shapes and varying scent emissions.

Given this futuristic setting, consider the following two cases:

> PURPLE'S MAD MADNESS: Purple, an Earthian predoctoral philosopher, goes to the clinic to undertake a routine occupational medical check-up. Once in the clinic, psychiatrists conduct the regular neuroimaging and genetic tests. The tests reveal that Purple has a specific alteration in their dopaminergic circuitry, one that is the hallmark of mental disorder SKEH-3.2, a sub-subtype of what was formerly known as schizophrenia. Specifically, SKEH-3.2 is characterized by a specific set of behavioral, cognitive, and phenomenological symptoms: rigidity in facial expressions with common social functions -typically smiling- accompanied by monotone and stuttering speech; a specific mix of persecutory and grandiose delusions, typically involving vengeful labor supervisors; and specific auditory hallucinations, typically involving a low-pitched metallic voice that usually warns one about the potentially threatening intentions of other people around and which claims to know more about one than oneself. However, Purple doesn't display any of these superficial symptoms. On subsequent clinical sessions, psychiatrists find out that Purple's neural alteration plays a wholly different causal role in Purple's perception, cognition, and action. Specifically, it provides Purple with an

extraordinary 5-octave vocal range, it facilitates jumping to conclusions when discussing about left-wing policies, and it's strongly linked to Purple's sudden cravings for pulpo a feira. Psychiatrists now strive to reach a conclusion. Does Purple have a mental disorder or not?

YELLOW'S KARNATAHCLANIARD MADNESS: Yellow, a Karnatahclaniard predoctoral astrobiologist, has gone on a trip to Earth to conduct a field research study on our habits and customs. However, something goes wrong with the space cruiser in which they travelled and Yellow ends up lost and alone in the middle of space. After four long years by themself (a considerable period in Karnatahclaniard's time as well), Yellow manages to make contact with a diplomatic carrier on its way to Earth and is finally rescued. As a result of the extreme isolation and constantly stressful situation that Yellow has gone through, they start to display what both Karnatahclaniard and Earthians would evaluate as bizarre patterns of experience, thought, and behavior. To begin with, Yellow displays certain rigidity in their "social gestures" (some particular body shapes). In addition, Yellow has developed a mixed delusional belief system, with both paranoid and grandiose elements: they have become convinced that the reason why the space cruiser collapsed in the first place is because their predoctoral supervisor was feeling threatened by Yellow's allegedly amazing intellectual talents and had devised an evil plan to get rid of them. In addition, these delusional beliefs are being constantly fed back by linguistic (i.e., visual-olfactory) hallucinations that typically warn Yellow about others' potential threatening intentions and pontificate on Yellow's "true" self. Earthian psychiatrists are puzzled: on the one hand, Yellow's superficial symptoms are strikingly similar to those that define the Earthian category SKEH-3.2; on the other hand, Yellow's bodily constitution is radically different from that of Earthians, so it doesn't even make sense to carry out the usual neuroimaging and genetic testing assessment procedures. So, does Yellow have a mental disorder or not?

These sci-fi examples or *thought experiments*[1] are just that: science fiction. However, we believe that they are useful inasmuch as they trigger some of the central questions that have occupied the debates on mental health philosophy, research, and practice during at least the last six decades. Roughly, these debates revolve around the determination of what counts as "mental disorder", "psychopathology", or, less technically, "madness", and what to

[1] In the philosophical literature, a thought experiment is commonly understood as an analytical device whereby a possible world (i.e., a counterfactual situation, different in various respects from how things are in the actual world) is depicted in order to test the strength of certain assumptions and theoretical commitments. They work by triggering possible counterintuitive consequences of such theoretical commitments, and they help to clarify what one might have to accept if one wants to keep the commitment to a certain view of how things are in the actual world (Brown & Fehige, 2022). Our two cases are essentially based on Lewis's (1980) famous "mad pain" and "Martian pain" thought experiments, although they also contain some elements from Schwitzgebel's discussion on "mad belief" (2012) and his "BetaHydrian valuing" example (2013, p. 83).

do about it. A classical philosophical distinction cuts across much of the field of debate: that between matters of *fact*, on the one hand, and matters of *value*, on the other. The fact/value distinction can be captured as a difference between *descriptive, factual*, or "*is-judgements*" (e.g., empirical claims regarding statistically normal biological functions, a species-typical genetic makeup, an organism's dynamics of interaction with the environment, etc.), and that of *normative*, *evaluative*, or "*ought-judgements*" (e.g., normative claims regarding the correctness or incorrectness, meaningfulness or meaninglessness, desirability or undesirability, etc. of certain courses of action, patterns of thought or experiences).

What are the relevant facts, what are the relevant values, and what differential role do they play in mental health assessment and treatment practices are the three overarching questions in the field of mental health. This general distinction, however, cuts across a myriad of different important issues. Our goal in this introductory chapter will be to present a rough sketch of them, as well as to specify the scope and object of this dissertation.

A myriad of questions: the philosophy of mental health

In essence, our thought experiments invite us to think about the primary subject matter of mental health assessment and treatment practices. On the one hand, we might question whether it makes sense to talk about some kind of "mad madness", i.e., a mental health problem whose typical profile diverged radically from what we would normally identify as "psychopathological". In our actual world, the assessment of an individual's mental health necessarily involves a reference to the person's actions, cognitions, and experiences; "having a mental health problem" (whether understood in diagnostic or case formulation terms) is primarily understood as displaying a clinically significant pattern of actions and reactions (American Psychiatric Association, 2013). In Purple's counterfactual world, things would be radically different: mental health assessment would have been reduced to a sufficiently thorough and sophisticated description of an individual's genetic and neural makeup. Is this scenario really plausible? Should Purple's extraordinary 5-octave vocal range or sudden cravings for pulpo a feira concern us, whatever their causes are?

On the other hand, we might also wonder whether it is sensible to talk about some extra-terrestrial or, more specifically, "Karnatahclaniard madness": could someone with such a radically different bodily constitution be diagnosed with a mental disorder[2]? Of

[2] Despite the differences between the concepts of "disorder" and "illness" that have been pointed out in the literature (typically, that the former is purely descriptive while the latter conveys an explanatory character; see Kupfer, 2002, p. 3; Spitzer et al., 1978/2018, p. 2), we'll here use the terms indistinctively, for our discussion is not affected by such distinction. In fact, we'll prioritize the use of "mental health problems", which is not as theoretically loaded as its conceptual counterparts.

course, Yellow's case is quite an extreme one, but we could ask a similar question regarding people with astonishingly different neural makeups –think, for example, of cases like Lorber's recovered hydrocephalics (Forsdyke, 2015; Lewin, 1980; see also Feuillet et al., 2007), where normal functioning was preserved in people with anomalously expanded ventricles (sometimes filling up to 95% intracranial space). In any case, both our sci-fi case and these other real cases pose an important question: what role do the causal underpinnings of certain behaviors and experiences play in their assessment as pathological? And what role should they play in their treatment?

In turn, debates about the primary subject matter of mental health care have been closely linked to discussions regarding the role of the socio-cultural context in which mental health assessment and treatment practices take place. Perhaps one of the reasons why it seems weird to think of Purple's vocal range and sudden cravings for pulpo a feira as pathological is that we find it hard to imagine a socio-cultural background where these behaviors and experiences were somehow condemnable. This invites us to think about the norms and values at play in our actual socio-cultural background: what norms and values do guide our clinical judgements -or which should guide them instead? We can also think about those at play in different social-cultural backgrounds from our own. Imagine that Karnatahclaniards viewed Yellow's phenomenological, cognitive and behavioral patterns as something relatively normal, e.g., as a "second adolescent phase", typical of young Karnatahclaniard academic researchers who have spent some time frictionlessly spinning in the (cosmic) void. Would it then make sense to diagnose Yellow with a mental disorder, or to try to convince Karnatahclaniards that their social norms are wrong or that they should change them in some way? Of course, the case at hand here might seem quite obviously improbable, but we could ask a very similar question regarding the not-so-old universalizing pretensions of certain approaches to mental health theory and practice. Once we reckon that different stakeholders (users, practitioners, governmental institutions, people from different cultural backgrounds, etc.) might have conflicting values, what does that tell us about the very concept of "mental disorder"? Would it be possible, or even desirable, to provide a definitive, universal, value-free notion of what constitutes a mental health problem?

All these questions set much of the scene for what has come to be known as the *philosophy of psychiatry*, although the term *philosophy of mental health* is perhaps more inclusive and appropriate (see Aftab, 2021; Banner & Thornton, 2007; Fulford et al., 2013a; Murphy, 2020; Thornton, 2007; see also Jaspers, 1913). Roughly, the philosophy of mental health can be defined as an applied interdisciplinary branch of philosophy whose double aim is to a) clarify the main conceptual commitments and problems of the different approaches to

mental health; and b) open up and advance new possible ways of conceiving, assessing and dealing with mental health problems. To do so, philosophers of mental health employ diverse conceptual tools and methods from different philosophical traditions (e.g., continental philosophy, analytic philosophy, etc.) and related subdisciplines (e.g., phenomenology, hermeneutics, the philosophy of science, the philosophy of mind, the philosophy of language, the philosophy of action, etc.) with the empirical evidence provided by the different basic and applied psychological sciences (see Aftab, 2021; Banner & Thornton, 2007; Fulford et al., 2013a; Graham, 2010b; Graham & Stephen, 1994; Kendler & Parnas, 2008; Murphy, 2020; Thornton, 2007; Varga, 2015, 2017).

As we might have glimpsed from our introductory thought experiments, the philosophy of mental health is a *complex* field -to say the least. In it, myriad different questions pertaining to different disciplines, spelled in different technical languages, and addressed from different conceptual frameworks, crisscross continuously. Four major themes can be identified though, some of which have already emerged in our introductory thought experiments. The first one corresponds to what we might call the *analogy problem*, or the problem of the analogy between somatic and mental health problems (e.g., see Boorse, 1975, 1997, 2014; Fulford & van Staden, 2013; Graham, 2010b; Kendell, 1975; Kendler, 2016; Szasz, 1961/1974; Thornton, 2007; Varga, 2015, 2017; Wakefield, 1992, 2007; see Fulford et al., 2013a). The central question here is the following: is there something intrinsically special to what falls under the rubric of "mental disorders", or are these mere variants of somatic (e.g., neurological) disorders? Typically, the analogy problem has been central to the longstanding debates about the legitimacy of medical approaches to mental health, which at least span from the 1960's -when Szasz (1960/1974) attacked the "myth of mental illness"- to the present -when the publication of the DSM-5 and its upcoming revision in 2022 have caused a new spate of criticism.

Another related theme involves the *boundary problem*, or the problem of the demarcation[3] between "mad and bad", as Fulford & van Staden (2013, p. 393) put it. The pressing question here is how to tell apart what may be assessed in a given community as "wrong", "undesirable", or "bizarre", from what may be rightfully assessed as "pathological" (e.g., see also Aftab & Rashed, 2020; Boorse, 1975, 1997, 2014; Graham, 2010b; Kendler, 2016; Kingma, 2013; Leoni, 2013; Thornton, 2007; Varga, 2015, 2017; Wakefield, 1992, 2007; see Fulford et al., 2013a). Many have taken this problem to involve the telling apart of facts from values; while

[3] In the philosophy of science, the "demarcation problem" refers to the problem of the distinction between "scientific" and "non-scientific" or "pseudo-scientific" disciplines or methods of inquiry. This has nothing to do with the demarcation problem in mental health theory and practice. To avoid confusion, we will use "boundary problem" hereafter to refer to the latter.

"wrong", "undesirable", or "bizarre" convey an evaluative force, it's assumed that "ill" or "pathological" should be spelled out in purely descriptive terms. Hence, many think of the analogy and boundary problems as mere variations of the same theme: if the diagnosis of somatic disorders is supposed to be "value-free", then assuming a strict analogy between mental and somatic disorders would amount to saying that mental health assessment practices are also purely descriptive. However, we might question both a) whether health assessment practices (of either mental or somatic disorders) really are or will ever be value-free, and hence whether accepting the analogy would somehow make obvious what tells apart "mad" (or "ill", more broadly) from "bad"; or b) whether rejecting the analogy automatically commits one to the view that either mental or somatic health assessment practices are necessarily "value-laden" (see Fulford & van Staden, 2013; Thornton, 2007).

The boundary problem is also related to debates around the *continuity thesis* (e.g., see Bentall, 2003; Bortolotti, 2010, 2012; Eysenck, 1959, 1964; Froxán-Parga, 2020; Hayes et al., 1999, 2001; Layng & Andronis, 1984; Lindsley, 1964; Rosenfarb, 2013; Skinner, 1953; Sturmey, 2020; Wong, 1996, 2006, 2014; Tumulty, 2012), i.e., the idea that psychopathology lies on a continuum with non-clinical behaviors and experiences. Supporters of the continuity thesis sometimes yield it as an argument against the analogy between mental and somatic health problems, as if the defense of the continuity thesis would reveal the fallacious character of the analogy; likewise, some detractors yield it as an argument for the analogy. In doing so, both proceed as if what characterized "real" (i.e., somatic) illnesses was a sharp discontinuity between health and pathology; an assumption that has been sometimes challenged by certain conditions whose pathological character no one disputes (e.g., diabetes), but for which no clear boundary can be pointed out (e.g., Engel, 1977).

Finally, even if agreement over their value-laden or value-free character was reached, there would still be the problem of determining how best to causally explain, predict, and intervene on mental health problems. We might refer to this as the *priority problem*, related to the question as to whether some scale of analysis (e.g., the biological, the psychological, the social, etc.) should be prioritized in the conceptualization, assessment, or treatment of mental health problems (e.g., see Andreasen, 1997, 2001; Craddock et al., 2008; de Haan, 2020a, 2020c; Insel & Cuthbert, 2015; Nielsen & Ward, 2018; Kendler, 2005; Kendler & Parnas, 2008; Pérez-Álvarez, 2004, 2012; see also Fulford et al., 2013a). In empirical terms, the "priority wars" among therapeutic approaches have been closely linked with disputes regarding the relative efficacy of pharmacological vs. psychosocial procedures in the intervention with people with different mental health problems; in this respect, a recent umbrella-review of meta-analyses (Leichsenring et al., 2022) suggests no clear winner.

In conceptual terms, these priority wars usually concern which should be the proper unit of analysis in mental health care -whether the individual's brain, their behaviors, cognitions, and experiences, or rather their social milieu. During the 1960's-1980's, these conceptual debates were especially acute, but the emergence of the now pervasive biopsychosocial model seemed to put an end to them -at least for a while. However, a new puzzle emerged then, which has been growing in relevance during the last decades: the *integration problem* (e.g., see Aftab, 2021; Aftab & Nielsen, 2021; Bolton & Gillett, 2019; Ghaemi, 2009, 2010; de Haan, 2020a, 2020b, 2020c; Kendler, 2005, 2016; Kendler & Parnas, 2008; Matthews, 2013; Pilgrim, 2015; Van Oudenhove & Cuypers, 2014; Walter, 2013; see Fulford et al., 2013a). Once we accept the kind of holistic approach advanced by the biopsychosocial model, how can we reconcile the different causal-explanatory projects of different approaches to mental health problems, which locate the relevant causal phenomena at different scales of analysis? In other words: how can we achieve a conceptually coherent and therapeutically enhancing integration of mental health approaches without falling into an eclectic chaos?

Apart from the aforementioned fact/value distinction, two other conceptual distinctions cut across these four major themes. Since they help to clarify the many different angles from which one may approach mental health philosophy, research, and practice, we'll review them here briefly before introducing the present dissertation. The first one is the *personal/subpersonal* distinction (see Pinedo-García & Noble, 2008; Pinedo-García, 2014, 2020), which distinguishes between modes of description and explanation; specifically, between a *rationalizing*[4] mode and a non-rationalizing or "purely causal" one. While the former involves the consideration of the individual as a rational or intelligible agent (i.e., one which acts or fails to act on the grounds of certain reasons and motives) the latter just depicts the individual's behavior in arational terms, i.e., as a neither rational nor irrational creature whose behavior is just a causal effect of some particular combination of factors. Another way to put this distinction is to differentiate between reasons and causes, i.e., between the logical or rational conditions for acting or reacting in certain ways and the empirical conditions that are causally related to certain patterns of activity. Note that the use of the word "person" here conveys a special status which may or may not be granted to every organism (e.g., somewhat similar to the use of the term in certain legal contexts) and not as a synonym of "individual" or "subject". Debates about the primary subject matter of mental health care -

[4] "Rationalizing", as we'll use the term here, just refers to the attempt to make a person's actions and reactions intelligible (i.e., meaningful, subject to assessments about the possible reasons for acting or reacting in such ways). This must be distinguished from a different, pejorative sense of the term, whereby "rationalizing" amounts to providing reasons for one's actions *which are not* the actual reasons why one acted in a certain way (for instance, when a person acts badly and then tries to excuse their behavior to themselves).

whether it mainly has to do with a person's reasons or motives to act in certain ways or rather with the biological, environmental, or social causes of such behaviors- have been recurrent (e.g., see Szasz, 1960/1974, 2004, 2011; see also Schaler, 2004).

This distinction sometimes overlaps with another one: the one that is often established between *lower-order* and *higher-order scales of analysis* (see de Haan, 2020a, 2020b; Engel, 1977; Kendler & Parnas, 2008). It's common ground that effective interventions should be based on well-established empirical evidence regarding the causal processes involved in the development and maintenance of mental health problems. However, deciding where to look for these causal processes is not a straightforward matter, as the "priority wars" reveal; our choices will vary depending on what we take to be our basic or primary unit of analysis. Should we start from the analysis of the individual's neural activity or should we instead start from the analysis of the individual's interaction with the environment? In fact, should we privilege some given level of description and explanation at all, or should we instead aim at a more holistic causal understanding of the causes of mental health problems?

Although sometimes conflated, here we'll treat these distinctions as conceptually orthogonal. For example, a higher-order, yet subpersonal explanation of a given event would involve taking a higher-order unit of analysis (e.g.., the organism-environment system) as a starting point and attempting to account for it in arational, purely causal terms. Alternatively, a lower-order, yet personal explanation would involve taking a lower-order unit of analysis (e.g., the organism or some of its parts) and describe or explain it in rationalistic terms. Certain behavior analytic explanations of mental health problems would constitute an example of the former, while some cognitivist accounts of the functioning of the brain would be an example of the latter.

Situating the present dissertation

This dissertation is a piece of work within the field of philosophy of mental health. As such, its interdisciplinary character is reflected through and through; in its thematic scope (which ranges from broad philosophical problems to specific clinical concerns), conceptual framework (which aims to integrate positions across philosophy and psychology), methods (which combine conceptual analysis with the review of empirical literature), or even its writing style and imaginary audience. In this introduction, we'll lay out the range of topics to be addressed, the main line of argument, as well as the structure of this dissertation.

Thematic scope

Regarding the topics of interest, our discussion will mainly delve with two inter-related classical philosophical problems which, from our perspective, lie at the core of the four major

themes of the philosophy of mental health that we've mentioned above. These are the *problem of mind* and the *problem of normativity*. The former comprises a series of inter-related issues regarding the relationship between mind and body, mind and world, and mind and language. We'll mainly focus on its ontological aspect: what kind of *thing* is the mind -if any? Is it a spooky phantom, a neural homunculus, a confusing byproduct of folk explanations of behavior? Does it cause behavior or is it epiphenomenal? The relevance of these questions to the field of *mental* health is obvious. To determine how (or even *if*) minds can have health-related problems, we must first have a clear grasp of what minds are.

By contrast, the problem of normativity is related to the place of values and norms in a naturalist understanding of the world. Are there intrinsically valenced properties of the world, e.g., intrinsically moral, aesthetic, logical, or, more broadly, *normative* in some particular way? Are these only "in the eye of the beholder"? And what role do our social backgrounds play here, if any? Once again, these questions are central to the field of mental *health*: if "health" or "disorder" are at least partially evaluative concepts, as many authors now agree (e.g., see Fulford & Van Staden, 2013; Thornton, 2007, 2014; Varga, 2015, 2017), we must understand what the values and norms involved amount to, and how these might fit within a naturalist approach to mental health.

In addition, we'll analyze the impact of these discussions in one of the mental health problems that has received most attention: delusions. Delusions and other psychotic phenomena have attracted the interest of mental health scientists and philosophers for a number of reasons. A primary reason is the enormous health, social, and economic burden associated to the diagnosis of schizophrenia. Recent prevalence estimates situate it around 0.32% worldwide (23.6 million people approximately), with varying results across age cohorts (e.g., 0.49% for those between 20-54 years old), regions, (e.g., USA = 0.52; Spain = 0.35%), etc. (see Institute for Health Metrics and Evaluation, 2022). Despite its relatively low prevalence, the Global Burden of Disease Study (GBD) ranked it 19th among the leading causes of disability worldwide in 2017 (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018) and 22rd within 20-45 years-old adults in 2019 (GBD 2019 Diseases and Injuries Collaborators, 2020), accounting for a 12.2% of the health burden associated to mental disorders (GBD 2019 Mental Disorders Collaborators, 2022). In addition, its associated economic burden has been estimated to range globally from 0.02% to 1.65% of the country's Gross Domestic Product (e.g., USA = 0.15 – 0.61%; Spain = 0.26%) (Chong et al., 2016). Related social and human ills include the increased risk of mortality, common human rights violations, stigmatization and the concomitant social and professional exclusion frequently faced by people with a diagnosis of schizophrenia (see WHO, 2022). Since the presence of delusions is one of

the primary symptoms that may grant a diagnosis of schizophrenia (APA, 2013), targeting the processes giving rise to distressful delusional experiences is a top priority.

Another reason why delusions merit special consideration is that they have been historically considered the "hallmark of insanity" (see Berrios, 1991, 1996, p. 87; Young, 1999). Already in the 18th century, for example, Battie (1758, pp. 5-6) considered "deluded imagination" to be "not only an indisputable, but essential character of madness". This tendency to place delusions, hallucinations, and other psychotic phenomena at the conceptual core of psychopathology can still be felt today. In this sense, schizophrenia has often figured at the center of debates around the analogy, boundary, and priority problems. On the one hand, it has usually been treated as the crucial gauntlet for both supporters and detractors of the medical model, as if the legitimacy of psychiatry as a medical discipline hinged on whether schizophrenia in particular could be considered a properly medical condition or rather amounted to mere "disapproved conduct" (Szasz, 1976, p. 311; see also Laing, 1960/2010; Kendell 1975, 2004; Schaler, 2004). Relatedly, delusions and other psychotic phenomena have been traditionally viewed as beyond the reach of psychosocial interventions. Although this impression has been gradually changing (Bentall, 2003, NICE 2009, 2014), psychological interventions are still often conceived as having a merely adjuvant role, and many still have reservations regarding their efficacy (e.g., Lynch et al., 2010). By contrast, in this dissertation we want to advance arguments in favor of the idea that, when conducted under certain conditions, psychological interventions have much to offer.

Synopsis of the argument

We'll begin our discussion by investigating how the problems of mind and normativity have typically appeared reflected in periodic debates regarding the proper subject matter of the mental health sciences. Following the usual narrative, we'll trace them back to the often-abhorred Cartesian theory of mind. As we'll see, much of the recent conceptual history of mental health can be understood as a continuous strive to overcome the many edges of this theory, with a particular -and perhaps exaggerated- emphasis on the problem of dualism. We'll claim that, in their rejection of dualism, many scientific approaches to mental health have implicitly or explicitly leaned towards two possible strategies: *reductivism* or *eliminativism*. Roughly, while the former implies the identification of mental properties -including mental health problems- with non-mental, "natural", or "physical" properties, the latter involves denying their existence altogether.

Despite their commonality, these two strategies and the range of mixed options in between ultimately fail to provide a satisfactory answer to the problems of mind and normativity at once; in particular, we'll argue, they fail to account for the normative dimension

of mental health, and therefore, given the tight link between mind and normativity, for what's specifically "mental" about mental health problems. Moreover, we'll see that, when considered in general, both lead to untenable *self-defeating* forms of naturalism. In line with recent proposals (e.g., see Aftab, 2021; de Haan, 2020a, 2020c, 2021; Fulford & Van Staden, 2013; Thornton, 2007), this dissertation will constitute an effort to resist reductivist and eliminativist tendencies, without letting up the goal of establishing a non-Cartesian naturalist conceptual framework for mental health care. In fact, we'll argue that the pull towards reductivism and eliminativism results from an implicit commitment to the "logical mould" of Cartesianism (Ryle, 1949/2009, p. 9); specifically, it stems from an underlying commitment to *descriptivism*, or the idea that our mental-state ascriptions (i.e., sentences of the form "*S* has or is in $\mathcal{X}$ mental state") ultimately describe or represent some state of affairs (i.e., some particular combination of objects, events, properties, or relations among them) (see Chrisman, 2007). There's indeed a long tradition in the psy-sciences of conceptualizing mind and language as primarily representational devices -as *mirrors of nature*, in Rorty's (1979) words. In the case of mental language, these two mirrors often face each other: mental-state ascriptions are then taken to represent representational machines; folk psychology, or the practice of interpreting each other in terms of our mental states, is thus viewed as a *theoretical* effort at mirroring each other's mirroring, "glassy essences" (Rorty, 1979)[5]. This representationalist conceptions are even more firmly rooted in the field of mental health, where psychological problems are often described in terms of maladaptive, distorting, and distorted representations; in fact, the kind of phenomena that characterizes most mental health problems (e.g., anxious anticipation, depressive rumination, delusions and hallucinations, etc.) often serve themselves as a major inspiration for the mirroring metaphor.

In this dissertation, we'll contend that, in order to develop a fully non-Cartesian approach to mental health care -one that resists the pull towards dualism, on the one hand, and reductivism or eliminativism, on the other- we must first abandon the commitment to descriptivism and the view of folk psychology as a pre-scientific, mirroring theory. We'll propose instead a *non-descriptivist* approach to the problems of mind and normativity, in an effort to provide an alternative conceptual framework for mental health theory and practice -a "philosophy of mental health without mirrors", we might say. In particular, our approach

[5] Despite our use of Rorty's metaphor, and although we share his anti-representationalism and its philosophical background, we won't stick to his own approach. Probably, our approach is more closely related to Price's (2011) "naturalism without mirrors" and his global expressivist view of language, although we won't endorse it in particular either. Rather, our discussion will primarily draw from a pragmatist reading of Wittgenstein's and Ryle's work, based itself on the work of some analytic philosophers at the University of Granada whom bring to bear various strands of post-Rortyan analytic thinking to the analysis of mental language (see Chapter 4).

draws from a pragmatist approach to the philosophy of mind and language, grounded in a Wittgensteinian and Rylean understanding of mental-state ascription practices. The main tenet of this approach is that these practices have a primarily evaluative, not descriptive function. Folk psychology is thus not some kind of pre-scientific theory, aimed at the goals of predicting and controlling behavior or its underlying causes; rather, it features in rationalizing accounts of each other's behavior and experiences, which aim to make it meaningful or intelligible and to assess it with regard to different normative standards.

Once we've laid out the whys and wherefores of our non-descriptivist approach, we'll put it to work in the analysis of some contemporary debates regarding the intervention with people with delusions. In particular, we'll focus on an ongoing debate concerning the *doxastic status of delusions*, i.e., their usual conceptualization, assessment, and treatment as beliefs or belief-like states (in particular, as *irrational* or *strange* beliefs). This characterization of delusions as beliefs is commonplace in the literature: it's the usual definition in the DSM-5 (APA, 2013, p. 87), as well as in cognitive models of delusions (e.g., Alford & Beck, 1994; Coltheart et al., 2011). Different parties have advanced arguments for and against this default doxastic definition, highlighting its potential benefits and shortcomings. Those against it (i.e., *anti-doxasticists*) usually converge in pointing out that the complexities of delusional phenomena are not well captured by doxastic definitions, and that this, in turn, could be limiting the development of new scientific theories and clinical procedures (e.g., Currie, 2000; Schwitzgebel, 2012). *Pro-doxasticists*, on the contrary, mainly argue that understanding delusions as beliefs has two major advantages: a) that it reflects how currently prevailing scientific theories understand them, leaving us in a better position to understand their causes and possible treatments; and b) that it provides a way to understand the intelligibility of delusional experiences, hence encouraging attributions of agency to people with delusions and providing some sort of conceptual barrier against unjust or abusive treatment practices (e.g., Bayne & Pacherie, 2005; Bortolotti, 2010).

Here we'll address how the adoption of a non-descriptivist perspective may open up new answers to this debate and promote better ways to intervene with people with delusions and other psychotic experiences. Specifically, we'll claim that, from our non-descriptivist approach, questions regarding the doxastic status of delusional phenomena should be regarded as orthogonal to questions regarding the causes behind their development and maintenance: whether delusional patterns of actions and reactions can be correctly interpreted in terms of beliefs needn't have any significant implications for the analysis of their causes. Instead, we'll advance a different defense of doxasticism: one which emphasizes its ethical, rather than scientific virtues. In this sense, we'll argue that the kind of doxasticist

approach at play in cognitive models of delusions not only faces several conceptual problems, but might also be detrimental to intervention: by putting undue emphasis on putative internal causal factors, cognitive models deflect attention away from environmental sources of control that play a major role in the development and maintenance of delusional phenomena.

Instead, we'll encourage the exploration of non-cognitivist approaches to the intervention with people with delusions. In particular, we'll review those stemming from functional analytic approaches to psychology, whose core tenet is that psychological intervention must always start from the functional analysis of *behavior*, understood in terms of the functional relations that are established between an individual's responses and their natural and social environment (see Skinner, 1953, 1974; see also Froxán-Parga, 2020; Hayes et al., 2001). We'll defend that this functional analytic model provides the proper framework for conducting interventions with people with delusions with an individually tailored, formulation-based, and causal-interventionist focus. Finally, we'll point out how a non-descriptivist approach to mental-state ascriptions may help to overcome certain limitations of functional analytic approaches to delusions by providing a better response to traditional objections raised by cognitive models. Hopefully, this might serve as a starting point towards a more general non-descriptivist and non-cognitivist approach to mental health.

Before we move on, a few caveats are in place. Firstly, some may think that the final connection between our Wittgensteinian and Rylean approach to mental-state ascriptions with the defense of functional analytic approaches to psychological intervention comes as no surprise. After all, both have been classically associated to "behaviorist" views of the mind; while the former would correspond to the so-called "logical behaviorist" approach to the philosophy of mind, the latter would correspond to the often-unqualified "psychological behaviorist" approach to the philosophy of psychology.

However, we've explicitly avoided the connection among these two approaches under the common label of "behaviorism". The main reason is precisely that we wanted to avoid such associations. To begin with, both "behaviorism" and its "logical" and "psychological" variants are umbrella-terms that fail to capture the radical points of divergence among the multiple different explanatory projects that fall under them (e.g., see Zilio et al., 2021); the overwhelmingly different readings of Wittgenstein's and Ryle's work, in the case of the former, or the anti-thetic relation between methodological and radical behaviorists, in the case of the latter, are just some common examples of this. Adding to that, while Wittgenstein's and Ryle's kind of non-descriptivism can be seen as an attempt to spell out the logical or otherwise normative relations that determine the meaning of psychological predicates,

functional analytic-oriented researchers are primarily concerned with the analysis of the causal or functional relations that produce and maintain behavior, regardless of its normative character. To be sure, it's true that there are certain undeniable affinities between our particular reading of Wittgenstein's and Ryle's work and the functional analytic approach to psychology; for instance, both draw from the idea that minds are not special (nor non-special) sorts of things which cause behavior, and both share, to some extent, the pragmatist intuition that it's our social niches what ground our linguistic practices. However, here we'll rather emphasize their differences, and we'll stress how our non-descriptivist approach yields a better understanding of our folk-psychological interpretative practices; one which, as we view it, enriches functional analytic approaches, liberates them from certain misguiding assumptions, and may lead to some practical benefits.

Secondly, we would like to stress that the main contributions of this dissertation are negative, rather than positive. Our efforts are primarily devoted to clarifying how the "mental" in mental health problems *should not* be conceived of, and although we highlight some of the potential implications of our approach, the elaboration of a full-fledged positive characterization of how they *should* be conceived of is relatively lacking. Nonetheless, we still think that our effort to conduct a systematic criticism of current approaches is a good first step towards the development of a truly integrative and conceptually sound approach to mental health research. We hope the reader finds our approach and arguments appealing for that purpose; if so, then this dissertation will have proved its value.

Finally, we would like to note that, although this whole dissertation is about mental health and the experiences of people with mental health problems, it's minimally informed regarding the political perspective of those whose lives have been significantly traversed by the experience of mental health problems and mental health solutions -which have often been part of the problem as well. We've included several references to this topic, as well as some of the author's own experiences, which appear reflected at some points of the dissertation. We've also attempted to systematically avoid the use of derogatory or pathologizing language, employing instead expressions that reinforce the agential status of people with mental health problems. However, references to user-led research or the mad pride and neurodiversity movements are scarce, to say the least. Therefore, the main conclusions of this dissertation are still pending a properly thorough analysis of their potential implications for the political struggles against sanism and other forms of structural inequalities.

Structure of the dissertation

Our dissertation will be divided in two main parts. In Part I (Chapters 1, 2, 3, and 4), our main goal will be to lay out the philosophical underpinnings of the main theoretical approaches to

mental health, as well as to introduce our alternative non-descriptivist framework. In Part II (Chapters 5, 6, 7, and 8), our goal will be to explore the conceptual and practical benefits of our non-descriptivist framework for the intervention with people with delusions, highlighting its implications for the conceptualization, assessment, and treatment of delusional phenomena.

Chapter 1 will serve to introduce the problems of mind and normativity as they've appeared in the recent history of mental health. We'll begin by presenting the classical debates among the different *therapeutic models*. We'll first introduce the medical model (section 1.1.), distinguishing between two main interpretations of this approach: a minimalist interpretation, according to which mental disorders are merely diagnostic kinds, and a strong interpretation, according to which they are natural kinds. We'll review the main critical approaches to mental health (section 1.2.), putting a special emphasis on Szasz's attack on "the myth of mental illness". These criticisms mainly focus on the analogy and boundary problem. By contrast, classical psychological models (section 1.3.) attacked the medical model on account of its neglect of environmental factors in the development and maintenance of psychopathology. In particular, we'll primarily focus on first-wave and second-wave behavior therapy, the former comprising behavior therapy and applied behavior analysis (section 1.3.1.), and the latter referring to cognitive behavioral therapy (section 1.3.2.). We'll then introduce what is seen by many as the prevailing contemporary approach to mental health: the biopsychosocial model (section 1.4.). We'll see how, despite its "mainstream ideology" character, many have recently criticized this model on account of its undue eclecticism. After that, we'll review three contemporary approaches to mental health (section 1.5.), which diverge on the emphasis they put on different scales of analysis: a) third-wave biological psychiatry, which puts a stronger emphasis on "brain circuitry" as the integrative unit of analysis (section 1.5.1.); b) contemporary functional analytic models, which emphasize the role of the person's context in the development and maintenance of mental health problems (section 1.5.2.); and c) the enactive approach to psychiatry, which aims to integrate different scales of analysis in mental health theory and practice (section 1.5.3.). We'll conclude by pointing out how the problems of mind and normativity underlie the four major themes of the philosophy of mental health, as well as the debates among therapeutic models.

In Chapter 2, we'll address these two problems. We'll begin by tracing them back to the Cartesian theory of mind (section 2.1.); in particular, we'll claim that the former results from Descartes's attempt to account for the latter (section 2.1.1.). We'll then characterize Cartesianism through a series of inter-related ontological and epistemological commitments (sections 2.1.2. and 2.1.3.), namely *factualism*, *mental causalism*, *intellectualism*, and

representationalism. In addition, we'll claim that the whole Cartesian theory of mind is built upon an implicit commitment to descriptivism, or the idea that the primary function of mental language is to describe events, objects, properties, or relations among them (section 2.1.4.). This view of the mental ultimately leads to a series of conceptual puzzles, among which we'll focus on the *mind-body problem*. We'll then introduce the main contending approaches to this issue (section 2.2.), which draw from a shared standard view of folk psychology as a causal-explanatory practice (section 2.2.1.). Drawing from a common commitment to ontological naturalism, three different kinds of naturalist approaches to the mind can be distinguished (section 2.2.2.): a) ontologically conservative approaches, which share the assumption that mental objects and properties can be reconceptualized in naturalistic terms; b) ontologically revisionary approaches, which assume that a mature science of behavior will eventually dispose of part of our mentalistic talk; and c) ontologically radical approaches, which assume that the mental is individuated by a series of necessarily non-natural properties and hence is incompatible with a naturalist worldview. After sketching out their different varieties, we'll analyze how they've appeared reflected in the different therapeutic models seen in Chapter 1 (section 2.3.). Finally, we'll claim that the reason why they cannot provide a satisfactory account of the mental aspect of mental health is that they all fail to account for the tight connection between mind and normativity (section 2.4.). This, we'll advance, is due to their failure to reject the underlying descriptivist approach to folk psychology.

Chapter 3 will focus on the analysis of the "dogma of descriptivism" in its various forms, as well as on how it unnecessarily constraints our conception of the place of mind on nature. We'll begin by examining this dogma (section 3.1.), situating its roots in the "standard" or "mindreading" conception of folk psychology, i.e., the idea that folk-psychological interpretation subserves a primarily theoretical or causal-explanatory function (section 3.1.1.). We'll then distinguish several varieties of descriptivism (section 3.1.2.), placing a special emphasis on the two possible versions of the descriptivist dogma: a) an affirmative, *shallow* version, which amounts to stating that all indicative sentences or a particular subset of them (e.g., mental, logical, ethical, and similar expressions) describe or represent some state of affairs; and b) a conditional, *deep* version, which implies that if and only if a sentence describes or represents some state of affairs, then it's truth-apt (i.e., it can be assessed in terms of its truth or falsity). We'll then see how a naturalized version of the latter (section 3.1.3.) leaves only two possible ways out of the mind-body problem: a) reductive compatibilism, according to which mental-state ascriptions are identical to descriptions of material events and thus their truth-evaluability is compatible with naturalism; or b) non-reductive incompatibilism, according to which mental-state ascriptions cannot be translated to descriptions

of material events and thus their truth-evaluability is incompatible with naturalism. None of them, as we'll see, is a viable option (section 3.2.); in particular, we'll claim that, despite their respective virtues (section 3.2.1.), their commitments to either reductivism or incompatibilism eventually lead to a self-defeating kind of naturalism, i.e., one which defeats its own logical axioms (section 3.2.2.). After considering and rejecting the possibility of alternatively endorsing non-naturalism about the mind (section 3.2.3.), we'll formulate the descriptivist's paradox, i.e., the *puzzle of translatability*, whereby we seem forced to choose between two flawed self-defeating forms of naturalism or a self-defeating normativism; both, we'll claim, fail to provide a proper account of the mind-body problem and the problem of normativity at once (section 3.2.4.). The challenge will thus be to find a way out of this paradox through a non-reductive, yet compatibilist form of naturalism.

In Chapter 4 we'll address this challenge. Specifically, we'll defend that the solution resides in adopting a non-descriptivist, post-ontological account of the place of mind on nature. We'll begin by introducing our approach (section 4.1.). After initially mapping the different varieties of non-descriptivism (section 4.1.1.), we'll endorse a Wittgensteinian, pragmatist kind of non-descriptivism, characterized by three core assumptions: a) the idea that the meaning of an expression is given by its possible uses in different "language games", which depend on the inferential connections that it has with other expressions within a language; b) that these are ultimately grounded on the different social practices in which we're systematically trained by our community; and c) that there need not be any common, necessary condition to all language games (section 4.1.2.). We'll then apply this non-descriptivist framework to the analysis of our folk-psychological interpretative practices (section 4.2.). Firstly, we'll reject descriptivism in its affirmative form, pointing out a number of reasons against the idea that mental-state ascriptions describe or represent some particular state of affairs (section 4.2.1.). Instead, we'll defend that mental-state ascriptions don't play a descriptive, but an *evaluative* and *regulative* function (section 4.2.2.); they do not primarily figure in nomological accounts of behavior, aimed at its causal explanation, prediction, and control, but rather in rationalizing accounts, whose goal is to make one's actions intelligible and evaluable in terms of their accordance with different normative standards (i.e., of rationality, morality, well-being, etc.). Finally, we'll also see how Wittgenstein's and Ryle's work allows us to challenge the deeper, conditional form of descriptivism (section 4.2.3.). From this perspective, mental-state ascriptions *are* truth-evaluable, but their truth or falsity is not given in terms of their representational capacity; instead, it depends on a myriad of social norms. Our pragmatist kind of non-descriptivism thus encourages a post-ontological view of the mind; rather than thinking in terms of bizarre metaphysic pluralities, it endorses

a pluralistic view of the criteria that competent language users employ to decide on the truth or falsity of different expressions (section 4.2.4.). This approach thus affords a non-reductivist, yet compatibilist kind of naturalism about the mind, i.e., one that preserves the truth-aptness of mental-state ascriptions while maintaining the commitment to ontological naturalism. This way, it avoids both the mind-body problem and the normativity problem.

Once we've laid out the key features and advantages of our proposal, in Part II we'll put it to work in the analysis of the debate around the doxastic status of delusions (i.e., their conceptualization as beliefs) and its clinical and scientific implications. In Chapter 5, we'll present the main arguments for and against doxasticism about delusions. We'll begin by introducing the main contending positions in the debate around the typology problem, i.e., the problem of what kind of mental state delusions are (section 5.1.). After presenting the standard doxasticist approach to delusions, we'll review its main antidoxasticist criticisms, which stem from two main theories of belief: interpretivism and functionalism, according to which beliefs are individuated in terms of their stereotypical rational or causal profiles, respectively (section 5.1.1.). Accordingly, antidoxasticists deny a belief-status to delusions on the grounds that many people with delusions exhibit inexcusable deviations from stereotypical belief-like causal roles or rationality constraints (section 5.1.2.). We'll then review the two main strategies to account for these criticisms (section 5.2.): a) revisionist doxasticisms, whose defense of doxasticism about delusions draws from the revision of the functionalist and interpretivist frameworks (section 5.2.1.); and b) non-revisionary doxasticism, whose defense of doxasticism draws from the rejection on functionalism and interpretivism and the endorsement of an alternative, cognitive-phenomenological theory of belief (section 5.2.2.). Finally, we'll present the two main desiderata behind the defense of doxasticism (section 5.2.3.): a) the scientific desideratum, according to which doxasticism leaves us in a better position to understand the factors contributing to the development and maintenance of delusions, hence explaining and promoting the implementation of successful interventions; and b) the ethico-political desideratum, according to which doxasticism yields a way to render delusional phenomena intelligible and hence provides a further barrier protection against potential deagentializing practices and the concomitant risk of unjust treatment.

In Chapter 6 we'll examine whether doxasticist proposals can live up to their own desiderata. Firstly, we'll focus on revisionist doxasticisms, considering whether they can meet the scientific desideratum (section 6.1.). We'll review its proposed revision of the notion of "belief" (section 6.1.1.), which is partially based on the addendum of a *ceteris paribus* clause to interpretivism and functionalism; in particular, they contend that delusions can be characterized as beliefs because the deviations from the stereotypical causal or rational profile

of belief can be excused by non-standard features of the context of ascription. We'll then introduce an objection to this strategy (section 6.1.2.), which will lead us to the conclusion that what revisionist doxasticists need for their defense to work is to embrace a strong account of the context-relativity of the truth value of belief ascriptions (section 6.1.3.). In doing so, revisionist doxasticists might preserve the doxastic conception of delusions, but the resulting doxasticism would be of little use for scientific purposes. We'll then turn to non-revisionist doxasticism (section 6.2.). We'll first review its underlying cognitive-phenomenological theory of belief, which was precisely designed to offer defense of scientific doxasticism, or the conception of belief at play in cognitivist accounts of delusions (section 6.2.1.). After laying out its dual descriptivist commitments, we'll conclude that, no matter how it's construed, this approach is essentially unable to meet the ethico-political desideratum (sections 6.2.2. and 6.2.3.). Finally, we'll turn back to the question of the doxastic state of delusions, and consider the main implications of our non-descriptivist approach to this debate (section 6.3.). We'll claim that our non-descriptivist approach allows us to see why doxasticism is in better shape than its competitors to account for how our folk-psychological interpretative practices in fact work (section 6.3.1.), and how exactly this provides a further barrier against undue deagentializing practices, hence yielding a preferrable conceptualization of delusions in ethico-political terms (section 6.3.2.). In addition, we'll show how non-descriptivism allows for a different and more robust defense of doxasticism, which is able to avoid the pitfalls of possible eliminativist counterarguments (section 6.3.3.).

In Chapter 7 we'll analyze whether scientific doxasticism, or the conceptualization of delusions at play in cognitive models of delusions, does really improve our understanding of delusions in scientific terms. To do so, we'll first introduce the main tenets of scientific doxasticism. In particular, we'll review two main approaches: the cognitive behavioral therapy for psychosis (CBTp) and the cognitive neuropsychiatric understanding of delusions (section 7.1.). We'll see that these approaches are characterized by a shared understanding of the hypothetical cognitive factors allegedly at play in the development and maintenance of delusions. These include a series of cognitive biases (e.g., jump-to-conclusions, externalizing attributional style, and flawed Theory of Mind), which hypothetically arise from and in turn feed back into underlying maladaptive cognitive schemas (e.g., negative self-schemas). We'll then review the evidence on the efficacy of CBTp interventions with people with delusions (section 7.2.). As we'll see, the evidence gathered so far is ambiguous with regard to the efficacy of CBTp on delusions; furthermore, it doesn't seem like its effect on delusional phenomena is explained by the introduction of changes in the hypothesized cognitive mechanisms. After that, we'll point out that the explanatory problems of scientific doxasticism

might be partly due to its allegiance to a flawed Cartesian conception of mental states and what it means to act "in accordance with a certain belief" (section 7.3.). In particular, we'll claim that it's the commitment to an intellectualist view of the mind which unfoundedly curtails the range of potential explanations available and the therapeutic methods that may be employed. By forcing research and intervention to look for hypothetical internal mechanisms, this intellectualist commitment diverts attention from the environmental sources of behavioral control. We'll conclude that, despite we agree with some cognitivist thinkers that psychological interventions should be based on an individually-tailored, formulation-based, and causal-interventionist approach to mental health problems, a better example of such an approach might be provided by non-cognitivist models.

Chapter 8 we'll review such non-cognitivist approaches to delusions, specifically focusing on functional analytic models. After reviewing some common characteristics of the functional analytic conceptualization of mental health problems (section 8.1.), we'll distinguish two main strands: "traditional" behavior analysis, which draws from a seemingly more "orthodox" understanding of radical behaviorism, and Acceptance and Commitment Therapy (ACT), which draws from functional contextualism and its core post-Skinnerian approach to human language and cognition, Relational Frame Theory (RFT). We'll then review the main characteristics of their conceptualization of delusional phenomena as well as the evidence supporting their efficacy in the intervention with people with delusions. On the one hand, traditional behavior analytic approaches have primarily conceptualized delusions as non-normative verbal behaviors, proving to be highly efficacious in their reduction (section 8.2.). On the other hand, ACT conceptualizes delusions as inflexible verbal rules primarily maintained by their avoidance function, and emphasizes the need to shift focus from "problem reduction" models of recovery to the enhancement of the strategies used by the person to cope with their experiences. However, even when measured by its own standards, the evidence of their efficacy in the case of delusions is far from conclusive (section 8.3.). Finally, we'll comment on how our non-descriptivist approach to mental-state ascriptions may offer some ways to improve the efficacy and perceived utility of functional analytic approaches to delusions (section 8.4.). After reviewing the typical responses by functional analytic researchers to some objections traditionally raised by competing approaches, we'll claim that these responses contain a somewhat residual commitment to intellectualism, which might explain some of the problems of functional analytic approaches (section 8.4.1.). In particular, we'll see how it might lead to: a) the identification of delusions with exclusively verbal behaviors, which hinders the perceived utility of traditional behavior analytic approaches (section 8.4.2.); or b) an excessive focus on potential verbal sources of behavioral control, which

might hinder the efficacy of ACT interventions with people with delusions (section 8.4.3.). Instead, our non-descriptivist approach stresses the distinction between norm-following behavior (i.e., acting in accordance with certain norms) and rule-governed behavior (behavior which is the causal product of verbal rules) (section 8.4.4.). In doing so, it encourages the adoption of values-based strategies for the determination of intervention goals, a focus shift from verbal behavior to overall patterns of behavior (whether verbal or non-verbal, covert or overt, etc.), and the use of Functional-Behavioral Assessment methods to determine, on a case-by-case basis, the specific causes of target behaviors and how best to intervene on them.

Finally, in Chapter 9, we'll draw the main conclusions of this dissertation. Firstly, we'll provide a summary of the present work, highlighting its main contributions (section 9.1.). After that, we'll discuss several possible lines for future research (section 9.2.). In particular, we'll sketch out some further reflections on how our non-descriptivist approach could be extended to the analysis of the four major themes of the philosophy of mental health, i.e., the analogy, boundary, priority, and integration problems (sections 9.2.1. and 9.2.2.).

PART I
Non-descriptivism and the philosophy of mental health

Chapter 1
A medicine of mind?

The history of mental health research and practice is an ongoing history of fierce tensions among different therapeutic models, acrimonious debates spreading far beyond the realm of clinical practice, grave accusations wielded back and forth among contending positions; a history of conceptual confusion, gloomy methods, devasting secondary effects, despicable abuses; a history of big words and silencing practices; and a history of sanism, classism, racism, sexism, and queerphobia. Yet it's also an ongoing history of creative syntheses among opposing parties, fruitful interdisciplinary partnerships, mutual recognition and therapeutic alliances; a history of conceptual clarification, user-led research, life quality improvements; a history of survival, de-stigmatization, re-signification, re-dignification; and a history of diversity, mad pride, and mad identities. It's a dense, complex, and heated history, and a full-fledged account of it is obviously beyond the scope of this dissertation. Instead, we just aim at providing an outline of a specific part of that history: namely, that related to the most prominent therapeutic models proposed and discussed from the second half of the 20th century onwards.

Critically, we won't assess these therapeutic models in terms of the efficacy of their related treatments -namely, because many of them are attempts to conceptualize the *same* treatment procedures from different theoretical standpoints, and some of them don't have yielded specific procedures of their own. Rather, we'll focus on the discussion of their conceptual frameworks. In particular, in this chapter we'll see that the main differences among them are expressed in the way they address the four major themes of the philosophy of mental health: a) the *analogy problem*, or the problem of the analogy between somatic and mental health problems; b) the *boundary problem*, or the problem of the distinction between "bad" (i.e., social deviancy) and "mad" (i.e., psychopathology proper); c) the *priority problem*, related

to whether some scale of analysis must enjoy causal or constitutive priority over others regarding the conceptualization or treatment of mental health problems; and d) the *integration problem*, or the problem of the conceptual gaps between causal-explanatory approaches drawing from different scales of analysis.

Our main goal will thus be to introduce the different therapeutic models in relation to these four problems. We'll begin in section 1.1. with an exposition of the *medical model*, the hegemonic approach to mental health research and practice since at least the fall of psychoanalysis and the rise of modern taxonomies. We'll distinguish two different interpretations of this model: a minimal interpretation, just committed to the description of psychological problems in medical terms, and a strong interpretation, which makes stronger assumptions about their underlying nature. While the former is almost universally shared, the latter characterizes the much-disputed *biomedical* model, which we'll identity with second-wave biological psychiatry.

In the remaining sections, we'll see how the conceptual history of mental health can be understood in terms of the main historical lines of criticism against this model, which focus particularly on two different sources of attack: one related to the analogy and boundary problems, the other one related to the priority and integration problems.

In section 1.2., we'll address several criticisms of the first kind, putting a special emphasis on Szasz's critical approach. As we'll see, Szasz's attacks on the "myths" of institutional psychiatry clearly reveal the core conceptual tensions behind the analogy and boundary problems.

In section 1.3., we'll review the main psychological models of mental health, whose criticisms revolve around the priority problem. We'll focus on "first-wave" and "second-wave" behavior therapies, that is: a) behavior therapy and applied behavior analysis, which can be distinguished by their respective roots in methodological and radical behaviorism; and b) cognitive behavioral therapy, resulting from the merger of behavior therapy with cognitive therapy.

In section 1.4., we'll explain how the biopsychosocial model emerged in an attempt to provide a synthetic solution to psychiatry's schisms and the debates around the priority problem. We'll see how it came to be the "mainstream ideology" for the next decades, and why, despite its popularity, it didn't succeed in establishing a proper multi-level approach to mental health. The problem, as we'll see, lies in its practical and theoretical eclecticism, which yields an unfitting integrative framework.

In section 1.5., we'll review three contemporary alternative approaches to mental health theory and practice, which constitute an extension of the medical, psychological, and

integrative models of previous decades: a) third-wave biological psychiatry, which draws from a precision medicine approach to mental health; b) contemporary functional analytic approaches, which, despite their differences, share the functionalist and selectionist perspective of their antecessors; and c) the enactive approach to psychiatry, which derives from the recent application of post-cognitivism to the field of clinical practice.

At the end of this section, we'll argue that the four major topics of disagreement among therapeutic models (i.e., the analogy, boundary, priority, and integration problems) are the specific expressions in the realm of mental health of two core conceptual problems: a) the problem of mind, which consists of a series of inter-related issues regarding the relationship between mind and body, mind and world, and mind and language; and b) the *problem of normativity,* or the problem of the place of values and norms in a naturalist understanding of the world. These two problems will constitute the primary object of discussion in upcoming chapters.

Finally, in section 1.6. we'll summarize the main contents of this chapter.

1.1. The medical model

We've begun the present chapter by stating that the medical model of mental health problems is the mainstream, hegemonic understanding of psychological suffering. This claim, we think, would be widely shared. But, if that's the case, this seems to be at odds with what many practitioners report as their actual theoretical background -typically, the biopsychosocial model (see Craddock et al., 2008; Ghaemi, 2010; Shah & Mountain, 2007); a mismatch for which there's even some empirical evidence (Colombo et al., 2003; Fulford & Colombo, 2004; see also Fulford & van Staden, 2013, p. 393-394). This points to the existence of a gap between what many practitioners explicitly endorse as their framework models and those that they seem to implicitly follow in actual research and clinical practice -or, as we'll put it in Chapter 4 (section 4.2.2.), between the therapeutic models they *say* they endorse and those that they *express* in action. Maybe this mismatch is merely due to lack of communication among different practitioners. Maybe it reflects the structural constraints and limitations that many practitioners face in actual practice, with medical units often in charge of the management of mental health care resources. Or maybe it's due to lack of clarity regarding the basic commitments of what we understand as the "medical model". Probably it's a combination of factors; however, here we'll focus on the last one.

In this sense, Murphy (2006, 2008, 2009, 2013, 2020) establishes a useful distinction between the *minimal interpretation* and the *strong interpretation* of the medical model. According to the minimal interpretation, the medical model of mental health problems just

entails their consideration *qua* medical entities. Specifically, it considers them as *syndromes*, i.e., as constellations of signs and symptoms that presumably tend to co-occur and which have some predictive value. Thus defined, mental *disorders* are merely conceived of as *diagnostic kinds* (Tabb, 2017), i.e., as medical categories with certain utilities (e.g., predictive, organizational, etc.). In Murphy's (2013) words, "the minimal interpretation is therefore recognizably medical in terms of the information it collects, the concepts it employs, and the practices it supports, but it makes few, if any, commitments about what is really going on with the patient" (p. 967). By contrast, the strong interpretation of the medical model implies a deeper commitment regarding the nature and causes of the observed syndromes. Specifically, it conceptualizes them as essentially identical to a series of underlying pathophysiological processes. Its core theoretical commitment is that mental disorders are *natural kinds* (i.e., categories which presumably "carve nature at its joints"): in particular, they pick out underlying neurobiological anomalies (e.g., functional or anatomical alterations in the brain). Thus understood, the strong interpretation of the medical model would correspond to what is usually called the *biomedical* model of mental disorders. The two construals of the medical model are deeply entrenched in the history of mental health theory and practice, and have been a source of ongoing debate as of yet.

Despite common accusations to the contrary, traditional nosological tools like *Diagnostic and Statistical Manual of Mental Disorders* (DSM) or the *International Classification of Diseases* (ICD) endorse the minimal, not the strong interpretation. There are several reasons for this. Firstly, traditional diagnostic tools were initially designed to satisfy primarily administrative purposes: the main motivation behind the creation of these diagnostic manuals was to address diverse practical necessities, such as conducting epidemiological studies, facilitating communication among practitioners from different theoretical backgrounds, providing an assessment tool for diverse insurance-related matters, etc. (see Kupfer et al., 2002; Leoni, 2013; Tabb, 2017, 2020). These still are important functions of mental health nosologies. With the advent of evidence-based research, they also provided a common language to establish comparisons among clinical procedures (APA Presidential Task Force on Evidence-Based Practice, 2006; APA Task Force on Promotion and Dissemination of Psychological Procedures, 1995).

Secondly, the minimal interpretation seems to have been favored for two main theoretical reasons: a) the lack of consensus regarding the appropriate definition of basic medical concept such as "disorder", "illness", or "disease" (e.g., Boorse, 1975; Spitzer et al., 1978; Szasz, 1961/1974; see also see Fulford & van Staden, 2013; Kupfer et al., 2002); and b) the lack of consensus regarding the appropriate theoretical framework or "school of thought" from

which mental health problems should be conceptualized (e.g., Kendell, 1975; Eysenck, 1959, 1960, 1963; see also Matthews, 2013; Kendler 2005). Historically, these concerns were especially acute during the period comprised between the development of the DSM-II (American Psychiatric Association, 1968) and the years following the development of the DSM-III (APA, 1980) and its revised version (DSM-III-R; APA, 1987). Let's see this in more detail.

Up until the 1960's, the mainstream model of psychiatric assessment and treatment was based on psychoanalytic assumptions and theories. However, due to a complex mix of scientific and social factors, psychoanalysis went through a relatively rapid decline. Among these factors, we may highlight the beginning of the wide circulation of psychiatric drugs, the boost of research on the genetic bases of mental disorders, the growing concern of institutional psychiatry regarding its own status as a branch of medicine, or the increasing need to standardize the provision of mental health resources after the Second World War (González-Pardo & Pérez-Álvarez, 2007; Kupfer et al., 2002; Leoni, 2013; Tabb, 2017, 2020). The DSM-II, developed in the midst of this ebullient context, was largely criticized at the time for its strong psychoanalytical imprint, and its related lack of validity and reliability (Eysenck, 1964; Haslam, 2013; Kendell, 2004; Spitzer et al., 1978).

It was against this social and scientific background that the APA appointed the DSM-III Task Force to develop a subsequent version of the diagnostic manual which properly addressed the above-mentioned concerns. The development of the DSM-III was strongly influenced by the logical-empiricist philosophical framework of Hempel's *operationalist* approach to taxonomy (Hempel, 1965; see also Haslam, 2013; Tabb, 2017; Thornton, 2013; Fulford et al., 2013b). In 1959, Hempel read a paper on the taxonomy of mental disorders at the APA's Work Conference on Field Studies in the Mental Disorders. His operationalism essentially came down to two methodological recommendations for the development of a sound psychiatric taxonomy: a) firstly, to define the taxonomic categories in clear operational terms, so that the system's "intersubjective uniformity" (i.e., inter-rater reliability; see Tabb, 2017) could be properly assessed; and b) once a sufficiently reliable taxonomic system was achieved, psychiatry should aim at establishing scientific laws and theories regarding the etiological processes behind each mental disorder. Hempel's operationalism later inspired the neo-Kraepelinian approach to diagnosis advocated by many of the leading figures of institutional psychiatry at the moment (Klerman, 1978; Spitzer et al., 1978, 1978/2018; see also Kupfer et al., 2002; Tabb, 2017; Haslam, 2013). These authors retrieved Kraepelin's syndrome-based approach to research on psychiatric and neurological disorders, which was similar to Hempel's approach to taxonomy; according to Murphy (2009), "Kraepelin saw classification by clinical description as an interim measure designed to satisfy the practical requirements

of contemporary physicians [and which] could also provide a fruitful heuristic for subsequent pathological and aetiological inquiry" (p. 110). Hempel's operationalism and Kraepelin's syndrome-based approach configured the minimal interpretation of the medical model characteristic of neo-Kraepelianism (Murphy, 2008, 2013), which has prevailed in institutional psychiatry until at least the last decade.

Drawing from these premises, Feighner et al. (1972) elaborated what was later known as the "Feighner criteria" for the diagnosis of mental disorders, which were subsequently developed by Spitzer et al. (1978) in their "Research Diagnostic Criteria". These, in turn, laid the foundations of the DSM-III (APA, 1980), which is commonly described as a turning point and a first milestone in the development of a properly scientific mental health research and practice. Following its Hempelian and neo-Kraepelinian conceptual bases, the DSM-III was primarily aimed at securing proper levels of inter-rater reliability; emphasis was made on providing reliable descriptions of psychological symptoms, rather than on attempting to classify disorders in terms of their presumed etiological basis. In this sense, it "sacrificed validity for reliability" (Tabb, 2015, p. 1047) and officially adopted an uncommitted view of the exact nature of mental disorders (e.g., see Spitzer et al.'s, 1978/2018 definition of "mental disorder" and "medical disorder"). This uncommitted stance was well-reflected in the DSM-III through the provision that "there is no assumption that each mental disorder is a discrete entity with sharp boundaries (discontinuity) between it and other mental disorders, as well as between it and No Mental Disorder" (APA, 1980, p. 6; see also APA 1987, p., xxii), subsequently modified to clarify that mental disorders weren't assumed to be "*completely* discrete [entities] with *absolute* boundaries" (APA, 1994, p., xxii, emphasis added; see also APA, 2000, p. xxxi). The definition of "mental disorder" also reveals this non-committal attitude. In the revised version of the third edition, it states:

> In DSM-III-R each of the mental disorders is conceptualized as a clinically significant behavioral or psychological syndrome or pattern that occurs in a person and that is associated with present distress (a painful symptom) or disability (impairment in one or more important areas of functioning) or with a significantly increased risk of suffering death, pain, disability, or an important loss of freedom. In addition, this syndrome or pattern must not be merely an expectable response to a particular event, e.g., the death of a loved one. *Whatever its original cause,* it must currently be considered a manifestation of a *behavioral, psychological, or biological dysfunction in the person*. Neither deviant behavior [...] nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the person, as described above. (APA, 1987, p. xxii)

This minimalist approach to the definition of mental disorders has been the official stance of important mental health institutions in the Western world, such as the National Institute of Mental Health (NIMH) and the American Psychiatric Association (APA). The DSM-IV and DSM-IV-TR contained almost identical definitions, just adding that mental disorders must not be "an expectable *and culturally sanctioned response* to a particular event" (APA, 1994, p. xxi; APA, 2000, p. xxxi; emphasis added). Finally, the DSM-5 made a first (although somewhat timid) step in the direction of nosological validation, by univocally defining mental disorder as "a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior *that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning*" (APA, 2013, p. 20, emphasis added). It also included a reference to research on several kinds of validators and even admitted that current diagnostic categories won't necessarily map well onto such validators (APA, 2013, p. 20); however, it nonetheless stated that:

> Until incontrovertible etiological or pathophysiological mechanisms are identified to fully validate specific disorders or disorder spectra, the most important standard for the DSM-5 disorder criteria will be their clinical utility for the assessment of clinical course and treatment response of individuals grouped by a given set of diagnostic criteria. (APA, 2013, p. 20)

These minimalist assumptions contrast sharply with the main guiding assumptions of a great deal of psychiatric practice and research back then. The truth is that, already at the time of the publication of the DSM-III, a majority of psychiatric researchers and practitioners were *de facto* committed to a stronger or *biomedical* conception of mental health problems. According to this strong interpretation of the medical model, mental disorders the result of specific neurobiological alterations (Murphy, 2006, 2008, 2009, 2013, 2020). The main historical root of the biomedical conception of mental disorders lies in Griesinger (1817–1868), who famously stated that mental disorders were essentially disorders of the brain (see Kendell, 2004, p. 3; Walter, 2013, p. 1). Following Shorter (1998), Walter (2013) takes Griesinger's work as foundational to the "first wave of biological psychiatry", which took place during the second half of the nineteenth century and where the grounds of the biomedical conception of psychological problems were first established.

More important to our present discussion was the "second wave of biological psychiatry", which took began in the aftermath of the Second World War, fully flourished during the 1970's, and spanned, at least, until the last decades of the twentieth century. Second-wave biological psychiatry can be identified with what we might call the *classical* biomedical

model of psychological problems. In a nutshell, its basic assumptions are the following (see Kandel, 2005; Kupfer et al. 2002; Murphy, 2009, 2013, 2020; Shorter, 1998; Walter, 2013):

a) Mental disorders are natural kinds; they pick out underlying biological (typically, genetic and neurobiological) abnormalities (e.g., alterations in neurotransmission circuits).
b) Mental health research should be mainly directed at the discovery of the specific biomarkers of each mental disorder (i.e., the specific genetic and neurobiological processes that explain its characteristic symptomatic picture).
c) Pharmacological treatments directly target the biological "root of psychopathology"; by contrast, the effectiveness of psychosocial interventions is mainly explained through their mediated effect on the (neuro)biological substrate.

Two major forces drove its emergence during the 1950's: a) research on the genetic basis of what are known as "Severe Mental Disorders" (e.g., schizophrenia); and b) the development of the first psychiatric drugs (Kallmann, 1946; Karasu, 1982; Kety et al. 1968, 1971; Lehmann & Hanrahan, 1954; van Praag, 1972; see also Kupfer et al., 2002; Shorter, 1998; Walter, 2013). These two research fields constituted the primary sources of evidence for the subsequent "neurobiological hypotheses" of mental disorders, e.g., the dopamine hypothesis of schizophrenia (see Brisch et al. 2014; Kendler & Schaffner, 2011). Critically, as Kupfer et al. (2002, p. 32) note, "these hypotheses were largely derived post hoc from discoveries related to the pharmacologic actions of antidepressant and antipsychotic drug treatments". In other words: these "first-generation" anti-psychotics and anti-depressants were not designed after having previously detected specific brain pathologies underlying each mental disorder; rather, it was the discovery that certain drugs produced an amelioration of the symptoms of certain mental disorders which provided inductive support for the biomedical assumption that mental disorders were mainly due to specific neurobiological alterations (Deacon, 2013; Deacon & Beard, 2009; González-Pardo & Pérez-Álvarez, 2007; Kendler & Schaffner, 2011; Kupfer et al., 2002; Moncrieff, 2015a, 2015b; Pérez-Álvarez & García-Montes, 2007; Whitaker, 2015). As Moncrieff (2015b) has pointed out, instead of a "drug-centered model" of the effect of psychiatric drugs (i.e., one according to which psychiatric drugs, like other drugs, exert global effects on the person) second-wave biological psychiatry assumed a "disease-centered model of drug action" -or "magic bullet model", as Whitaker (2011) puts it-, according to which psychiatric drugs exert their effects on abnormal brain states and processes, the symptomatic relief associated to drug intake is explained by its action on such target

processes, and where "therapeutic effects" can be clearly distinguished from "side effects" (Moncrieff, 2015, p. 214; see also González-Pardo & Pérez-Álvarez, 2007; Shorter, 1998; Whitaker, 2011).

These neurobiological hypotheses have largely framed research on the validity of traditional nosological tools. What is more, some of them have deeply permeated the folk understanding of psychological problems, as it's the case with the dopamine hypothesis of schizophrenia or the serotonin hypothesis of depression (González-Pardo & Pérez-Álvarez, 2007). Even official organisms, which explicitly commit to a minimalist conception of the medical model, have progressively endorsed stronger biomedical assumptions; hence the claim in the DSM-IV that "the term *mental disorder* unfortunately implies a distinction between 'mental' and 'physical' disorders that is a reductionistic anachronism of mind/body dualism" see APA, 1994, p. xxi) or the increasing appeals to research on nosological validators in the DSM-5 (APA, 2013). It remains to be seen what the imminent publication of the DSM-5-TR will bring in this regard. In any case, despite the official allegiance to the minimal interpretation of the medical model, it seems that mental health research, practice, and public discourse have been largely framed by biomedical assumptions.

All in all, both versions of the medical model have been ongoingly and relentlessly criticized -often in absence of a proper distinction between its possible interpretations. From the critical and psychological models of the 1960's (see sections 1.2. and 1.3.) and Engel's biopsychosocial model in the late 1970's and 1980's (see section 1.4.), to the most recent contemporary approaches to psychiatry (including what has been called the "third wave of biological psychiatry"; see section 1.5.), all these distinct and even antagonistic approaches to mental health practice have found some common ground in their analyses of the main pitfalls of the medical model. The main concerns regard the following issues: a) the conceptual validity of the very concept of "mental disorder", usually related to concerns about its "mythical" or "invented" character, as well as to the questioning of the analogy between mental and somatic disorders (Pérez-Álvarez & García-Montes, 2007; Szasz, 1961/1974; see also Thornton, 2007); b) the lack of reliability of traditional nosological tools, with comorbidity and diagnostic instability being the norm rather than the exception (Cooper, 2014; Deacon & McKay, 2015; Keshavan et al., 2011, 2013; Markova & Berrios, 2012; Tandon, 2013); c) their lack of validity, related to the long-standing yet fruitless search of specific biomarkers for each mental disorder (Borsboom et al., 2019; Deacon, 2013; Deacon & McKay, 2015; Kendler & Schaffner, 2011; Keshavan et al., 2011; Kupfer et al., 2002; Lacasse & Leo, 2015; Peele, 2015); and d) the long-term inefficacy and secondary effects of pharmacological treatment, among which the increased risk of chronicity is of special importance (Deacon, 2013; Deacon &

McKay, 2015; El-Mallakh et al., 2011; Lader et al., 2009; Moncrieff, 2015a, 2015b; Rummel-Kluge et al., 2010 Qaseem et al., 2016; Whitaker, 2011). In addition, the medical model has also raised ethical and political concerns related to a variety of topics, such as the risk of over-medicalization due to overdiagnosis, the pathologizing and medicalization of non-normative forms of life, the invasiveness of certain treatment procedures, the threat to autonomy posed by involuntary commitment, derived forms of personal and social damage (e.g., epistemic injustice), etc. (see Bueter, 2019; Carel & Kidd, 2014; Crichton et al., 2017; Deacon & Baird, 2009; Deacon & McKay, 2015; Kvaale et al., 2013; Moncrieff, 2015a, 2015b; Pescosolido et al., 2010; Schomerus et al., 2012; Whitaker, 2011).

A fundamental question behind many of these criticisms is the following: can psychiatry (or any other clinical discipline) be meaningfully conceived of as a "medicine of mind"? In the next section we'll see what are the historical roots of this central question, and how it has inspired harsh criticisms among contending positions since, at least, the second half of the twentieth century.

1.2. Critical mental health: the "myths" of the medical model

Undoubtedly, an essential part of the history of psychiatry is the history of its critics, which have typically come to be grouped under the hazy label of "anti-psychiatry" -a term originally coined by Cooper (1967). We've decided to avoid this way of referring to them for a number of reasons, namely, because: a) it obscures the wide heterogeneity of schools of thought and disciplines from which such criticisms have been historically raised; b) it conveys a view of psychiatry and its opponents according to which psychiatric research and practice can only be properly understood under a unified model (presumably, the biomedical one), and where criticizing such model automatically confers one's approach an "anti-" or "outlier" status; c) the term itself was rejected by some of the most prominent critics of psychiatry (e.g., Szasz, 2009); and d) some of the criticisms that have been historically raised against psychiatry (e.g., the criticism against the medical understanding of psychological problems) also apply to other clinical disciplines, such as clinical psychology, where the medical model is also widely -although less consensually- accepted (see González-Pardo & Pérez-Álvarez, 2007). Instead, we've decided to use the term "critical mental health" (see Cohen, 2018) to refer to the multiple critical approaches to mental health theory and practice, since it avoids the above-mentioned problems and, in addition, it is more inclusive than "critical psychiatry" (see Middleton & Moncrieff, 2019).

The founding works of the diverse lines of criticism against institutional psychiatry simultaneously appeared during the 1960's. In 1961, two groundbreaking works appeared

that would shake the conceptual and historical foundations of modern psychiatry. On the one hand, Szasz's (1961/1974) *The Myth of Mental Illness* shook its most cherished theoretical assumptions: by attacking the logic behind the conceptualization of psychological suffering in medical terms and describing involuntary treatments as deprivations of liberty and even "crimes against humanity" (p. 268), Szasz's most famous work intended to be a hammer blow against psychiatry's self-perceived status as a scientific enterprise and an actual branch of medicine. On the other hand, Foucault's (1961) *Histoire de la folie* (later translated as *Madness and Civilization: A History of Insanity in the Age of Reason*) radically subverted the deceptive and self-indulgent common tale of the historical origins of modern psychiatry. On his view, mental health care came to be viewed as the exercise of often hidden and asymmetrical power relations which, in his view, were typically masked under the veil of humanitarianism.

A year before, Laing (1960/2010) had published *The Divided Self*; which questioned the alleged incomprehensibility of schizophrenic phenomena and vindicated the existential-phenomenological approach to the meaningful analysis or "empathic understanding", to use Jasper's (1913/1963) term, of psychotic experiences. According to Laing, the view of psychotic problems as incomprehensible or unintelligible phenomena was essentially grounded in the individualistic focus of modern psychiatry; instead, he proposed a framework for analyzing and uncovering the ultimately contextual origin of these experiences, which rendered them as intelligible or rational responses to distressful and alienating social settings.

Somehow in line with Laing's emphasis on the social and contextual dimension of psychopathology, although drawing on a radically different theoretical framework, many behavioral psychologists also raised criticisms against psychiatry and the medical model during the 1950's and the 1960's. Eysenck's (1960), *Handbook of Abnormal Psychology* (cited in Kendell, 1975; Fulford & van Staden, 2013) was viewed as the major psychologically-oriented attack to psychiatry and the medical model at the time, but other works by Skinner (1953, 1957) also proved to be highly influential. Contrary to Laing's work and the existential-phenomenological tradition, these authors' emphasis on the environmental aspect of psychological problems wasn't aimed at *understanding* them in intelligible terms; rather, they aimed to *explain* them in causal terms. Specifically, they emphasized the need to focus on environmental sources of psychological distress, and defended the therapeutic efficacy of behavior modification methods.

Psychiatry also came under attack from sociology during the 1960's. Drawing from previous work by Goffman (1963), Scheff (1966/1999) published *Being Mentally Ill* in 1966, where he laid the foundations for one of the most influential critical approaches to psychiatry: labeling theory (see also Scheff, 1974; Link et al., 1989) which explored not only the

conceptual relations between the notions of "mental disorder" and that of "social deviance", but also the possible *causal* relations between them.

Ever since these groundbreaking works appeared, both psychiatry's status as a branch of medicine and the medical model of psychological problems have been the object of harsh criticism. In this and the following sections, we'll mainly address those that are most tightly related with the four major themes of the philosophy of mental health: the analogy, boundary, priority, and integration problems. In particular, we'll focus on discussions around these issues during the first decades of the second half of the 20th century. In this section we'll discuss the analogy and boundary problems, and we'll leave the discussion of the priority and integration problems for sections 1.3. and 1.4.

To begin with, the analogy problem, or the problem of the conceptual relation between *mental* and *somatic* health problems, has been one of the most pressing issues for any approach that conceptualizes psychiatry as a branch of medicine. A great deal of psychiatric research, practice, and public discourse (including anti-stigma campaigns; see Pescosolido et al. 2010) takes the analogy for granted: on this view, mental health problems are, ultimately, just like somatic health problems. This typically justifies the ascription of the sick role to those diagnosed with mental health problems, with the subsequent diminishing of the perceived responsibility (and autonomy) of the person regarding the experiences, thoughts, and behaviors that are viewed as the effect of the underlying condition.

The most famous and widely discussed attempt to problematize this analogy is Szasz's relentless and uncompromising attack on *the myth of mental illness* (Szasz, 1960, 1961, 1961/1974), which spanned over more than 50 years (Szasz, 2011). For Szasz, the use of medical terms like "illness" in relation to the mental amounted to a perverse "metaphor" (Szasz, 1961/1974, p. ix), and the medical conception of psychological suffering was, essentially, *mythical*. He explicitly equated the explanation of problematic behavior in terms of "mental illnesses" to the explanation of moral and social deviance in terms of "witchcraft" during the Middle Ages (Szasz, 1960, p. 1; Szasz, 1961/1974, p. 182); for Szasz, such explanations were merely fictional devices that obscured the proper understanding of psychological (or social) problems, and that ultimately subserved the purpose of psychiatrists' professional self-aggrandizement and legitimation.

Szasz's core argument against the notion of mental illness draws from a particular understanding of the concept of "illness" itself, which first appeared in his 1960 paper and which he subscribed to during his whole career (e.g., Szasz, 2011). According to Szasz (1960, p. 114):

> The concept of illness, whether bodily or mental, implies *deviation from some clearly defined norm.* In the case of physical illness, the norm is the structural and functional integrity of the human body. Thus, although the desirability of physical health, as such, is an ethical value, what health *is* can be stated in anatomical and physiological terms. What is the norm deviation from which is regarded as mental illness? This question cannot be easily answered. But whatever this norm might be, we can be certain of only one thing: namely, that it is a norm that must be stated in terms of psychosocial, ethical, and legal concepts. (Szasz, 1960, p. 114)

Specifically, the concept of illness that Szasz had in mind was first stipulated by the German pathologist Rudolf Virchow (see Szasz, 2003; see also his replies in Schaler, 2004), who identified illnesses not with the conflation of different signs and symptoms, but with their actual pathophysiological basis; illness, or disease, was thus defined as the deviation from some given pattern of anatomical or functional integrity of the body. For Szasz, this "purely materialist-scientific notion of disease" (2003, p. 12) provided the criterion on which to decide whether something is an illness or not; by contrast, as he pointed out, psychiatric diagnoses were essentially dependent on psychosocial and legal norms that regulate the distinction between appropriate behavior and misconduct.

In this sense, the assimilation of psychological suffering within the realm of medical language was for Szasz just a self-enhancing masquerade that hid the essentially psychosocial, not medical, nature of psychiatry and psychological problems and disguised what he saw as illegitimate coercive practices as "psychiatric care". For him, the origin of this illegitimate conceptual change was to be found in Charcot's formulation of hysteria as a medical condition, which was linked to the distinction between somatic or organic illnesses and mental or functional illnesses. This was accompanied with a distinction between two kinds of medical enterprise: while "organic" medicine (e.g., neurology) dealt with the pathological processes that caused the observable symptoms, the main subject matter of "mental" medicine was the "ill" behavior itself. For Szasz, this distinction was based on a fallacy, since "*the behavior*, *per se*, *cannot, as a matter of definition, be a disease*" (Szasz, 2003, p. 12), and it revealed the pseudoscientific and "quack" character of psychiatry: while illnesses were amenable to scientific *discovery* and *testing* through the analysis of the objective anatomy and physiology of cells, tissues, and organs, so-called "mental illnesses" were just *invented* or *declared* as such (Szasz, 1961, p. 12; see also González-Pardo & Pérez-Álvarez, 2007). It's important to note here that Szasz didn't deny that research could eventually discover the neurobiological basis of, say, schizophrenia; what he pointed out is that, if such discovery was made, that would only *prove* his point, i.e., that schizophrenia was, after all, a physical illness -the only possible kind of illness for him (Szasz, 1961/1974; see also Schaler, 2004).

Szasz's antagonism to the medical model of psychological problems earned him the enmity and hostility of many of his contemporaries, who were often quick to dismiss his approach as "pseudoscientific"; some of them even actively tried to undermine his professional career (see Schaler, 2004). This is somewhat curious, since his main intention actually was to dispel what he viewed as the major obstacle for psychiatry to become an actual science. In his view, mental health theory and practice shouldn't aspire to become a naturalistic enterprise nor define its subject matter in the same terms as the natural sciences did. For him, the subject matter of psychiatry wasn't "mental illness", but just "personal, social and ethical *problems in living*" (Szasz, 1961, p. 262, emphasis added), i.e., problems encountered by human agents in the course of life, namely due to the complexities of our social world. These problems, in Szasz view, amounted to the contradictions between the *norms* that a person wants to follow and those that the person in fact follows, either due to external constraints or to sheer "lack of will" (see also Chapter 2, section 2.4.).

This view of the subject matter of psychiatry and other clinical disciplines informed what he viewed as the proper theoretical framework for such disciplines. By defining the subject of treatment in *personal* terms, he rejected all the different deterministic or *subpersonal* accounts of human affairs (whether this determinism was spelled out in psychoanalytic, neurobiological or behaviorist terms); for him, the naturalist pretensions of psychoanalysis, biomedicine, and behaviorism were all illegitimate and pseudoscientific attempts to enhance each group's professional status by assimilating their theories to those of the natural sciences. Instead, he viewed problems in living as problems of *persons* (not brains, nor organisms) in finding *meaning* in their *free* and *autonomous* actions. Thus, what psychiatry needed to be an actual scientific enterprise was to reject any determinist theoretical framework, recognize the irreducibly normative and personal character of its subject matter and consequently develop a proper *theory of personal conduct*; one that explained human behavior in personal terms (i.e., in terms of meaning, agency, free will and autonomy). This, in turn, would help practitioners to develop better ways of working with people and helping them out in detecting and solving the potential normative contradictions underlying their psychological suffering.

Szasz's focus on the irreducible *personal* character of mental health problems was shared by Laing's (1960/2010) existential-phenomenological project, which aimed at reconceptualizing mental disorders as problems in sense-making; i.e., as problems in the person's capacity to find meaning in their lived world. However, despite this surface similarity, both projects were diametrically opposed, as were their ideological viewpoints and political agendas. On the one hand, Laing and his followers drew from a Marxist understanding of the

relation between an agent and their social environment, as well as of their capacity to make sense of their lived world, and thus emphasized the role of a person's social and cultural background in shaping an agent's meaningful interactions with the environment (see Laing & Cooper, 1964; see also Fulford et al., 2013b); by contrast, Szasz explicitly loathed this view and its Marxist grounds (Szasz, 1961), described Laing's and Cooper's anti-psychiatry as "quackery squared" (2009), and favored instead an individualistic, ultraliberal view of people as a rational, autonomous, and radically free-willed agents whose problems were mainly due to themselves. For him, the nationalization and alleged "socialist" character of the American and British mental health services constituted no less than a legitimation of the State's control over the individual, and involuntary psychiatric treatment was nothing more than the illegitimate and coercive exercise of psychiatric power via the implementation of "psychiatric slavery" and "psychiatric rape" (Szasz, 2009, p. ix). Instead, his underlying moral and political project was to put agency -understood in steadfast liberalist terms- back in the individual; and, with it, all the responsibility for whatever their problems were. No matter what their environmental constraints might be; ultimately, any attempt to excuse or limit the person's responsibility for their own psychological problems would be to "infantilize" the individual.

We think that anyone with a minimal sense of empathy for how a person's history and social context can affect their experience and behavior would be prone to dismiss Szasz's individualistic views. But despite the respective appropriateness of Szasz's steadfast liberalism or Laing's underlying existential Marxism, the truth is that their emphasis on the personal and psychosocial aspect of psychological problems has deeply influenced debates around the proper subject matter of mental health care; in fact, as we'll see in upcoming chapters, their observation of the intimate conceptual connection between mental and personal language goes straight to the heart of the core philosophical problems in the field of mental health.

Apart from problematizing the analogy between mental and somatic health problems, Szasz's and Laing's observations are also strongly linked with the problematization of the *boundaries* of mental health problems. The boundary problem is central to the scientific aspirations of the medical model: to tell the "mad" apart from the "bad" -that is, to tell psychopathology apart from mere social deviancy- is viewed as crucial to justifying the ascription of the "sick role", with its subsequent amelioration of moral responsibility. In fact, this has been usually considered a historical moral achievement of modern psychiatry; the often-recalled and glorified scene of Pinel entering the Bicêtre Hospital to "free the mad from their chains" typically serves as a symbol of the nineteenth-century triumph of scientific, value-

free, medical reason over the religious superstition and moral condemnation of insanity that allegedly characterized the previous centuries.

Several critical authors, including Szasz (1961) and, especially, Foucault (1961), viewed this as just another foundational myth of modern psychiatry. Foucault (1961) pointed out that the progressive medicalization of insanity, forged throughout the 17th and 18th centuries, never really abolished the deep and intricated relationship between madness and moral or social deviancy, rooted in the dialectics of exclusion/salvation that had before characterized the management of leprosy. The official pathologizing of non-normative sexual orientations until at least 1973 and of dissenting gender identities until at least 2013 is a painfully recent proof that, at least on some occasions, "mad" is just a particular kind of "bad", i.e., that psychiatric diagnoses often mask underlying moral judgements and social prejudices. For Szasz, this masking enabled the exercise of psychiatric coercion on the part of the State to maintain control over deviancy; for Foucault, this monitoring, surveillance, and control functions were woven deeply into the very historical origins of psychiatry (see also Leoni, 2013). This tells us that the disentanglement of "psychopathology" and "social deviancy" is, at least, not a straightforward task. Whether we accept the framing of psychological problems in terms of medical entities or not, we're faced with the problem of setting *normative* boundaries between what counts as morally wrong or morally sanctionable, and what counts as "pathological" or "mad" proper.

Drawing from similar observations, Scheff (1966/1999) articulated his sociological criticism of the medical model and proposed his labeling theory of mental health problems. Crucially, for Scheff, moral condemning attitudes and social prejudices weren't just conceptually inseparable from the concept of psychopathology; in addition, they also played a *causal* role in the development and maintenance of psychopathological behaviors over time. In this sense, his work lies in between the boundary and priority problems. Labelling theory draws from sociological studies on the negative, self-fulfilling effect of stereotypes and stigmatization on socially deviant behavior. The hypothesized process goes as follows: firstly, "mad behavior" and other kinds of normatively disruptive conduct are *labeled* in a certain society due to their undesirability or morally outraging character; in turn, this labelling and the social attitudes accompanying it eventually affect the labeled individual's self-concept, inducing them to conform to the social expectations attached to the label. In this sense, labeling theory placed the problem-label relation upside-down; if for biological psychiatrists the label picked out a series of underlying pathological processes, labeling theorists took this relation the other way around: it was the labeling process itself which would ultimately cause

the stabilization and chronification of mental health problems (see Scheff, 1974; Link et al, 1989; see also Fulford et al., 2013).

Scheff's labelling theory laid the grounds for the analysis of the social or socio-psychological factors influencing the development and maintenance of mental health problems. However, the main attacks on the medical model -at least with regard to the priority problem- have come from the field of psychology. In the next section, we'll delve into the historical origins of psychological models of mental health problems. In we'll focus on the behavioral approach to psychology, which has yielded the most renown and empirically-supported therapeutic procedures to date.

1.3. Psychological models and the priority wars

As we've seen in the previous section, Szasz's theory of personal conduct and Laing's existential-phenomenological approach coincided in emphasizing the psychosocial nature of mental health problems. This emphasis has been repeatedly construed as implying that a proper account of mental health problems should prioritize the psychological and social variables that *causally* explained them, instead of adopting a narrow or exclusive focus on biological factors (e.g., see Engel, 1977). To be sure, this is a straightforward misinterpretation of Szasz's and Laing's objections: neither was trying to replace the causal language of neurobiology by the causal language of psychology or sociology. Instead, both (and especially Szasz) viewed mental health practices as necessarily bounded to personal-level, rationalizing practices, not aimed at the discovery of the biological, psychological, nor social *causes* of mental health problems, but rather at exploring the agent's *reasons* for acting in certain ways and the *meanings* of such actions.

Notwithstanding this misreading, the truth is that many researchers and practitioners have criticized the medical model from this particular angle. Scheff (1966/1999) provides a sociological example of this kind of critical approach. Psychological models of mental health, on the other hand, defend that a proper understanding of psychopathology requires the understanding of the *psychological* processes involved in its production and maintenance. The emergence of these psychological models around the 1950's started the "priority wars", or the disputes around the proper scale of analysis for addressing mental health problems. What "psychological" amounts to, however, is a matter of everlasting dispute. Here we'll focus on those psychological models that are grounded on a *behavioral approach* to psychopathology, which we take to describe a loose set of psychological approaches that, at least in their origins, defended the causal explanation and treatment of psychopathology through the manipulation of basic learning principles, namely: a) classical conditioning (also called

"Pavlovian" or "respondent"), involving the establishment of S-S relations that elicit responses already present in the organism's repertoire to new stimuli; and b) operant conditioning, involving the establishment of (at least) two-term R-S relations whereby the probability of a response varies as a function of its consequences (see Ayllon et al., 1965; Ayllon & Haughton, 1964; Ayllon & Michael, 1959; Eysenck, 1959, 1960, 1964, 1972; Eysenck & Rachman, 1965/2013; Ferster, 1966, 1972, 1973; Ferster & DeMyer, 1962; Kanfer & Saslow, 1965; Lazarus, 1958; Lazarus & Rachman, 1957; Lindsley, 1956, 1962, 1964; Rachman, 1958, 1959; Skinner, 1953; Wolpe, 1952, 1954, 1959).

This unitary presentation of the behavioral approaches to psychopathology nonetheless hides the multiple differences between them, some of which were present from the beginning and are important to understand their subsequent separate developments. For the sake of clarity, here we'll make use of Hayes's (2004) distinction between the first, second, and third "waves" of behavior therapy; in this section we'll focus on the first two, and we'll discuss the third one in section 1.5.2.2.

1.3.1. First-wave behavior therapy: behavior therapy and applied behavior analysis

Following Hayes (2004), here we'll use "first-wave behavior therapy" to refer to the set of behavioral approaches to clinical practice that emerged during the 1950's, that employed methods and procedures derived from the experimental principles of learning theory, and which lasted approximately until the rise of cognitive models during the 1970's. However, despite this nomenclature is useful to analyze the historical evolution of the different behavioral approaches, it still fails to capture important conceptual differences among them. To begin with, we're talking here about the subsequent waves of *behavior therapy*, but this label has more than one meaning, and none of them exhausts what falls under the behavioral model of psychopathology. Some authors treat "behavior therapy" as a synonymous with *behavior modification*[6], and take both to describe in a unitary way what Hayes (2004) refers to as the first-wave of behavior therapy (see also Eysenck, 1972; Kazdin, 1982). Here we'll use "behavior modification", but not "behavior therapy", in that sense; instead, we'll use the latter to refer to a specific subset of approaches within first-wave behavior therapy: those based on the philosophy of *methodological behaviorism,* characteristic of S-R psychology (see Guinther & Dougher, 2013; Dougher & Hayes, 2004; Pérez-Álvarez, 1996). In this sense, behavior therapy was mainly developed in England and South Africa by researchers working under this paradigm (e.g., Eysenck, 1959, 1960, 1964, 1972; Eysenck & Rachman, 1965/2013;

[6] In fact, these two terms were originally used to refer to two separate things, with "behavior therapy" describing the set of techniques derived from classical conditioning and "behavior modification" describing those derived from operant conditioning (Froxán-Parga, 2020). However, we won't make that distinction here.

Lazarus, 1958; Lazarus & Rachman, 1957; Rachman, 1958, 1959; Wolpe, 1952, 1954, 1959; see also Kazdin, 1982).

But behavior therapy was just one of the two coexisting tendencies within first-wave behavior therapy. The other one is *behavior analysis* (see Ayllon et al., 1965; Ayllon & Haughton, 1964; Ayllon & Michael, 1959; Ferster, 1972, 1973; Ferster & DeMyer, 1962; Kanfer & Saslow, 1965; Lindsley, 1956, 1959, 1962, 1964; Skinner, 1953; see also Cooper et al., 2019; Froxán-Parga, 2020; Sturmey, 2020), grounded not on methodological, but on *radical behaviorism* (Skinner, 1945, 1953, 1974, 1977, 1981, 1990; see Baum, 2011; Chiesa, 1994; Moore, 1981, 2001, 2008; see also Dougher & Hayes, 2004; Guinther & Dougher, 2013). Contrary to behavior therapy, this approach was initially developed in America, in close relation to Skinner's operant psychology. Behavior analysis is broadly divided into its "basic" and "applied" branches, i.e., *experimental behavior analysis* and *applied behavior analysis*, where the latter conveys the application of the behavior analytic principles studied by the former to the intervention on a range of socially relevant matters (Cooper et al., 2019; Sturmey, 2020); in particular, here we'll be concerned with the applied behavior analytic approach to psychopathology. Although applied behavior analysis as a distinct branch didn't appear until the 1960's (Rutherford, 2003), the first behavior analytic approaches to human behavior had already begun in the 1950's, typically in controlled settings (e.g., mental health hospitals) (e.g., Lindsley, 1956, 1959); at the risk of being technically anachronistic, here we'll also refer to these first investigations as "applied behavior analysis" (see also Cooper et al., 2019).

To be sure, these two first-wave strands had some important characteristics in common. When behavior modification started to thrive during the 1950's and 1960's, behavior therapists and applied behavior analysts joined forces in their challenge to the prevailing medical-psychoanalytic explanations of psychological problems, which they viewed as either empirically unsound or straightaway "explanatory fictions" (see Ayllon et al., 1965; Eysenck, 1959, 1960; Eysenck & Rachman, 1965/2013; Kanfer & Saslow, 1965; Rachman, 1958, 1959; Skinner, 1953, 1957; Wolpe, 1952, 1959); instead, they vindicated the experimental soundness and therapeutic potential of the recently developed behavior modification techniques (e.g., systematic desensitization, flooding, extinction, differential reinforcement, etc.), based on the experimental paradigms of classical and operant conditioning (see Eysenck, 1959, 1964; see also Cooper et al., 2019; Froxán-Parga, 2020; Kazdin, 1982; Sturmey, 2020). In addition, they jointly defended what is now known as the *continuity thesis*, or the idea that the difference between "pathological" and "non-pathological" forms of behavior is just a matter of degree, i.e., of variation along the quantitative dimensions (e.g., frequency, intensity, duration, etc.) of the behavior of interest. In this sense, behavioral psychologists defended a shift

from "categorical" models of mental health problems (e.g., DSM), to "dimensional" approaches (e.g., Eysenck, 1959, 1964).

All in all, first-wave behavior therapy was unitarily characterized by the defense of behavioral methods of assessment and treatment of mental health problems. Where supporters of the medical model saw disordered minds, first-wave behavior therapists saw "dysfunctional" or "inappropriate" patterns of behavior (Ayllon et al., 1965; Ayllon & Haughton, 1964; Ayllon & Michael, 1959; Eysenck, 1959, 1964; Kanfer & Saslow, 1965; Lindsley, 1956, 1963; Skinner, 1953). Crucially, these adjectives weren't understood as neither pointing to underlying biological dysfunctions nor absolute moral principles; rather, they were understood in contextual terms, i.e., as pointing to the "mismatch" between the person's behavior (or some dimension of it) and their social context. The aims of first-wave behavioral approaches also cut across the traditional "neurotic/psychotic" divide: from "neurotic" conditions (e.g., phobias and other anxiety-related problems) to the psychotic disorders that have traditionally been considered as the incontestable exclusive domain of psychiatry (e.g., schizophrenia), the full range of psychopathology could be understood, explained, and dealt with in terms of the learning processes involved in the origin and maintenance of all kinds of behavior (e.g., Ayllon & Haughton, 1964; Ayllon & Michael, 1959; Eysenck, 1964; Lindsley, 1956, 1963; Wolpe, 1952). Finally, many first-wave behavior therapists understood that *behavioral assessment* was inseparable from *behavioral treatment*: successful interventions needed to draw from a previous analysis of the environmental variables that controlled the individual's behavior; no psychological "magic bullet" would work for every possible case because each problem behavior was to be understood in terms of the particular conditions that produced and maintained it (see Froxán-Parga, 2020).

The philosophical differences between both first-wave strands, however, would eventually blow up this partnership. Behavior therapy, as we've said, was based on methodological behaviorism. The foundation of this form of behaviorism is commonly attributed to Watson (1913, 1919). In this sense, Watson's behaviorist "manifesto" (1913) can be read as advancing two main thesis, respectively related to the proper methods and subject matter of a truly scientific psychology: a) a methodological thesis, namely, that introspection was unreliable and that the only legitimate research method should be the systematic observation of overt behavior and its environmental determinants; and b) a metaphysical thesis, namely, that mental states and processes were mere explanatory chimeras, and that all behavior could be properly explained without resorting to mentalistic explanations (see Cooper et al., 2019; Dougher & Hayes, 2004; Guinther & Dougher, 2013; Froxán-Parga, 2020; Moore, 1981, 2013; Skinner, 1974; Zilio et al. 2021).

Methodological behaviorism is commonly tied to the former tenet, whose wide adoption leaded to the replacement of introspection reports by systematic observation of behavior as the proper method of psychology (Moore, 1981, 2013; Skinner, 1974, 1984; see also Froxán-Parga, 2020). Watson's metaphysical views about the mind, however, were less appealing, and many psychologists rejected it. The subsequent development of operationalism and the rise of neo-behaviorist theories (Tolman, 1928; Hull, 1945; see also Greenwood, 2015) maintained Watson's methodological maxim, but rejected his metaphysical views; instead, they viewed the measurement of publicly observable behavior as *precisely* the proper method to study mediational mental states and processes through inferential, hypothetico-deductive methods (Moore, 1981). As Moore (2013, p. 203) puts it:

> Methodological behaviorism adheres to a symbolic referential conception of complex verbal behavior and a particular conception of operationism. According to these conceptions, (a) words are things that symbolically refer to other things; (b) psychological terms are hypothetical constructs that, when given partial operational definitions, may be inferred to symbolically refer to or represent causal mental variables; and (c) the job of psychology is to use observables as proxies to stand for causal mental variables so that those variables may be investigated. (Moore, 2013, p. 203)

Thus S-R psychology, although initially grounded on the principles of classical and operant conditioning, progressively gave way to the introduction of hypothetical constructs regarding putative inner states of the organism (whether cognitive or bodily) that presumably filled the explanatory gaps between environmental stimuli and the individual's responses to them. Habits, drives, inner associations, cognitive maps, representations, etc. became an explanatory necessity under the influence of neo-behaviorism. This set the stage for the subsequent development of second-wave behavior therapy (i.e., cognitive behavioral therapy or CBT), which resulted from the merger between behavior therapy and cognitive therapy (see Bandura, 1969, 1974; Beck, 1963, 1964, 1979; Ellis, 1958, 1962; Kazdin, 1982; Lazarus, 1968, 1977; Mahoney, 1977a, 1977b; Mahoney & Kazdin, 1979; Meichenbaum, 1977).

But before we move on to second-wave behavior therapy, let's first go back to the other major strand within the first wave: applied behavior analysis. Of the two first-wave strands, behavior therapy was way more popular; applied behavior analysis, by contrast, had "far fewer adherents" (Dougher & Hayes, 2004, p. 14). Part of the reason was that while the former was widely implemented in ambulatory settings with verbally competent users, the application of the latter was more commonly implemented in more controlled inpatient settings -a seemingly trivial difference that, however, eventually had deep implications, as we'll

see in section 1.5.2.2. However, while behavior therapy was eventually assimilated by CBT, applied behavior analysis remained an independent psychological approach to clinical practice which still continues today. Due to its continued importance and its centrality in the discussion of contemporary psychological models (section 1.5.2.), here we'll review its core commitments in more detail.

One major reason why applied behavior analysis wasn't eventually assimilated by second-wave behavior therapy was its background philosophy of psychology. As we've said, behavior analysis didn't draw from methodological behaviorism, but from radical behaviorism (Skinner, 1945, 1974; see also Chiesa, 1994; Cooper et al., 2019; Froxán-Parga, 2020; Moore, 2008; Sturmey, 2020). The "radical" in radical behaviorism refers to Skinner's application of behavior analysis to the philosophy and science of psychology themselves, proposing to understand them in terms of verbal behavior, rather than in terms of logical or inferential connections among propositions (Skinner, 1945; see also Moore, 2008; Schneider & Morris, 1987). This leads radical behaviorists to three major inter-related commitments: a) the dismissal of "truth-as-correspondence" and "truth-as-coherence" criteria in favor of a more pragmatic "truth-as-effective action" criterion for assessing scientific analyses; b) the dismissal of hypothetical-deductive scientific methods and the strict preference for the systematic observation of behavior, emphasizing the goals of prediction and control over theoretical correspondence; and c) the application of these assumptions to the analysis of the use of subjective terms, which grounds the rejection of "private events" as something distinct from behavior itself and which enjoys a central role in the causal explanation of behavior (Skinner, 1945, 1953, 1957, 1963, 1974, 1990; see also Chiesa, 1994; Moore, 2008; Zilio et al., 2021). Taking all these features into account, radical behaviorism entails the conceptualization of behavior as the *radix* of psychology, i.e., as the proper subject matter of psychological analysis, and not just a proper method of inquiry into other more central topics, e.g., cognition (see also Dougher & Hayes, 2004; Follette et al., 1996; Froján-Parga, 2020; Guinther & Dougher, 2013; Kohlenberg et al., 1993; Madden et al., 2016; Pérez-Álvarez, 1996, 2012; Sturmey, 2020).

Let's see this in more detail. To understand the core commitments of radical behaviorism and behavior analysis, it's useful to contrast the behavior analytic account of behavior with that of S-R psychologists. S-R psychology was essentially mechanistic: it retained a view of the organism's behavior as the product of a series of mechanic relations of cause and effect connecting the external stimuli with the organism's overt responses -hence the eventual introduction of hypothetical constructs (see MacCorquodale & Meehl, 1948) to fill the gaps in the description of such mechanical machinery. In doing so, behavior itself was progressively displaced from the focus of analysis, relegated to a mere method to investigate what

was now considered the proper subject matter of psychology: the organism's internal states, the "glue" between world and action (see section 1.3.2.). Behavior analysts identify this kind of mechanistic explanations with *structuralist* approaches to psychology, which are primarily interested in the description of the organism's internal structures that ultimately give rise to their responses (see Cooper et al., 2019; Skinner, 1974; Sturmey, 2020).

By contrast, drawing from Darwinism and American pragmatism (James's, in particular), behavior analysts endorse a *functionalist*[7] and *selectionist* view of behavior (Skinner, 1953, 1957, 1971, 1974, 1981, 1990; see also Barrett, 2015, 2019; Chiesa, 1994; Cooper et al., 2019; Froxán-Parga, 2020; Hayes, 2016; Moore, 2008; Sturmey, 2020). Skinner conceived of behavior as the product of three main types of variation and selection by consequences: a) natural selection, which operates at a phylogenetical scale and would be responsible for species-characteristic behaviors; b) operant selection, which operates at an ontogenetical scale and whereby the environment selects variations in the behavior of individual organisms; and c) cultural selection, derived from the second type, which would operate at a social scale and would account for the variation and selection of cultural practices (Skinner, 1953, 1971, 1981, 1990; see also Alonso-Vega et al., 2020; Cooper et al., 2019; Sturmey, 2020). On this view, behavior is no longer equated with an organism's mechanical reactions to a series of environmental stimuli (which would unavoidably require a reference to intermediate causal steps); rather, it's understood in terms of the *functional relation* that is established between an organism's responses and the varying environmental *contingencies* that control them. In this sense, behavior analysts view behavior as a function of the varying probabilistic relations established among environmental events and between these and the organism's responses (see Skinner, 1953; see also Cooper et al., 2019; Moore, 2008; Sturmey, 2020).

Drawing from this functionalist and selectionist approach, behavior analysts reject the need to postulate hypothetical constructs to account for behavior. Behavior needs to be analyzed *at its own scale of analysis*: that of the functional relations established between the organism's responses and their selecting contingencies (Skinner, 1950, 1963, 1977, 1981, 1990; see also Barrett, 2015, 2019; Chiesa, 1994; Cooper et al., 2019; Froxán-Parga, 2020; Hayes, 2016; Moore, 2008; Pérez-Álvarez, 1996, 2004; Sturmey, 2020). That's why radical behaviorists maintain the irreducibility of behavior analysis to "lower-order" explanations. On the one hand, they view descriptions of whatever physiological processes occurring "underneath"

[7] "Functionalism" here refers to the "functional psychology" approach that derived from James's (1890/1981) work, which opposed structuralist approaches to consciousness around the end of the 19th and beginning of the 20th centuries, eventually leading to the development of behaviorism (see Schneider & Morris, 1987). Therefore, it must not be confounded with functionalist approaches to cognitive science (see Chapter 2, section 2.2.2.1.), developed much later, and which provided the grounds for the rise of cognitivism.

the organism's behavioral patterns as informative to explain *how* behavior occurs (i.e., what are the material conditions for it to even take place), but not *why* it occurs. On the other hand, behavior analysts (at least under a common reading of radical behaviorism) dismiss explanations that appeal to mental or cognitive constructs as mere "explanatory fictions" which preclude the analysis of the environmental sources of behavioral control (see Baum, 2011; Chiesa, 1994; Moore, 2009; Schnaitter, 1984; Skinner, 1945, 1950, 1953, 1971, 1974, 1977, 1990).

Despite appearances to the contrary, however, the common claim that "behavior analysts denied the existence of mental events" needs qualification (see also Chapter 8, section 8.4.). Behavior analysts didn't deny the existence of thoughts (construed as "inner speech"), feelings, nor many other "occurrent" mental states (see Chapter 2, section 2.1.); rather, they denied their conceptualization as *mental* events, and the reliability of introspective methods -although they didn't dismiss them as Watson and methodological behaviorists did (e.g., Skinner, 1974). In addition, in line with their functionalist and selectionist approach, they denied the explanatory necessity to appeal to mediational variables in order to explain behavior. In other words: what they denied was that so-called "private events" had any special character that granted their categorical distinction from behavioral events; rather, they conceptualized them as embedded in the larger "behavioral stream", and rejected any kind of explanatory strategy that artificially "brought investigation to an end" (Skinner, 1974, p. 19) by appealing to hypothetical internal events.

This points to one of the major differences between behavior therapy and applied behavior analysis. As we've seen, both targeted medical-psychoanalytic theories; however, while the former mainly did it on account of their empirical flaws and the unoperationalizable character of their concepts, the latter also targeted their overall *internalism*. In other words: applied behavior analysts weren't just concerned with the "intra-psychic" explanations of psychoanalysis, but with the internalist assumptions that characterize the medical model in general, both in its first psychoanalytic formulations and in its posterior neurobiological construal. According to this explanatory internalism, the root of psychopathology is to be ultimately found in some internal state of affairs, whether intrapsychic conflicts, cognitive processes, or neurobiological states. By contrast, behavior analytic approaches to clinical practice assumed that behavior, whether psychopathological or not, must be understood in irreducibly contextual terms (Cooper et al., 2019; Dougher & Hayes, 2000; Follette et al., 1996; Froján-Parga, 2020; Guinther & Dougher, 2013; Hayes, 2004, 2021; Hayes et al., 2012; Kohlenberg et al., 1993; Madden et al., 2016; Sturmey, 2020). That's why some claim that behavior analysis offers a *genuinely* psychological model, i.e., one which not only describes

its subject matter in psychological terms, but which establishes a scale of analysis of its own and defends its explanatory primacy for the study of psychopathology (e.g., Pérez-Álvarez, 2004, 2012; Froxán-Parga, 2020).

This is linked to the core behavior analytic assessment tool: the *functional analysis* of behavior (see section 1.5.2.1.; Cooper et al., 2019; Froxán-Parga, 2020; Sturmey, 2020). On this view, psychological treatment must be based on an individually tailored assessment of the environmental contingencies that maintain the problem. The problem's *topography* (i.e., its morphology, quantitative and qualitative dimensions, how it may be described in folk-psychological terms, etc.) becomes of secondary importance; it's its *function* what matters most for the behavior analytic oriented psychologists (Cooper et al., 2019; Froxán-Parga, 2020; Sturmey, 2020).

In a nutshell, these are the defining features of the behavior analytic approach to psychopathology: a) the rejection of mechanicist views of behavior and its conceptualization in functional and selectionist terms, i.e., in terms of the functional relation established between an organism's activity and the environmental contingencies that control it; b) the conceptualization of cognitive and experiential phenomena (e.g., feelings, thoughts, sensations, etc.) in behavioral terms, that is, as subtypes of behavioral *explananda* (and not as their internal causes); c) the conceptualization of mental health problems as irreducibly behavioral problems, to be analyzed at their own scale of analysis; and d) the use of functional assessment or functional analytic tools to determine the environmental contingencies maintaining problem behaviors and designing intervention plans accordingly. In section 1.5.2., we'll see that the main contemporary psychological approaches to psychopathology consist of developments of the behavior analytic tradition.

1.3.2. Second-wave behavior therapy: cognitive behavioral therapy

The advent of second-wave behavior therapy, or cognitive behavioral therapy (CBT) (Bandura, 1969, 1974; Beck, 1979; Mahoney, 1977a, 1977b; Mahoney & Kazdin, 1979; Meichenbaum, 1976, 1977; Kazdin, 1982) during the 1970's was mainly due to two main reasons: a) the growing general criticism against the somewhat poorly defined specter of behaviorism, as well as the insatisfaction with its account of verbal behavior, commonly traced back to Chomsky's critique of Skinner's (1957) *Verbal Behavior*; and, relatedly, b) the insatisfaction with behavioral explanations of clinical change in outpatient contexts (see Dougher & Hayes, 2004; Hayes, 2004). Against this background, the conceptual differences between behavior therapy and behavior analysis eventually led to a schism within first-wave behavior therapy: while the former, grounded on methodological behaviorism, progressively started to incorporate cognitive or mediational variables in their explanations of psychopathology and clinical change

(e.g., Bandura, 1969; Lazarus, 1968), the latter remained committed to the functionalist and selectionist explanatory framework of radical behaviorism (see Hayes, 2004).

Cognitive models of mental health problems and therapy would eventually emerge as an attempt to fill the explanatory gaps of behavior therapy and its S-R psychology model, as well as to reinstate the explanatory primacy of mental events, rejected by behavior analysts (e.g., Bandura, 1969, 1974; Beck, 1963, 1964, 1970, 1979; Ellis, 1958, 1962). As Hayes (2004, p. 642) puts it:

> Behavior therapists knew they needed to deal with thoughts and feelings in a more direct and central way. In the context of the failure of both associationism and behavior analysis to provide an adequate account of human language and cognition, the seeds planted by early cognitive mediational accounts of behavior change (e.g., Bandura, 1969) quickly flowered into the cognitive therapy movement (e.g., Beck, Rush, Shaw, & Emery, 1979; Mahoney, 1974; Meichenbaum, 1977). [...]. Early cognitive behavior therapies addressed cognition from a direct, clinically relevant point of view. Certain cognitive errors seemed characteristic of patient populations, and research proceeded directly to the identification of these errors and the methods needed to correct them. Hayes (2004, p. 642)

Initial works by the founders of cognitive therapy appeared during the 1950's and the 1960's (see Beck, 1952, 1961, 1963, 1964, 1970; Beck & Alford, 1967/2009; Ellis, 1958, 1962; see also Bandura, 1969), where the key assumption of the cognitive approach was laid out: namely, that world and action aren't directly related, and that the individual's interpretations or *representations* of the world necessarily mediate between the environment and its effect on the individual's behavior (e.g., see Beck, 1970; Dobson & Dozois, 2010; Knapp & Beck, 2008). Thus, despite their many differences (see Dobson & Dozois, 2010; Knapp & Beck, 2008), cognitivist approaches to psychopathology shared the fundamental assumption that a properly explanatory account of behavior needed to tackle one's interpretations of the world, including oneself and others. Partly modelled on the guiding computer's metaphor of the emerging cognitive science, partly modelled on the intrapsychic self of the psychoanalytic tradition (see Beck, 1952, 1961, 1970, 1979), the mind, with its self-creating powers, unresolved issues, inner conflicts, and distortions, was again reinstated as the primary locus of psychopathology and clinical change. The different cognitive processes hypothesized to causally mediate between world and action -e.g., Bandura's (1969, 1974) self-efficacy, Beck's cognitive schemas (see Beck & Alford, 1967/2009), Ellis's (1962) irrational ideas, etc.- can be viewed as mere variations of this common theme: the causal centrality of the inner mind in the interpretation of external events and the production of behavior. Due to its historical importance, here

we'll focus on Beck's cognitive model to exemplify the main tenets of cognitivist models of psychotherapy.

Beck's approach was first developed in his renown cognitive theory of depression (see Beck, 1961, 1963, 1964, 1979; Beck & Alford, 1967/2009), which was later applied to many other mental health problems (Knapp & Beck, 2008). Already in 1963, Beck laid the grounds of his theory. Drawing from a comparison between people with depression and people with other mental health problems, Beck (1963) identified the main themes of depressive thought, e.g., "low self-esteem, self-blaming, overwhelming responsibilities, and desires to escape" (Beck, 1963, p. 326). In addition, he characterized these negative cognitions in terms of their "formal properties", namely their automatic, involuntary, and ruminative character, as well as their self-perceived plausibility. But, most importantly, he identified a series of *cognitive distortions* or rationality failures -common to *all* nosological groups-, whose pathological nature he identified with their *systematic* character, e.g., "arbitrary inferences", "selective abstractions", and "overgeneralizations" (see Beck, 1963, p. 328; see also Beck, 1964, 1979; Beck & Alford, 1967/2009). Although initially focused on the person's self-concept, Beck eventually extended his analysis to cover the person's distorted interpretations of external events and of the future. This leaded, in 1967 (see Beck & Alford, 1967/2009; see also Beck, 1979), to the formulation of his famous *cognitive triad*, which comprises three core *cognitive schemas* or *structures* (Beck, 1964), the deepest roots of psychopathology: negative views about the world, negative views about oneself, and negative views about the future.

Beck's cognitive model was thus organized in a hierarchical manner, with core negative schemas at the "deep" level giving rise to ruminative, automatic, and involuntary negative thoughts at the "surface" level; these negative thoughts, in turn, would cause emotional distress and behavioral disruptions. Cognitive distortions (i.e., the person's systematically faulty cognitive processes) would bidirectionally mediate the relation between the deep and surface levels by a) prompting negative and distressing appraisals automatically; and b) in turn, feeding them back to the person's core cognitive schemas (Beck, 1964, 1979; Beck & Alford, 1967/2009; see also Hyland & Boduszek, 2012; Knapp & Beck, 2008). This model of the individual's faulty information processing machinery provided a way to distinguish clinical and non-clinical cases (Beck & Alford, 1967/2009). It also laid the ground for his subsequent development of cognitive therapy (Beck, 1979), and the development of contemporary cognitive procedures. The chief example are cognitive restructuring techniques (e.g., Socratic dialog; see Froxán-Parga et al., 2018), whose main goal is to help people identify and challenge their negative automatic cognitions and faulty cognitive processes, and teach them

alternative ways of appraising events to ultimately modify their cognitive structures (see Knapp & Beck, 2008).

For behavior therapists, these hypothetical cognitive mediation processes provided the key to explain clinical changes following so-called "talk therapy" that couldn't be otherwise explained. In this sense, "methodological behaviorism provided a ready means for the transition from the first to the second wave of behavior therapy" (Hayes, 2004, p. 642); namely, it provided behavior therapists with a scientifically respectable way to study mediational processes that could explain individual differences in therapy and fill the "explanatory gaps" between the in-session application of therapeutic techniques and clinical changes outside the clinical context. In addition, it contributed to the development of verbal-cognitive techniques (e.g., cognitive restructuring, self-instructional training, etc.) to deal with the person's self-defeating views of the world, themselves, and the future, as well as to boost the efficacy of traditional behavior modification procedures. Although some behavioral thinkers challenged the emerging cognitive models on account of their conceptual difficulties and explanatory redundancy (Ledwidge, 1978; Rachlin, 1977a, 1977b), or protested that cognitive variables had been already considered by behavior analysts and therapists (Wolpe, 1978), the so-called "cognitive revolution" in psychotherapy (Mahoney, 1977b) proved to be unstoppable: cognitive (i.e., second-wave) behavior therapy had arrived to stay. Although in sections 1.5.2. and 1.5.3. we'll see some contemporary non-cognitivist approaches, many still think of classical CBT as the canon regarding psychological models of mental health problems.

Some have questioned whether this particular psychological model constitutes a real challenge to the medical model of mental health problems (González-Pardo & Pérez-Álvarez, 2007). To be sure, many cognitive thinkers have explicitly rejected biomedical perspectives and emphasized the need to tackle psychological problems at a scale of analysis of their own, primarily defined in terms of the person's cognitions and coping strategies (e.g., see Beck, in Bentall, 2003, p. xi). However, the truth is that CBT leaves the medical model largely unchallenged, at least in its minimal interpretation: the common use of medical jargon and comparison methods to establish CBT's efficacy, as well as the tendency towards manualization and standardization of clinical procedures reveal the medicalization of CBT's psychological model. But the major problem lies in its internalist assumptions: by accounting for mental health conditions in terms of primarily internal *deficits* -whose description in psychological terms may or may not be reducible to the language of neurobiology, but which nonetheless remain "within" the person- CBT retains the individualistic thinking that characterizes the medical model (González-Pardo & Pérez-Álvarez, 2007).

Thus far, we've seen how various critical approaches, developed during the 1960's and the 1970's, questioned the viability of the medical model, both in its minimal and strong interpretation. Although drawing from completely opposite points of view with regard to many matters, all critical models converged on their vindication, in one way or another, of the relevance of psychological and social affairs for attaining a proper understanding of mental health problems. Szasz and Laing, for instance, advocated for different kinds of *personal* approaches to psychopathology, analyzing psychological problems in terms of meaning and interpretability; on the other hand, first-wave and second-wave behavioral approaches to psychotherapy prioritized the psychological scale of analysis, viewing mental health problems in a continuum with other non-clinical phenomena and emphasizing the psychological causes of psychopathology (whether understood in behavioral or cognitive terms). In the next section, we'll see what was the response from psychiatry to this diverging attacks. We'll focus on Engel's biopsychosocial model, the most important historical attempt to find an integrative framework for the field of mental health.

1.4. The biopsychosocial model and the integration problem

As we saw in section 1.1., the official stance of the most important institutions in Western psychiatry has been the minimal interpretation of the medical model. The adoption of this relatively uncommitted standpoint, reflected in the successive editions of the DSM, was partly due to the inner tensions among different "schools of thought" within psychiatry and the mental health disciplines in general, as well as to the various criticisms raised against the medical model throughout the 1960's and the 1970's. Above all, it was supposed to prevent the adoption of a narrow theoretical and practical focus that could hinder research on the causes of mental health problems and on the development of effective intervention procedures.

At the end of the 1970's, this uncommitted approach took the form of an intendedly integrative theoretical framework: the biopsychosocial model, which was first fully articulated by Engel (1960, 1977, 1978, 1980, 1997; although see Ghaemi, 2010). This model pretended to be an end to the "priority wars" of previous decades. A few years before, Luborsky et al. (1975, p. 1003) had made their "dodo bird verdict" regarding the comparative effectiveness of different kinds of psychotherapies -a highly disputed verdict, on the other hand (see Eysenck, 1993); now, the biopsychosocial model also declared that "Everyone has won and all must have prizes", although this time in the realm of theory (Aftab & Nielsen, 2021). Allegedly, the priority problem was solved: now all three levels, the biological, the psychological, and the social, were deemed equally relevant to understand mental health problems.

Engel's holistic model pursued no less holistic ambitions; rather than just aiming to provide unifying framework for psychiatry, Engel's goal was to provide a "new medical model" for medicine as a whole. In his seminal paper, Engel (1977) states:

> I contend that all medicine is in crisis and, further, that medicine's crisis derives from the same basic fault as psychiatry's, namely, adherence to a model of disease no longer adequate for the scientific tasks and social responsibilities of either medicine or psychiatry. [...] Medicine's crisis stems from the logical inference that since "disease" is defined in terms of somatic parameters, physicians need not be concerned with psychosocial issues which lie outside medicine's responsibility and authority. (Engel, 1977, p. 129)

Therefore, his proposal was to provide a synthetic solution to what he saw as two antithetical models of illness: the psychosocial one and the biomedical one. He viewed the former as captained by Szasz, whose approach he described as advancing the "removal of the functions now performed by psychiatry from the conceptual and professional jurisdiction of medicine and their reallocation to a new discipline based on behavioral science" (Engel, 1977, p.129). In other words: he took Szasz's criticism as vindicating the psychosocial level of *causal* explanation for psychiatry, placing him side by side with behavior scientists (to their very likely mutual disgust). On the other hand, he viewed the biomedical model as an originally scientific model that had permeated the boundaries of scientific research, becoming "the dominant folk model of disease in the Western World" and thus finally acquiring "the status of a dogma" (Engel, 1977, p. 130). Both positions, according to Engel, disagreed about whether psychiatry could be based on a biomedical model, but both agreed that this model was the proper one for medicine.

The biopsychosocial model aimed to contest this latter common assumption (Engel, 1977, 1978, 1980). For him, the split consideration of the biological, on the one hand, and the "human" or psychosocial on the other, was a vestige of a *dualist* approach to the mind-body problem (see Chapter 2); instead, he assumed, medicine, and psychiatry as an irremediably medical discipline, ought to adopt a holistic approach to this relation, and thus *integrate* both science and humanism in theory and in practice.

In the realm of theory, Engel saw the allegedly "dualist" philosophical assumptions at the bottom of the biomedical model as the main responsible for its dogmatic character, which precluded the integration of empirical data from the biological, the psychological, and the social sciences (Engel, 1977, 1978, 1980). Instead, Engel wanted to implement a new philosophical framework, based on the biologist von Bertalanffy's General Systems Theory (see von Bertalanffy, 1950, 1968). This theory emphasized the need to complement the analytic

method characteristic of modern science with a more holistic view of the functioning of living systems in order to properly account for their self-organizing capacities and goal-directedness. In this sense, von Bertalanffy thought that it was necessary to consider both same-scale and cross-scale causal interactions within a biological system's components and between the whole biological system and the environment. This conflation of causal-explanatory projects and, most importantly, of subpersonal and personal languages within a single scientific and philosophical framework, provided the theoretical grounds for Engel's multilevel approach to mental health phenomena, whose basic tenets he described as follows:

> Each level in the hierarchy represents an organized dynamic whole [...]. Cell, organ, person, family each indicate a level of complex integrated organization [...] implies qualities and relationships distinctive for that level of organization, and each requires criteria for study and explanation unique for that level.
> [...] *Each system is at the same time a component of higher systems* [...]. Each system as a whole has its own unique characteristics and dynamics; as a part it is a component of a higher-level system. The designation "system" bespeaks the existence of a stable configuration in time and space, a configuration that is maintained not only by the coordination of component parts in some kind of internal dynamic network but also by the characteristics of the larger system of which it is a component part. Stable configuration also implies the existence of boundaries between organized systems across which material and information flow. (Engel, 1980, p. 536-537)

With regard to clinical praxis, Engel developed the practical implications of his model in subsequent works (Engel, 1978, 1980). He put a special emphasis on the educational implications of the model: medical students should be taught, from the beginning, the psychosocial aspects of health problems, so they could incorporate such knowledge to their daily practice as physicians (Engel, 1977, 1978, 1980). In Engel (1980), he attempted to show how exactly his multilevel approach could be applied to enhance medical treatment, not of a person with a mental health problem, but of a person with a heart attack. Although the patient luckily survived, Engel concluded from his case exposition that, perhaps, if the medical staff attending the patient had adopted a systems approach, the noxious effect of psychosocial variables on the course of the heart attack (e.g., the patient's personality style, which made him prone to misestimate the severity of his symptoms, or his anxiety during some of the medical procedures, which triggered ventricular fibrillation) could have been prevented.

We began section 1.1. by stating that many would agree that the medical model is the hegemonic conception of psychological problems, despite the fact that many practitioners

report ascribing to the biopsychosocial model (see Fulford & van Staden, 2013) and despite the fact that many steadfast advocators of the medical model see themselves as a minority (Craddock et al., 2008; Ghaemi, 2009, 2010; Shah & Mountain, 2007). These authors have argued instead that it's the biopsychosocial model which constitutes the "mainstream ideology of contemporary psychiatry" (Ghaemi, 2009, p. 3). In view of Engel's proposal, the alleged mismatch between self-professed and enacted models of care no longer seems contradictory. After all, his biopsychosocial model was, as he indeed intended it to be, a new medical model; one where the medical status of psychiatry remained unquestioned.

Despite its popularity, however, the biopsychosocial model has also been the target of sound criticisms, both concerning its theoretical and practical viability (Craddock et al., 2008; de Haan, 2020a, 2020b, 2020c; Ghaemi, 2009, 2010; Murphy, 2013; Matthews, 2013; Pilgrim, 2015; Van Oudenhove & Cuypers, 2014; see also Bolton & Gillett, 2019). To begin with, it's clear from how he lumped Szasz's approach and behavioral approaches together that his advocacy of the biopsychosocial model for medicine drew from a misreading of Szasz's critique (see Chapter 2, section 2.4.) and thus from a false premise. He rejected Szasz's attack on the medical status of psychiatry on the grounds that not only psychiatry, but the whole medical field should also incorporate the psychosocial in their *causal* theories of illness. Yet, as we saw in section 1.2., Szasz's claim that mental illnesses were mythical creatures wasn't based upon the observation that mental health problems were better causally explained at the psychological or social level, but upon the observation that *it didn't make sense* to even speak of minds (which he understood as exclusively definable in personal terms) as being sick or ill, except in metaphorical terms. Engel bypassed this critical observation and drew his biopsychosocial model on the unquestioned assumption that mental illness was the proper subject matter of psychiatry. Moreover, in the happiest of coincidences, not only psychiatry's medical status shouldn't be put in question; in fact, it was psychiatry itself the branch of medicine that should become a role model for the rest of medical fields, for it had first acknowledged the need to embrace a more holistic causal model.

Besides this conceptual flaw, the most pressing problem of the biopsychosocial model lies in its inability to provide a sound unifying framework to understand how exactly different scales of analysis relate to each other, both in theory and in practice[8]. As we said, Engel's

[8] Awais & Nielsen (2021) have recently claimed that interpretations and criticisms of the biopsychosocial model that emphasize cross-scale causal interactions overlook that Engel was primarily interested in providing "a framework that would bring the psychosocial and phenomenological dimensions of illness within the realm of medical and scientific inquiry", where "causes and risk factors are included (...), but they are not particularly privileged by Engel" (p. 9). We agree with this claim. However, we also think that Engel's explicit reliance on

multi-level framework pretended to put an end to the inter-disciplinary struggles of the previous decades, where a major theoretical concern was the priority problem. However, his attempted solution to this issue leaded directly to another one: the integration problem (de Haan, 2020a, 2020b, 2020c; Ghaemi, 2009, 2010; Kendler, 2005; Murphy, 2013; Matthews, 2013; Pilgrim, 2015; Van Oudenhove & Cuypers, 2014; Walter, 2013; see also Bolton & Gillett, 2019). Some critics point out that this was due to the overall *vagueness* with which Engel defined each of the relevant scales of analysis and the possible relations among them. Besides the mere formal recognition of the potential relevancy of all three scales to psychopathology, how exactly are the biological, the psychological, and the social to integrate with each other? How is the "mental" or "psychological", for instance, causally related to the physical (see Chapter 2)? In this sense, Engel didn't even provide a clear definition of what he understood as "psychological variables". Although his model was supposed to provide a reconciliation between biomedicine and "behavioral science", his own account of the mental and of the psychophysical laws relating the biological and the psychological seemed to be strongly influenced by his own psychoanalytic background (Engel, 1960; see also Ghaemi, 2009, 2010).

This has obvious implications for practice. In absence of a proper account of how each level ought to be defined and related to each other, Engel's approach comes down to a plea for theoretical and practical eclecticism (Ghaemi, 2009, 2010). The apparent wonders of such eclecticism, so many times conjured against "dogmatism" and "closed-mindedness" in mental health disciplines, fade quickly in the face of some of its possible consequences. These are best exemplified by one of the antecedents of Engel's approach, e.g., Meyer's psychobiology. Meyer, cited by Engel, also conceived psychobiology as the integrative and multilevel study of the psychical life, ranging from the physicochemical and the neurological to the psychological and social (Ghaemi, 2010; Scull & Schulkin, 2009). This approach grounded Meyer's own plea for theoretical and practical eclecticism, which allowed him to endorse and provide ongoing support for extremely different and often incompatible perspectives on psychopathology. On the one hand, he is often recognized for his work on the philosophy of occupational therapy; on the other, he was equally ready to support and endorse some of the most brutal biomedical-based procedures in the history of psychiatry. For example, he supported his former student Cotton's focal toxin theory of mental illness -according to which mental health problems were the result of infected organs releasing toxins that affected the brain-, as well as Cotton's preferred treatment methods: teeth removal, colectomy, and, when these failed, the "surgical extirpation of [other] offending organs", including "tonsils,

General Systems Theory, as well as his descriptions of how psychosocial factors may influence health outcomes, reveal that he was also explicitly invested in ontological concerns about cross-scale causal interactions.

spleens, stomachs, [...] uteri, and so forth") (Scull & Schulkin, 2009, p. 24; see also Ghaemi, 2010)[9]. Not to mention his support for Freeman's frontal lobotomy, which he warmly endorsed when the procedure was first presented in 1936, despite the even then obvious risks and few guarantees of success (Ghaemi, 2010).

To be sure, eclecticism *per se* needs not derive in such monstruous consequences, which were more driven by scientistic and neuro-centric views of mental health rather than eclecticism -although it helped legitimize them. In fact, we think that the kind of *pluralist* spirit behind Engel's approach, properly understood, is worth keeping (see Chapter 9, section 9.2.). However, this must be framed within a sound conceptual and scientific approach; otherwise, it leaves the door open for all possible kinds of pseudoscientific theories and procedures across all possible levels. Moreover, it also leaves the door open for reductivist construals of the biopsychosocial model, according to which the "integrative" scale of analysis would be the biological (neural) one (see section 1.5.1.).

In fact, this also seems to be what Engel had in mind sometimes. When listing "the requirements for a more inclusive scientific medical model for the study of disease", he explicitly begins by admitting that "the biochemical defect constitutes but one factor among many, the complex interaction of which ultimately may culminate in active disease or manifest illness" and that "the biochemical defect [can't] be made to account for all of the illness, for full understanding requires additional concepts and frames of reference" (Engel, 1977, p. 131). However, the rest of reasons that he gives for paying more attention to psychosocial variables don't reflect an actual acknowledgement of the possibility that psychosocial variables could be as relevant as biological ones in the causal explanation of mental health problems; instead, as Aftab & Nielsen (2021) argue, what mattered most to Engel had "more to do with psychosocial influences in the form of illness interpretation and presentation, sick role, seeking or rejection of care, the doctor-patient therapeutic relationship, and role of personality factors and family relationships in recovery from illness" (p. 9; see also Engel, 1977, p. 132). Moreover, when he addressed how psychological variables (e.g., the patient's perception of his relationship with the doctor) might influence the course of treatment, he understands this influence as mediated "by virtue of interactions between psychophysiological reactions and biochemical processes implicated in the disease" (Engel, 1977, p. 132). This contributes to paint a picture of psychosocial interventions as having a mediating or *adjuvant* role in the treatment of mental health problems; a picture that is reinforced by his sharp distinction between the "curing" and "caring" functions of the physician, whereby the former would

[9] In fact, according to Scull & Schulkin (2009), Meyer not only endorsed Cotton's brutal practices, but also prevented his subordinate Greenacre from exposing the ineffectiveness of Cotton's methods.

comprise the intervention on abnormal biochemical processes and the latter would involve "the more personal, human, psychological and social aspects of health and disease" (Engel, 1978, p. 170).

This implicit prioritization of the biological level in the causal explanation of and intervention on psychopathology can partially explain the above-mentioned discrepancy between the self-professed biopsychosocial orientation of many researchers and practitioners and their perception of the biomedical hegemony in mental health research and practice. In the next section, we'll see how this implicit prioritization has actually become explicit in the last few years, when the reliability and validity crisis of traditional nosologies has given rise to a newest medical approach, based on the main tenets of the so-called "third wave of biological psychiatry" (Walter, 2013) and captained by some of the most important mental health institutions in the Western world.

1.5. Contemporary approaches to mental health

Thus far, we've seen several reasons why the different critical approaches to mental health have questioned the viability of the medical model of mental health problems, both in its minimal and strong interpretation, namely: a) its framing of psychological problems in the language of medicine; b) the concomitant assumption that diagnoses of mental disorders not only describe the behavior and experiences of the people diagnosed with them, but actually point to "something wrong somewhere else"; and c) its narrow focus on neurophysiological accounts of psychological problems. We've also seen how the biopsychosocial model emerged in the late 1970's to provide an eclectic framework for both medicine and psychiatry that could preserve the medical status of the latter. In this section, we'll see how some of these disputes have evolved and have helped configure some of the most relevant contemporary approaches to mental health.

We'll begin with what Walter (2013) has called "the third wave of biological psychiatry". We'll then review some recent functional analytic approaches to mental health, which constitute the main current psychological alternatives to CBT. Finally, we'll show how the most recently developed enactive approach to psychiatry has revindicated the integrative project of the biopsychosocial model.

1.5.1. Precision medicine and third-wave biological psychiatry

As we've seen, the medical approach to mental health problems has been a steady subject of concern and criticism since at least the second half of the twentieth century. Despite the storm of criticisms it faced during the 1960's and the 1970's, it retained its hegemony as the mainstream paradigm for research and treatment. Ever since, the official stand of the most

important mental health agencies in the Western world remained committed to a minimalist interpretation of the medical model, which was later expressed in the adoption of a minimally committal biopsychosocial approach.

However, in recent years, these minimalist assumptions have come under attack from many different sources. One of the most important contending issues in this sense has been the reliability and validity crisis of traditional diagnostic tools (see section 1.1.). As we saw above, the Kraepelinian approach, consolidated with the publication of the DSM-III, privileged reliability over validity; following the Hempelian approach to taxonomy, it set itself the task of first securing a reliable taxonomy that could be later used for validation research (i.e., research on the actual etiology of mental disorders so classified). However, this approach has been steadily questioned since its inception, with authors such as Eysenck (1970, cited in Eysenck, 1983) already pointing out in the 1970's the apparent lack of reliability of many diagnostic categories and advancing the need for replacing the categorical approach with a dimensional one. About four decades later, the controversy surrounding the beforemath and aftermath of the publication of the DSM-5 in 2013 echoed these old concerns. The main difference was that, this time, the evidence on the reliability and validity problems of traditional diagnostic tools (e.g., high co-morbidity, low-reliability, diagnostic instability, arbitrary boundaries, lack of treatment specificity, lack of biomarkers, etc.) was overwhelming (see Cooper, 2014; Deacon & McKay, 2015; Keshavan et al., 2011, 2013; Lacasse & Leo, 2015; Markova & Berrios, 2012; Peele, 2015; Tandon, 2013; Whitaker, 2011).

This led many professionals and researchers, not only from "without" disciplines but also within institutional psychiatry, to raise serious doubts regarding the viability of the neo-Kraepelinian project. But this tsunamic "second wave" of criticism has not signaled the end of the hegemony of the medical model; much to the contrary, it has led many to endorse a stronger interpretation of it, i.e., one that readily conceptualizes mental health problems as brain disorders and that assumes that a proper psychiatric taxonomy must be primarily based on neuroscientific research. On this account, the reliability and validity problems of traditional nosologies don't lie in the medical model itself, but precisely on its minimal interpretation and on the historical reluctance of institutional psychiatry to fully embrace a neurobiological understanding of mental health problems (Andreasen, 1997, 2001; Cuthbert, 2014; Cuthbert & Insel, 2013; Insel, 2013, 2014; Insel et al., 2010; Insel & Cuthbert, 2015).

The historical and conceptual background of these "within" criticisms lies in the 1990's, or the so-called "decade of the brain", when public wonder about the promises of brain science underwent a major boom (Andreasen, 1997, 2001; Murphy, 2020; Varga, 2015; Walter, 2013). This decade marked the birth of third-wave biological psychiatry, which

Walter (2013) views as driven by two major forces: "progress in molecular neuroscience" and "the birth of cognitive neuroscience and neuroimaging" (p. 2). In a nutshell, the core idea behind third-wave biological psychiatry is the reconceptualization of psychiatry as an applied field of cognitive neuroscience. Psychiatry as applied cognitive neuroscience (see Andreasen, 2001) -or as "cognitive neuropsychiatry", as others have called the discipline (see Coltheart, 2007)- draws from the idea that cognitive neuroscientific and neuropsychological models of normal cognitive and neural functioning can be used to: a) map the specific information processing alterations that characterize different forms of psychopathology; and b) establish the neurobiological abnormalities behind such information processing anomalies. As Andreasen (1997) defines it:

> Contemporary psychiatry studies mental illnesses as diseases that manifest as mind and arise from brain. It is the discipline within cognitive neuroscience that integrates information from all these related disciplines in order to develop models that explain the cognitive dysfunctions of psychiatric patients based on knowledge of normal brain/mind function. [...] Finding the neural mechanisms of mental illnesses must be an iterative process; syndromal clinical definitions (or the phenomenotype) are progressively tested, refined, and redefined through the measurement of neurobiological aspects (Andreasen, 1997, pp. 1586-1587).

The rising force of third-wave biological psychiatry since the 1990's was consequently followed by an ongoing attrition of the neo-Kraepelinian project characteristic of the minimal medical model. Already in 2002, Kupfer (Chair of the DSM-5 Task Force), Regier (Vice-Chair) and First pointed out in their *A Research Agenda for DSM-V* the "need to explore the possibility of fundamental changes in the Neo-Kraepelinian diagnostic paradigm" (Kupfer et al., 2002, p. xviii), due to the above-mentioned reliability and validity problems of the DSM. They suggested that "research exclusively focused on refining the DSM-defined syndromes may never be successful in uncovering their underlying etiologies" and that "for that to happen, an as yet unknown paradigm shift may need to occur" (p. xix). However, despite these initial considerations, the final version of the manual ended up prioritizing, once again, "clinical utility" over validity (see section 1.1.).

The inner tensions of institutional psychiatry exploded shortly before the publication of the DSM-5. Three weeks before its publication, the then director of the NIMH published an entry on the NIHM's blog criticizing the DSM-5 for the above-mentioned validity problems of its "symptoms-based" approach (Insel, 2013). There he announced that "NIMH [would] be re-orienting its research away from DSM categories [...] supporting research projects that look across current categories –or sub-divide current categories– to begin to

develop a better system" (Insel, 2013, par. 6; see also Cuthbert, 2014; Cuthbert & Insel, 2013; Insel, 2014; Insel et al., 2010; Insel & Cuthbert, 2015; Walter, 2013; Tabb, 2020). The alternative research framework proposed was the Research Domain Criteria (RDoC) initiative. Echoing in its name "the rationale for developing the Research Diagnostic Criteria in the 1970s that led to the innovative DSM-III for clinical use" (Insel, 2010, p. 748), the RDoC initiative was devised by the NIHM in 2008 "to explore ways of incorporating such methods as genetics, neuroimaging, and cognitive science into future diagnostic schemes based upon behavioral dimensions and neural systems" (Cuthbert, 2014, p. 28). This way, the NIHM became the first psychiatric institution to officially endorse the strong interpretation of the medical model and the basic tenets of third-wave biological psychiatry.

The RDoC initiative can be primarily characterized by its brain-centered focus, its multi-level perspective, and its bottom-up approach to clinical categorization. Firstly, it "conceptualizes mental illnesses as brain disorders", assuming that "in contrast to neurological disorders with identifiable lesions, mental disorders can be addressed as disorders of brain circuits" (Insel et al., 2010, p. 749). However, it explicitly considers other relevant scales of analysis within its research program. In that sense, the RDoC model can be seen as the cathartic culmination of the contradiction between biopsychosocial advocacy and biomedical practice that we mentioned above: in line with the former, it adopts a multilevel approach to mental health phenomena; in line with the latter, it prioritizes the "brain circuit" scale of analysis as a means of integration among scales, so as to avoid the perils of eclecticism. Finally, instead of proceeding in a Hempelian manner (i.e., first securing the reliability of higher-order taxonomies and then proceeding to determine their neurobiological validators), this new biomedical model is defined by a bottom-up approach, based on the assumption "that data from genetics and clinical neuroscience will yield biosignatures that will augment clinical symptoms and signs for clinical management". This turns neo-Kraepelinianism upside down: focusing first on the search of proper validators for more specific units of analysis will eventually bring about more reliable taxonomies at the higher-order level. In sum, Insel et al. (2010) describe their approach as follows:

> The primary focus for RDoC is on neural circuitry, with levels of analysis progressing in one of two directions: upwards from measures of circuitry function to clinically relevant variation, or downwards to the genetic and molecular/cellular factors that ultimately influence such function. [It considers] different levels of analysis, from genetic, molecular, and cellular levels, proceeding to the circuit-level (which, as suggested above, is *the focal element of the RDoC organization*), and on to the level of the individual, family environment, and social context.

> Importantly, all of these levels are seen as affecting both the biology and psychology of mental illness (Insel et al., 2010, p. 749, emphasis added).

RDOC's emphasis on validity over reliability has been recently compensated by a newest approach to nosology: the Hierarchical Taxonomy Of Psychopathology (HiTOP) (Kotov et al., 2017, 2018, 2020). HiTOP's solution to the reliability problems of the DSM is to reject the arbitrary boundaries between normality and disorder, as well as between disorders, which result from a categorical approach to psychiatric taxonomy; instead, it proposes to adopt a dimensional approach, which can be traced back to Eysenck's (1970, cited in Eysenck et al., 1983) proposals. In particular, it advances a hierarchical, multi-dimensional nosological model, which results from the successive application of factor-analytic methods in what Kotov et al. (2017, p. 456) call "structural studies". These explore the actual statistical structure (i.e., the interrelation patterns) of psychopathological phenomena. The resulting hierarchical multi-level model includes subsequent dimensions ranging from "homogeneous components" (groups of closely related symptoms) and "maladaptive traits" to "superspectra", such as the "*p* factor" (i.e., a general factor assumed to underlie all kinds of mental health problems). Middle levels include "syndromes" (i.e., groups of components and traits), "subfactors" (i.e., groups of syndromes), "spectra" (i.e., larger groups of syndromes) (Kotov et al., 2017, p. 456)

RDoC and HiTOP are thus natural allies in their quest for a new approach to psychiatry. On the one hand, HiTOP provides a framework to openly investigate the multi-level *statistical* structure of mental health problems without self-imposed aprioristic diagnostic constraints. As Kotov et al. (2017, p. 459) themselves suggest, this would provide clearer "psychiatric phenotypes" across every statistically relevant dimension, which RDoC researchers could use as a roadmap for progressively establishing their "biosignatures" or "genotypes", i.e., the multi-level *etiological* structure of psychopathology. In turn, RDoC researchers would help to establish the exact nature of the statistical dimensions of the HiTOP approach, as well as new theoretical constructs that the quantitative phenotypical nosology should include. In an almost perfect match, the latter's emphasis on validity research is neatly balanced by the former's focus on reliability; as Kotov et al. (2017, p. 459) conclude, "these two efforts approach nosology from different perspectives, but are well positioned to advance toward one another to produce a unified system". Considered together, these two initiatives constitute the two-headed spearhead of the *precision medicine approach to psychiatry* (see Insel, 2014), or what we could call *precision psychiatry*. Precision medicine is an approach to health problems that aims at "the tailoring of medical treatment to the individual

characteristics of each patient" (National Research Council, 2011, cited in Tabb, 2020, p. 308). Similarly, precision psychiatry involves differentiating and targeting specific clinical population subgroups for research and intervention, with the aim of enhancing healthcare quality and maximize treatment efficacy and efficiency.

The main characteristics of this precision psychiatry approach afford straightaway solutions to the old problems of psychiatry. As we've seen, although HiTOP and RDoC focus on different aspects of psychopathology (i.e., its statistical and etiological structure, respectively), both adopt a hierarchical, multi-level, and dimensional approach. By encouraging a dimensional approach to psychopathology, both circumvent the boundary problem, since they posit no cut-off boundaries between non-clinical and clinical patterns of behavior, cognition, and action. In addition, by combining their multi-level approach with RDOC's emphasis on the brain circuitry level, these precision psychiatry approaches provide a clear answer to the analogy, priority, and integration problems: a) mental health problems are analogous to somatic health problems (specifically, they are "disorders of the brain"); b) all levels of analysis are potentially relevant to understand psychopathological phenomena; and c) the brain circuitry level is the one at which all the other relevant scales of analysis merge to ultimately produce mental health problems.

It's questionable though that these answers provide proper solutions to these problems. To begin with, assuming that there're no clear boundaries between clinical and non-clinical phenomena might provide a better understanding of how certain behavioral, cognitive, or experiential patterns are widely distributed throughout the whole population, but it doesn't answer the question as to why exactly *those* patterns, and not others, are pathological (and not, say, just deviant or non-normative). In addition, precision psychiatry doesn't solve the analogy paradox, pointed out by Szasz and other critical thinkers: even if we were to assume that mental disorders are brain disorders, why exactly should we simply accept that dysfunctions in brain circuitry amount to *mental* disorders, while disorders following identifiable brain lesions are somatic (i.e., neurological)? What's specifically mental about brain circuits? Finally, of course, we might also wonder why exactly brain circuitry, and not any other scale of analysis, should have causal explanatory or conceptual priority over the rest. In empirical terms, there's just no available evidence to support the idea that brain processes should be granted explanatory primacy over, say, psychological or social processes (see Borsboom et al., 2019; Deacon & McKay, 2015; Kendler, 2005; Kendler & Schaffner, 2011; Keshavan et al., 2011; Lacasse & Leo, 2015; Leichsenring et al., 2022; Peele, 2015; Whitaker, 2011). In conceptual terms, the priority of the neural scale of analysis just seems to reflect a preconceived and much-criticized Cartesian understanding of minds as inner

representational and computational systems, whose operations are primarily realized by the brain (see Chapter 2); a view that, as we'll see in upcoming chapters, faces serious problems.

In sum, precision psychiatry leaves the analogy and boundary problems unexplained; in addition, in its attempt to provide a way out of the integration problem, third-wave biological psychiatry unfoundedly prioritizes the "brain circuitry level" over other scales of analysis. We'll now see how, contrary to RDoC's neurocentrist model, contemporary approaches to clinical psychology have continued to focus on the psychological level, aiming to explain how the environment and the learning history of an agent can have an impact on their behavior, thought, emotion, and experience.

1.5.2. Functional analytic approaches and the third-wave of behavior-therapy

In section 1.3. we reviewed the main historical psychological models, which defend the conceptual or causal priority of psychological explanations in the understanding of psychopathology. Following Hayes's (2004) nomenclature, we distinguished between first-wave and second-wave behavior therapy, the first comprising behavior therapy and early applied behavior analysis, and the second referring to the merger of behavior therapy and cognitive therapy and the subsequent development of cognitive behavioral therapy (CBT). To be sure, CBT continues to be one of the prevailing approaches to psychopathology from a psychological standpoint, and so it'd perhaps make sense to discuss it as a contemporary psychological model. However, here we want to focus on some contemporary *functional analytic*[10] approaches that, drawing from common behavior analytic roots, dispute the hegemony of classical CBT. We'll focus on two major developments and their implications for the functional analytic approach to mental health, namely: a) the development of Functional Behavioral Assessment (hence FBA) methods; and b) the contemporary research on verbal and complex (e.g., symbolic) behaviors. In particular, we'll focus on the relevance of the latter for the development of *third-wave behavior therapy* (Hayes, 2004), and how this eventually leaded to a sort of schism within the field of behavior analysis (see Hayes, 2016, 2021). Due to its relevance in this sense, we'll focus on Acceptance and Commitment Therapy (hence ACT; Hayes et al., 1999), which constitutes a "post-Skinnerian" attempt to assimilate some of the main tenets of classical CBT within a functional analytic view of psychotherapy.

[10] We'll use "functional analytic" rather than "behavior analytic" to englobe the different approaches discussed here. This is mainly due to the explicit separation of some of these approaches from "traditional" behavior analysis (see below) (Hayes, 2016).

1.5.2.1. Functional Behavioral Assessment-based interventions

As we saw in section 1.3.1., the core assessment tool of applied behavior analysis is the functional analysis, which allows for the identification of the contextual variables controlling a person's behavior (Peterson & Neef, 2019). Functional analysis can thus be understood as a specific behavior analytic methodology for conducting what is now known as a "case formulation", i.e., "a hypothesis about the causes, precipitants, and maintaining influences of a person's (…) problems" (Eells, 2007, p. 4; see also Froxán-Parga, 2020; Froxán-Parga et al., 2019). However, the truth is that many early behavior analytic procedures were implemented in the absence of a particular hypothesis about the functions of target behaviors; rather intervention was typically reduced to "superimposing powerful arbitrary contingencies of reinforcement or punishment over existing but often unknown sources of reinforcement for problem behavior" (Hanley et al., 2003, p. 147).

In this sense, one of the major contemporary advances within the functional analytic tradition has been the systematization of functional analysis procedures (Iwata et al., 1982/1994, 1994; see Beavers et al., 2013; Froxán-Parga, 2020; Hanley et al., 2003; Hurl et al., 2016; Peterson & Neef, 2019). Now referred to as "Functional Behavioral Assessment", FBA methods comprise a variety of procedures for conducting "a pretreatment ideographic set of assessments which aim is to identify variables associated with the occurrence of a specific behavior, in order to develop an idiosyncratic intervention aimed at promoting behavioral changes" (Froxán-Parga et al., 2019, p. 1). The systematization of these methods of assessment was initially due to Iwata et al.'s (1982/1994, 1994) work on self-injury behaviors, where they first offered an experimental method for conducting behavioral assessments. Current literature distinguishes among three main kinds of FBA methods: indirect, descriptive, and experimental. Indirect FBA typically draws from interviews or questionnaires to establish the functional hypothesis guiding the intervention. Descriptive FBA, by contrast, entails the direct observation of the target behavior in its natural context. Finally, experimental FBA (which is also called "functional analysis" proper), involve the systematic manipulation of "antecedents and consequences to the target behavior, usually in a single-subject reversal or replication design, in order to identify social and non-social factors that may be influencing the target behavior" (Hurl et al., 2016, p. 73). Many FBA-based interventions combine several methods, with indirect or descriptive methods typically subserving as supplementary methods to establish the key hypotheses to be experimentally tested (see Beavers et al., 2013; Hanley et al., 2003; Froxán-Parga et al., 2019).

Although initially applied in the intervention with people with self-injury behaviors, these methods have been progressively applied to an increasing number of target behaviors,

from aggressive and stereotypical behaviors to sleep disturbances, rumination, or atypical verbalizations in people with psychotic experiences (Beavers et al., 2013; see Froján-Parga et al., 2019). In addition, when compared with interventions not based on a pre-treatment FBA, FBA-based interventions displayed significantly larger effect sizes on the reduction of problem behaviors, as well as displaying (non-significant) larger effect sizes on the increase of alternative ones (Hurl et al., 2016).

Notwithstanding the importance and therapeutic potential of FBA methodologies, the truth is that their use has been largely restricted to very specific issues -typically, problem behaviors related to developmental problematics (Beavers et al., 2013; Hayes, 2021). This is probably due to the fact that FBA methods have been more commonly employed within a "traditional" behavior analytic approach, which has a historical preference for highly controlled settings and easily operationalizable problems (Hayes, 2016; Zettle & Hayes, 1982). This has eventually leaded to a relative withdrawal of traditional behavior analytic approaches from the field of psychotherapy (i.e., "talk therapy") mostly implemented with verbally competent adults in less controlled, outpatient contexts. As we'll later see, this neglect of psychotherapy by traditional behavior analytic approaches is one of the main reasons why classical CBT has typically dominated the field (Hayes, 2004), and why some functional analytic researchers have eventually departed from "traditional" behavior analysis (Hayes, 2016, 2021).

1.5.2.2. Verbal and complex behavior: the birth of third-wave behavior therapy

The second major advancement of contemporary functional analytic approaches was the development of behavior analytic research on verbal and complex (e.g., symbolic) behavior. Skinner (1945, 1957, 1969) had already emphasized the need to focus on verbal behavior in order to understand complex human behaviors (e.g., scientific practices themselves). However, the major influence on the clinical field has come from research on the phenomenon of *stimulus equivalence* and the formation of *equivalence relations* (Sidman, 2009). In the formation of equivalence relations, certain stimuli acquire the functions (i.e., the particular effect on behavior) of other stimuli -what is known as "transfer of functions" (see also Pilgrim, 2019). So-called "derived relations" are important here, for they imply a transfer of functions between stimuli which haven't been explicitly paired or whose "matching" hasn't been explicitly trained (Barnes-Holmes et al., 2004; Hayes et al., 2001; Pilgrim, 2019; see also Alonso-Vega, 2021). For instance, after training a child to a) choose the image of a guitar (among other *comparison stimuli*) when presented with a guitar sound, and b) choose the written word "guitar" when presented with the image of a guitar, a number of non-trained or derived equivalence relations might emerge: the kid might spontaneously choose the guitar image

when presented with the same guitar image (i.e., *reflexivity*), choose the guitar image when presented with the word "guitar" (i.e., *symmetry*), and choose the word "guitar" when presented with a guitar sound (i.e., *transitivity*) (see Sidman, 2009).

Subsequent research on *nonequivalence relations* is also important here (Barnes-Holmes et al., 2004; Critchfield & Rehfeldt, 2019; Hayes et al., 2001; see Alonso-Vega, 2021). In a nutshell, nonequivalent relations are "those in which stimuli are related on some basis other than "sameness"", which are "a big part of how people make sense of, and function effectively in, the world around them" (Critchfield & Rehfeldt, 2019, p. 541). Examples of nonequivalence relations include comparison (e.g., "greater than") relations, opposition (e.g., "opposite than") relations, or deictic (e.g., "I-you") relations, which "specify a relation in terms of the perspective of the speaker" (Hayes et al., 2001, p. 38). Given that some of the terms used to describe equivalence relations -namely, "transfer of functions", "symmetry", and "transitivity"- no longer describe nonequivalence relations accurately, these are often replaced by "transformation of functions", "mutual entailment", and "combinatorial entailment", respectively (Hayes et al., 2001).

Research on stimulus relations is key to understand certain characteristics of so-called *rule-governed behavior* (i.e., behavior that is causally controlled by verbal rules) vs. *contingency-shaped behavior* (i.e., behavior that is maintained by ongoing environmental contingencies) (Skinner, 1969; see also Hayes et al., 2001; Kohlenberg & Tsai, 1991; Zettle & Hayes, 1982). Rules, in the behavior analytic tradition, are descriptions of contingencies, i.e., of relations between an individual's responses and certain actual or possible consequences (Skinner, 1969). Often -though not necessarily- formed after being exposed to such contingencies, these rules can come to exert a control function over behavior, sometimes to the point to which behavior becomes "insensitive" to actual operating contingencies (Zettle & Hayes, 1982). Overall, the interest in these phenomena -equivalence and nonequivalence relations, rule-governed behavior, etc.- lies in that it's commonly taken to provide a functional analytic understanding of core characteristics of human cognition and language, e.g., meaning and referentiality, inferential and logical connections among events, concept formation, or, critically for our discussion in upcoming chapters, intentional or *norm-following* behavior (Barnes-Holmes et al., 2004; Hayes et al., 2001; although see Tonneau, 2001).

These implications are of paramount importance for functional analytic approaches to clinical practice; in particular, they've played a major role in the development of third-wave behavior therapy, which encompasses a number of approaches to clinical psychology that emerged during the 1990's (e.g., Hayes et al., 1999; Kohlenberg & Tsai, 1991; see also Pérez-Álvarez, 2012). According to Hayes (2004), although "no one factor unites these new

methods", all of them "[emphasize] such issues as acceptance, mindfulness, cognitive defusion, dialectics, values, spirituality, and relationship [...]; their underlying philosophies are more contextualistic than mechanistic" (p. 640). Here we'll focus on one of the most renown third-wave approaches: Acceptance and Commitment Therapy or ACT (Hayes et al., 1999), primarily characterized by its attempt to reassimilate cognitivist explanations within a contextualist approach, rooted in a particular understanding of Skinner's radical behaviorism and behavior analysis (Hayes, 2004; Hayes et al., 2001; Hayes, 2016, 2021).

To understand the differences between more traditional forms of behavior analysis and contextual therapies like ACT, one must first understand the different contexts where they've been traditionally applied. As we've mentioned, more traditional forms of behavior analysis have been largely applied in contexts where there's a maximum degree of control, and where it's possible to directly manipulate the contingencies controlling a person's behavior -that's why early applied behavior analysis was largely conducted in inpatient settings (Zettle & Hayes, 1982). That normally allows practitioners to focus on target behaviors themselves and not their potential sources of verbal control. By contrast, third-wave therapies like ACT are mainly approaches to psychotherapy, typically implemented in outpatient settings, with verbally competent adults, where treatment is crucially based on the therapist-client *verbal* interaction. This is typically conceived of as a subfield of applied behavior analysis: *clinical behavior analysis* (see Dougher, 2004; Dougher & Hayes, 2004; Guinther & Dougher, 2013; Follette et al., 1996; Kohlenberg et al., 1993, 2002; Madden et al., 2016), whose main research question is the following: how is it possible that psychotherapy, which is primarily conducted through verbal means, can promote the maintenance and transfer of *in-session* clinical changes to *extra-clinical*, daily life contexts (Kohlenberg et al., 1993, p. 271)? Recall that this was one of the "explanatory gaps" in S-R psychology that conduced many first-wave therapists to endorse cognitive models of psychopathology. The "talk therapy" problem, therefore, had a pivotal role in the assimilation of behavior therapy into CBT (Hayes, 2004; Kohlenberg et al., 1991). A central research goal within clinical behavior analysis has thus been to understand, from a functional analytic, non-cognitivist point of view, *why* people change through psychotherapy, whether it's conducted by one or another kind of therapist. To answer that question, many clinical behavior analysts, to a lesser or greater extent, retain the cognitivist idea that cognition, at least sometimes, causes behavior; the difference lies in their reconceptualization of cognition in behavioral terms, and thus of cognitive-behavior causal links as behavior-behavior relations (e.g., rule-governed behaviors) (Hayes, 2016; Zettle & Hayes, 1982).

Now, different approaches within clinical behavior analysis differ as to how exactly reconceptualize cognitive phenomena in behavioral terms. Drawing from a "traditional" behavior analytic point of view, some approaches have attempted to address the "talk therapy problem" by relying on Pavlovian and operant conditioning principles, as traditionally defined. An example of such an approach is provided by the work of Froxán-Parga and collaborators (e.g., Alonso-Vega, 2021; Alonso-Vega et al., 2019; Calero-Elvira et al., 2013; Froján-Parga, 2011; Froján-Parga et al., 2006, 2008, 2010a, 2016, 2017, 2018, 2019, 2020; Montaño-Fidalgo et al., 2013; Pascual-Verdú et al., 2019; Pereira et al., 2019; Ruiz-Sancho et al., 2015). Drawing primarily from conceptual and non-experimental (e.g., observational) empirical methods, these researchers have attempted to explain both characteristic cognitive change processes (e.g., cognitive restructuring), as well as general therapeutic changes achieved through the verbal interaction in therapy, by hypothesizing the combined occurrence of operant and Pavlovian processes. Examples of these are: a) the verbal reinforcement of pro-therapeutic verbalizations and verbal punishing of anti-therapeutic verbalizations (e.g., Froján-Parga et al., 2016; Ruiz-Sancho et al., 2015); b) the shaping and chaining of verbal behavior towards target self-supporting verbalizations (e.g., Calero-Elvira et al., 2013; Froján-Parga et al., 2018); c) the use of motivating operations to increase the probability of occurrence of pro-therapeutic verbalizations (e.g., Froján-Parga et al, 2010b); d) the reinforcement of in-session instruction-following and the description of instruction-following in extra-clinical contexts (de Pascual & Trujillo, 2018); or e) the employment of Pavlovian pairings to maximize the probability of occurrence of new behaviors in extra-clinical contexts (Froján-Parga et al., 2017; Pereira et al., 2019).

In this sense, this line of research is in strict continuity with the tradition of behavior analysis. Other clinical behavior analysts (e.g., ACT proponents) by contrast, have emphasized the need to revise this tradition in order to provide a proper explanation of human language and cognition, as well as related clinical phenomena (Hayes, 2016, 2021; Hayes et al., 2001). This has eventually leaded to the above-mentioned schism within behavior analysis, with ACT advocates now endorsing *functional contextualism* -a version of radical behaviorism- and *Contextual Behavioral Science* (CBS), instead of traditional behavior analysis, as their philosophical and scientific frameworks, respectively (see Hayes, 2016, 2021; Hayes et al., 2012; see also Zettle et al., 2016). Since these differences will become relevant in upcoming chapters (see Chapter 2, sections 2.3.2., 2.3.5., and Chapter 8), we'll here review them briefly.

In a nutshell, functional contextualists share with "traditional" radical behaviorists their functionalist and selectionist view of behavior as the result of natural, operant, and cultural selection processes; however, functional contextualism makes some core

commitments explicit: a) it explicitly rejects mechanistic readings of Skinner's work on verbal behavior and private events (which they claim to drive some strands within traditional behavior analysis), endorsing instead a contextualist reading; b) it explicitly establishes prediction and control[11] as its pre-analytic goals; and c) it explicitly endorses a "pragmatic truth-criterion", which they read from American pragmatists, according to which truth is equated to "effective action" (as measured by their own pre-analytic standards; Hayes, 2016, 2021; Hayes et al., 2012; Vilardaga et al., 2009). CBS is grounded on these assumptions. It can be roughly described as a multi-level research framework characterized by its "willingness to create parallel language conventions for different analytic purposes, greater methodological flexibility, a refined perspective about the role of theory, and a pragmatic approach to treatment testing" (Vilardaga et al., 2009, p. 108), characteristics which have allowed researchers within this tradition to endorse a pragmatic attitude towards theorization about and employment of *middle-level* terms (see below), as well as traditional meta-analytic techniques to assess therapeutic efficacy (see Chapter 8).

Core to CBS is its "post-Skinnerian" theory of human cognition and language, RFT, which attempts to explain complex behavior (e.g., symbolic behavior, concept formation, etc.) in terms of the above-mentioned equivalence and non-equivalence relations (here named "relational frames") (Barnes-Holmes et al., 2001; Hayes et al., 2001). In particular, RFT aims to explain these derived relations not as a mere result, but as a particular kind of operant itself: *relational responding* (i.e., the individual's responding to an event in terms of its relations to other events) and more specifically, *arbitrarily applicable relational responding*, i.e., responding to events in terms of other events with which they bear no physical or structural resemblance, but with which they're related through social convention (e.g., relations between signs and referents). Defined as a "learned overarching behavioral class" -like generalized imitation, for instance (Hayes et al., 1999, p. 40)-, arbitrarily applicable relational responding is thought to capture human's direct experience of the world as filled with social *meanings* (behaviorally defined). Take, for instance, the following example by Hayes et al. (1999):

> It is worth noting that, defined in this way, most human behavior is verbal, at least to a degree. If we look at a tree and see a T-R-E-E, a "plant" that "photosynthesizes" and has particular "cell structures" and so on—then the tree is functioning as a verbal stimulus for the observer.

[11] Functional contextualists prefer the term "influence" over "control" for a number of reasons (see Vilardaga et al., 2009). Here we'll use them as synonyms, roughly indicating the selective and instantiating effects of variations in environmental conditions upon an organism's behavior.

> It is hard for humans to avoid the derived nature of stimulus functions in their world, because even "nonverbal" stimuli quickly become verbal in part when they enter into relational frames. (Hayes et al., 1999, p. 43)

Arbitrarily applicable relational responding, which RFT theorists take to be widespread in the human species, itself forces the adoption of the explicitly contextualist framework that characterizes functional contextualism; on this view, the "stimuli" aren't raw sensations nor given objects "out there" -functional contextualists explicitly avoid strong realist commitments (see, for instance, Barnes-Holmes, 2000); rather, "stimulus" and "response" are functional categories that can be flexibly applied by the analyst and whose successful applicability will be determined by the success of the analysis, measured in terms of its predictive and controlling abilities (i.e., the pragmatic truth-criterion) (Hayes, 2021).

ACT applies all of the above to the clinical field (Hayes et al., 1999). ACT differs from other strands of clinical behavior analysis in that it not only assumes that cognition (behaviorally defined) *sometimes* has an important role in psychotherapy; rather, it endorses second-wave behavior therapists' assumption about the causal *primacy* of cognition, although reconceptualized in contextualistic, relational terms (Hayes, 2004). For ACT, "mental representations" or "interpretations of reality" (see section 1.3.2.) just amount to responding to certain events in verbal, relational terms, i.e., in terms of other events with which they stand in functional relation (Hayes et al., 1999). These relations are commonly reflected in verbal rules, which may come to control behavior in spite of the actual contingencies operating in the environment. In sum, mental representations and interpretations of reality aren't taken to be internal objects mediating world and action; rather, they point to the socially-mediated, intrinsically *relational* experience of the world that humans (at least) enjoy.

Or suffer. For ACT's key idea is that relational behavior "gone-wrong" (i.e., self-defeating) and inflexible verbal rules are often at the core of psychological problems. In particular, it hypothesizes that a great deal of psychological suffering is maintained by people's active attempt to escape or avoid noxious experiences through a variety of ways, including verbal (i.e., relational) ways (e.g., rationalizing bad experiences). Known since the first-wave, such attempts to escape often have a backfiring effect: through negative reinforcement loops, they actually increase suffering and debilitate the person's ability to try other coping strategies. ACT conceptualizes this in terms of *experiential avoidance*. According to ACT, people often engage in inflexible coping strategies to escape noxious experiences; one of them is the so-called *cognitive fusion* with one's thoughts, i.e., one's relational responding to one's thoughts as if they were *literally* true (i.e., as if they "represented reality"). These and

other behavioral patterns are subsumed under the concept of *psychological inflexibility*. ACT is thus primarily directed at targeting the inflexible coping strategies involved in experiential avoidance and teaching the person other more flexible ways to cope with one's experiences.

To do so, ACT employs a variety of methods. First, the person is trained in a variety of *acceptance* and contemplative exercises (e.g., mindfulness), whose aim is to teach the person not to immediately react (relationally) to -and thus escape from- noxious experiences. Since cognitive fusion is taken to be one core avoidance strategy, ACT makes use of its core "pragmatic truth-criterion" to promote *cognitive defusion*, training the person not to take their own thoughts at face value and to value them in terms of their utility. In particular, this utility is measured against the person's *values* (i.e., their "pre-analytic" goals, so to speak), which are made explicit though therapy and establish the life-horizons that the person acquires a *commitment* with. Intervention efficacy is thus not to be measured in terms of overall "problem behavior reduction", nor in terms of changes in one's "cognitive contents"; rather focuses on altering the *function* of the person's relation to their own noxious thoughts and experiences, and tailors expected behavioral outcomes to each person's values.

Experiential avoidance, cognitive fusion, commitment (or lack thereof); all these terms are examples of what in ACT literature is known as *middle-level terms*, i.e., "theoretically-specific, non-technical term[s] that [have] not been generated within basic scientific research" and which lie "on a continuum between the analytic units of the basic science (of psychology) and folk-psychological terms (e.g., emotion, memory, stress, etc.)" (Barnes-Holmes et al., 2016, p. 367). Once again, ACT adopts a pragmatic attitude towards middle-level terms; insofar as they subserve the goals of prediction and control, let's keep them. However, many of these middle-level terms, as well as the *hexaflex* model of psychopathology they've given rise to (Wilson, 2007, cited in Froxán-Parga et al., 2020) have recently come under revision, precisely for their potential utility problems in certain contexts (Barnes-Holmes et al., 2016; see also Assaz et al., 2018). A related concern is that these middle-level terms and the hexaflex model is that they're often employed as a form of psychological assessment, thus virtually replacing the functional assessment of behavior as the primary assessment tool (Froxán-Parga et al., 2020). This, as we'll exemplify in Chapter 8, might be due to a certain residual cognitivist commitment, which may have detrimental effects in therapy.

So far, we've seen what contemporary psychological approaches look like. Despite the differences among them, they all share a core assumption: that mental health problems must be primarily understood in non-cognitivist, functional analytic terms, i.e., in terms of the person's interaction with the environment. In this sense, these models oppose the neuro-centric tendencies of third-wave biological psychiatry and assume the causal and

conceptual priority of the psychological scale of analysis, understood in functional and selectionist terms. For reasons that we'll delve into in the following chapters, we think that, although subject to several objections (common to other approaches reviewed here), these approaches offer the best currently available framework for mental health research and practice. In addition to their functionalist and contextualist approach to mental health problems, we're particularly attracted by some implications of these approaches regarding the analogy and boundary problems, namely, that a) psychological and somatic health problems are *disanalogous*, because the former only appear when actions are viewed through contextual lenses; and b) the boundary between social deviancy and mental health problems lies primarily in that the latter entail a primarily self-defeating character (i.e., a "going against one's values" character) (e.g., see Hayes et al., 1999; González-Pardo & Pérez-Álvarez, 2007) (see Chapter 9, section 9.2.1.). It's perhaps less clear which, if any, is the integrative project of these approaches; although some of them explicitly endorse a multi-level perspective (e.g., CBS; Vilardaga et al., 2009), these approaches are more invested in vindicating the conceptual and explanatory primacy of the behavioral scale of analysis rather than offering a detailed integrative framework. In the next and final section, we'll focus on a recent kind of approach, which shares with functional analytic models its contextual perspective, but which nonetheless is primarily directed at the integration among different scales of analysis: the enactive approach to psychiatry.

1.5.3. The enactive approach to psychiatry

In section 1.4., we saw how Engel's (1977, 1978, 1980, 1997) biopsychosocial model constituted the first attempt to provide a multi-level framework for research and intervention on mental health problems. It emerged as a conciliatory enterprise, aimed at putting an end to the theoretical struggles of the previous decade over the proper level of analysis for the study of psychopathology. However, its vague eclecticism led to the integration problem.

More recently, *postcognitivist* approaches to mental health have emerged as an attempt to provide a new integrative framework for the different disciplines working in the field of mental health research and practice (Cooper, 2017; de Haan, 2020a, 2020b, 2020c, 2021; de Jaegher, 2013; Dings, 2020; Drayson, 2009; Fuchs, 2007. 2009; Glackin et al., 2021; Hoffman, 2016; Krueger, 2020, 2021; Krueger & Colombetti, 2018; Krueger & Maiese, 2018; Nielsen, 2021; Nielsen & Ward, 2018, 2020; Roberts et al., 2019; Röhricht et al., 2014; Sneddon, 2002; Sprevak, 2011). The roots of these postcognitivist approaches lie in so-called "4E Cognition" or "5E Cognition" approaches to cognitive science (see Chapter 2, section 2.2.2.1.). Although varied in their philosophical frameworks (Newen et al., 2018), these approaches reject "traditional cognitivist" accounts (Menary, 2010) or "sandwich models" (Hurley, 2001)

of the perception-cognition-action triad; instead, they aim to understand cognition in terms of its *embodied*, *extended*, *enactive*, or *embedded* character (hence the "4E"[12]), i.e., its causal or constitutive dependence from a) the body, taken as a whole, and not just the brain; b) resources external to the individual (such as computing devices, notebooks, etc.); c) the organism's ongoing actions and active sense-making; and d) the natural and social environment with which an organism interacts and in which it lives. Sometimes a fifth "E" is added to emphasize the historical importance of Gibson's (1979/2015) *ecological* psychology, which predated by more than 20 years the apparition of contemporary 4E approaches (see Heras-Escribano, 2019).

Drawing from foundational works by Sneddon (2002), Drayson (2009), and Fuchs (2007, 2009) -as well as from more distant sources like Laing's existential-phenomenological approach (see section 1.2.)- postcognitivist approaches to mental health reject the "neuro-reductionist" tendencies of biological psychiatry and emphasize the embodied, extended, enactive, embedded, or ecological character of mental health problems. Although some of them still draw from the functionalist (see Chapter 2, section 2.2.2.1.) and computationalist framework that characterizes traditional cognitive science (e.g., classical extended cognition approaches; see Hoffman, 2016), others advocate for a more radical departure. It's the case of the recently articulated *enactivist* approaches to mental health, which emphasize the enacted, embodied, and embedded nature of mental health problems (e.g., de Haan, 2020a, 2020b, 2020c, 2021; de Jaegher, 2013; Nielsen, 2020, 2021a, 2021b; Nielsen & Ward, 2018, 2020). These approaches explicitly aim to develop a properly integrative framework for mental health theory and practice, which is able to solve the integration problem of the biopsychosocial model. Although there're also important differences between them (see de Haan, 2021, Nielsen, 2021), here we'll illustrate the approach focusing on de Haan's (2020a) enactive approach to psychiatry.

The main goal of de Haan's (2020a, 2020b, 2020c, 2021) enactive approach is to solve the integration problem. She rejects both mind-body dualism as well as the "neuro-reductionist" assumptions of the new and old biomedical models. On the other hand, although she appreciates the integrational project of the biopsychosocial model, she founds Engel's approach wanting for its relative lack of detail as to how exactly each scale of analysis relates to each other. In addition, de Haan (2020a) stresses the importance of considering the *existential* dimension of psychopathology (i.e., how a person relates to their own experiences) in

[12] Due to important philosophical differences between some of these approaches (e.g., radically enactive vs. extended cognition approaches; see Newen et al., 2018) some authors subtract some "E's" to the "4E" formula (e.g., Nielsen & Ward, 2018).

a separate way, which the biopsychosocial model fails to do. By contrast, de Haan distinguishes the existential dimension, and offers a detailed *emergentist* account of how all scales of analysis relate, from the molecular to the psychological, existential, and social (see Chapter 2, sections 2.2.2.1. and 2.3.6.). In this model, the dynamics of each scale of analysis require to be studied with scale-specific explanatory tools, without prejudice to the interdisciplinary study of the *upward* and *downward* causal links among levels. This way, the enactivist model aims to provide a way out of both reductionism and unqualified holism; to be integrative without being eclectic.

Despite its multi-level character in explanatory terms, the enactive approach to psychiatry does prioritize the level of the interactions between and organism and the environment when it comes to *conceptualize* mental health problems. In this sense, it goes hand in hand with functional analytic approaches to mental health (section 1.5.2.). However, while the latter often adopts a *subpersonal* approach, which primarily emphasizes the *causal* role of the social and environment on psychopathology, the enactive approach to psychiatry pays special attention to the exploration of the *existential* and *personal* experience of it. Let's see this in more detail.

De Haan's approach draws from a particular kind of enactivism: *autopoietic enactivism* (see Varela et al., 1991). The core idea behind autopoietic enactivism is the 'life-mind continuity thesis', or the idea that mentality is identical to the characteristic feature of open (i.e., living, biological) systems: their self-organizing and self-maintaining nature. In this kind of approach, there's no external, meaningless world onto which living beings project meaning via their computational, representational minds. Instead, organisms interact with their *lifeworld*: a world that is structured in a meaningful way for the organism due to its phylogenetical and ontogenetical history. Autopoietic enactivists like de Haan employ the notion of *sense-making* to point out that the meaningful structure of the lifeworld is not static nor given; on the contrary, living organisms, whose precarious life depends on their ability to detect possible resources and potential dangers in their ever-changing environment, continuously "bring forth" or enact norms of interaction that discriminate between *correct* (i.e., life-supporting) and *incorrect* (i.e., life-undermining) courses of action. In particular, de Haan (2020a) defines sense-making "as an organism's evaluative interaction with its environment", which "is an environmentally and temporally situated process that is a) essential to life, b) implies values, and c) is affective" (p. 7). It's in that sense that she claims that, "for enactivism, the central unit of analysis for understanding cognition is not an isolated individual agent, let alone its brain, but the *organism-environment-system*" (p. 7).

However, de Haan's approach is interesting because she distinguishes two kinds of sense-making: *basic sense-making*, characteristic of all living beings, and reflective or *existential sense-making*, which she views as unique to humans. She views this second kind of sense-making as constituting "a qualitative shift in the very nature of sense-making that comes from being able to reflexively relate to one's own experiences: what I call the existential stance". Specifically, she conceives of this existential stance, grounded on the ability to explicitly reflect "on oneself, one's experiences [and] one's environment" as one which transforms "the whole system to such an extent that it calls for distinguishing organism-environment from person-world interactions" (de Haan, 2020a, p. 9). In particular, this existential kind of sense-making is the reason why, for humans, the meaningfulness of the world entails not just an enacted "desire (…) for survival", but also -or even primarily- "for dignity, for living a *good* life" (de Haan, 2020a, p. 9).

Drawing from this core assumptions, de Haan conceptualizes mental health problems not as brain disorders, nor as mere patterns of behavior that are undesirable for either society or the individual, but as *problems of sense-making*, in the existential sense of the term (de Haan, 2020a, p. 11). De Haan employs this concept to draw an explicit disanalogy between mental and somatic health problems: according to the author, the difference between the former and the latter is that, although somatic health problems can produce problems of sense-making, these are only "secondary effects", while mental health problems are *primarily* characterized by the disruption of the capacity of an agent to engage in a *meaningful* way with their environment. As she puts it:

> Psychiatric disorders are thus *enacted*: they dissolve if one succeeds in changing one's way of interacting with the world. Secondary effects of somatic disorders on sense-making in contrast do not disappear by interacting with the world in a different way. As disorders of sense-making, psychiatric disorders are not of the brain, not even of the body, but of *persons*; that is, of bodily, social, and reflective beings. Persons, moreover, whom we cannot understand in isolation from their interactions with and embeddedness in their sociocultural worlds. From an enactive perspective then, if we want to understand psychiatric disorders, we should look at persons *in interaction with their specific worlds*. (De Haan, 2020a, p. 11)

We think that there's much to praise in de Haan's enactive approach. It gathers many of the central claims of the other therapeutic models that we've seen here. In line with Laing's and Szasz's critical approaches of the 1960's, it vindicates the analysis and conceptualization of mental health problems in personal terms (see section 1.2.). In addition, it adopts the multi-level perspective first vindicated by Engel (section 1.4.) and then adopted by third-

wave biological psychiatry (section 1.5.1.), although within a more detailed theoretical framework than the one provided by the former and explicitly avoiding the latter's commitment to neuro-reductionism (yet without neglecting the importance of the neurobiological factors). Finally, it explicitly adopts what clearly looks like the kind of functionalist, selectionist, and contextualist framework endorsed by functional analytic approaches (see sections 1.3.1. and 1.5.2.); in this sense, de Haan's claim that mental health problems "dissolve *if one succeeds in changing one's way of interacting with the world*" (de Haan, 2020a, p. 11; emphasis added) points to a possibly fruitful partnership between enactivism and functional analytic approaches (for a similar remark, see Barrett, 2015, 2019).

Despite our sympathy for this kind of approach, however, we think that there's still a core problem with this kind of approach; one which is in fact shared by all the approaches that we've seen here so far. The problem is that, ultimately, all of them fail to *properly* distinguish between the subpersonal and personal realms of analysis. As already Szasz, Laing, and other critical mental health theorists pointed out, the *mental* aspect of mental disorders can only be properly understood in personal terms (i.e., in terms of agency, meaning, intentionality, etc.). Many of the approaches reviewed here seem to eschew questions about agency and meaning altogether (e.g., second-wave and third-wave biological psychiatry, more traditional functional analytic approaches). Others do address these questions (e.g., the biopsychosocial model, functional contextualism, enactivism), but ultimately conflate them with their *causal* -and hence subpersonal- accounts of human affairs, no matter whether these are spelled out in multi-level, relational, or sense-making terms.

Two core issues are at stake here: *the problem of mind*, which itself comprises a series of problems regarding the relation between mind and body, mind and world, and mind and language; and *the problem of normativity*, which refers to the problem of the place of values and norms in the contemporary scientific worldview. As we view it, these two core issues traverse the four major topics of the many debates between competing therapeutic models: the analogy, boundary, priority, and integration problems. Analogies between mental and somatic disorders are often drawn on the grounds of the presumed identity between mind and body; disanalogies often stem from its rejection. Boundaries are set between disorder, normalcy, and deviancy, which ultimately come down to norms of one or another type. Priority wars between neuro-reductivists and environmentalists begin on account of the different scales of analysis where the mental and its normative features are searched for, or where their indispensability is contested. And solutions to the puzzles of integration stand or fall on whether minds and their normative features can be properly accounted for. As Walter (2013) has observed, these two problems arise in the field of mental health in relation

to discussions on the two main aspects of the notion of "mental disorder": a) the "mental" aspect, related to the presumed ontological and explanatory status of mental states and processes in each therapeutic model; and b) the "disorder" aspect, related to the conditions that each approach establishes for something to count as "pathological". In Chapter 2, we'll delve into these problems, pointing to their origins in the Cartesian theory of mind.

1.6. Conclusion

In this chapter, we've sketched out the history of the different therapeutic models that have been most widely discussed since the second half of the 20th century. We've approached this history through the lenses of the four major themes of the philosophy of mental health: the analogy problem, the boundary problem, the priority problem, and the integration problem.

We've begun with what seems for many to be the prevailing therapeutic model: the medical model, distinguishing between its minimal and strong interpretations (Murphy, 2009). The minimal interpretation, adopted by official institutions until at least the last decade, is just committed to the description of mental health problems in medical terms, as "diagnostic" kinds. On the contrary, the strong interpretation (which we've identified with the biomedical model), was characteristic of second-wave biological psychiatrists, who viewed mental disorders as natural kinds, i.e., as, in essence, neurobiological disorders.

We've then seen that, during the 1960's and 1970's, a complex mix of critical approaches emerged to defy the medical model of mental health problems, both in its minimal and strong versions. Among the many criticisms raised against the medical model, Szasz's problematization of the analogy between mental and somatic disorders has probably been the most far-reaching of all. Szasz, together with authors like Laing, vindicated the personal level of analysis as the proper one to understand what he viewed as "problems in living". In addition, he and others, like Foucault, pointed to the mythical character of psychiatry's construal of its own historical origins, and questioned the legitimacy of the power relations constitutive of psychiatric practice.

Another important strand of criticism against the medical model came from behavioral approaches to clinical psychology. Following Hayes's (2004) nomenclature, we've distinguished between first-wave and second-wave behavior therapy. The first-wave comprises behavior therapy (based on S-R psychology and methodological behaviorism) and applied behavior analysis (based on behavior analysis and radical behaviorism), which vindicated the efficacy and conceptual soundness of behavior modification procedures based on the experimental paradigms of respondent and operant conditioning. As we've seen, applied behavior analysis developed into an independent approach to clinical practice which

established behavior (construed as the functional relation between an organism and the environment) as the proper subject matter of psychology, and functional analysis as the proper assessment method. Behavior therapy, by contrast, eventually merged with cognitive therapy, giving rise to second-wave behavior therapy (CBT). CBT reinstated the causal primacy of cognition in the understanding of psychopathology and behavior change.

These critical approaches ignited what we've called the "priority wars" in mental health theory and practice. In the late 1970's, Engel proposed his biopsychosocial model as an attempt to offer a new holistic medical model, which could appease the tensions between competing explanatory frameworks. Both biological and psychosocial variables were now deemed important to provide satisfactory healthcare. Although it eventually became the dominant ideology in mental health theory and practice, its lack of specificity regarding how the biological and the psychosocial relate has led to the integration problem.

Despite scientific and philosophical advances, contemporary approaches to mental health struggle with similar conceptual issues. Third-wave biological psychiatry, originated in the 1990's during the so-called "decade of the brain", has gained traction during the last decade. The RDoC and HiTOP initiatives, which constitute the spearhead of the precision medicine approach to psychiatry, have emerged to address to the reliability and validity problems of traditional nosologies. While the former provides a multi-level, yet brain-centered approach to the etiology of mental health problems, the latter adopts a hierarchical-dimensional approach to research on the statistical structure of psychopathology.

Contrary to RDoC's emphasis on the brain circuitry level, contemporary functional analytic approaches establish the organism-environment interaction as the core unit of analysis for mental health research and practice. We've highlighted two major advancements of these approaches: a) the systematization of Functional Behavioral Assessment methods; and b) the development of experimental research on verbal and complex behaviors. The latter laid the grounds for the emergence of many third-wave behavior therapies and clinical behavior analysis, which attempt to explain why in-session changes following primarily verbal interactions can transfer to extra-clinical contexts. In this sense, clinical behavior analysts attempt to reconceptualize classical cognitive constructs (e.g., mental representation) in behavioral terms. More "traditional" behavior analytic approaches deny the causal primacy of cognitive behavior, and reinterpret cognitive techniques and the verbal interaction in therapy Pavlovian and operant terms, as traditionally defined. By contrast, post-Skinnerian strands within third-wave behavior therapy (i.e., Acceptance and Commitment Therapy) assume the causal primacy of cognition, although reconceptualize it in terms of arbitrarily applicable relational responding.

Most recently, postcognitivist approaches to mental health have emphasized the embodied, embedded, enactive, or extended character of mental health problems. We've focused on de Haan's enactive approach to psychiatry, whose main goal is to provide an integrative framework for mental healthcare. This enactive approach draws from the life-mind continuity thesis, which views living creatures as mental creatures, whose core feature lies in their sense-making interactions with the world. In particular, de Haan emphasizes the existential dimension of sense-making abilities in the case of humans, which leads to a shift from the "organism-environment system" to the "person-world system" as the relevant unit of analysis. In this sense, de Haan's enactive approach conceptualizes mental health problems as problems of sense-making, i.e., as problems in a person's capacity to find meaning in their dynamic interactions with the lifeworld.

As we've seen, the contemporary history of the mental health is a complex, shaky, and confused mix of bombastic theses, harsh antitheses, and attempted syntheses. At the end of this chapter, we've identified two major conceptual issues that seem to lie at the core of the debates among the different therapeutic models: the problem of mind and the problem of normativity. In Chapter 2, we'll delve into these two problems, paying special attention to the former. We'll also highlight the intimate link between mind and normativity, which is, from our perspective, what any appropriate philosophical framework for mental health should take into account.

Chapter 2
The mental in mental health: from ontology to semantics

Chapter 1 intended to provide a glimpse of the intricated and convoluted history of the mental health disciplines. On the one hand, their *actual history* (i.e., the history of the actual development of mental health institutions and practices, of their darker and lighter figures, as well as of the users and survivors of such institutions and practices) is a shadowy one –to say the least–-. Heroes and villains merge constantly; Enlighted chain-breakers soon turn into perverse engineers of contemporary forms of social control (Foucault, 1961/1965; Szasz, 1961/1974); charitable and open-minded leaders are revealed as unscrupulous abettors of monstruous intervention procedures (Ghaemi, 2010; Scull & Schulkin, 2009); cold-hearted, mindless behavior analysts are vindicated as deeply committed critical thinkers (Goddard, 2014). On the other hand, their *conceptual history* (i.e., the history of the different therapeutic models and the conceptual problems at the core of the debates among them) is no less confusing. It's complicated to delineate the central commitments of the different therapeutic models, for the different proposals are often brimming with conceptual lacunae, unclarifiable ambiguities, and "straw man" depictions of opposite approaches. Yet this kind of everlasting hermeneutical effort to rationally reconstruct the history of mental health theory and practice is key to promoting ethical, conceptual, and technical advances. The way different stakeholders conceptualize mental health problems impacts which methods and strategies they deem worth researching, developing, and implementing, as well as the roles, rights, and duties that each is ascribed in therapeutic settings and the broader social context (Bolton & Gillett, 2019; Fulford, et al., 2013; Lazare, 1973).

This chapter aims to contribute to this hermeneutical effort; in particular, our main goal will be to provide an overview of the philosophical commitments underpinning the different therapeutic models, as well as to provide a plausible account of the origins of such

commitments. In doing so, we hope to cast some light on some of the intricated debates that we saw in the previous chapter.

At the end of Chapter 1 we claimed that these debates revolve around two entwined issues, which can be separately considered as a matter of differential emphasis on either the problematic character of the "mental" or the "disorder" aspects of the notion of mental disorder: the problem of mind and the problem of normativity. In the field of mental health, the former typically arises in relation to questions about the scientific status of mentalistic descriptions and explanations of psychological problems. Cognitive therapists talk about the "cognitive distortions" or "irrational beliefs" that cause psychological disturbances (e.g., Beck, 1963, 1964; Ellis, 1958, 1962); cognitive neuropsychiatry is concerned with the disruptions of inner information processing mechanisms that give rise to psychiatric disorders (e.g., David & Halligan, 1996, 2000; Ellis, 1998); and, of course, contemporary diagnostic manuals describe many mental disorders in mentalistic terms, e.g., irrational beliefs, motivational flaws or consciousness alterations (APA, 2013). But what do these mental concepts amount to? From a *naturalist* point of view, how do these mental events and processes stand in relation to non-mental events and processes?

On the other hand, the problem of normativity in mental health arises in relation to the assessment and description of certain conditions *qua* disordered or pathological, which leads to questions about the place of values in mental healthcare (see de Haan, 2020a; Fulford & van Staden, 2013; Thornton, 2007; Varga, 2015, 2017; Walter, 2013). Are mental health problems pathological just in virtue of some natural fact, such as statistically deviant characteristics that cause biological disadvantages (Kendell, 1975) or reduce functional abilities (Boorse, 1975, 2014)? Or are mental health research and practice inexorably bonded to social norms, values, and conventions, as Szasz (1960, 1961/1974, 2011) and other critical thinkers pointed out (see Laing, 1960/2010; Scheff, 1966/1999; see also Fulford & van Staden, 2013; Graham, 2010b; Thornton, 2007; Varga, 2015, 2017)? If so, does this affect the scientific character of mental health practice?

In this chapter, we'll try to show how the contending therapeutic models that we saw in Chapter 1 are grounded on more general philosophical approaches to the problems of mind and normativity. Due to its central role in theoretical discussions within the field of mental health, we'll mainly focus on the former; in particular, we'll put a special emphasis on the ontological aspect of the problem of mind, i.e., the *mind-body problem*. However, the relevance of the problem of normativity will become evident when we discuss the origins of this problem and the suitability of its possible solutions.

Following a standard narrative, we'll first trace these problems back to Descartes's (1641/2008) theory of mind; after that, we'll present some contemporary approaches to the philosophy of mind developed as solutions to the mind-body problem, focusing on those which have had a direct influence on the field of mental health. Despite the relatively standard character of this narrative, here we'll approach it from a not-so-standard angle; one which, at least in discussions among mental health researchers and practitioners, is not the usual way of addressing the problem of mind. While these discussions have usually revolved around the ontological and -to a lesser extent- epistemological puzzles of Cartesianism and their impact on the conceptualization, assessment, and treatment of mental health problems (e.g., what's the metaphysical status of mind, how is it causally linked to the body, how are minds related to the world, etc.), here and in the following chapters we'll encourage a shift from ontological to primarily semantic concerns; that is, from discussions about the metaphysical status of minds to the analysis of the function and meaning of *mental-state ascriptions* (i.e., attributions of mental states to others or to oneself, such as "They believe that Soto Asa is the most charming and talented Spanish trap musician" or "I hope that Cecilio G is not dead yet"). This strategy will allow us to see more clearly what different approaches to the mental in mental health have in common, and where their problems begin.

The structure of the chapter is as follows. In section 2.1., we'll introduce the problem of mind, as it appears in the Cartesian account of the relation between the mind, the body, and the world. We'll see how this problem results precisely from Descartes's attempt to provide a solution to the problem of normativity. The Cartesian solution will be characterized in terms of its core ontological, epistemological, and semantic commitments, and we'll show how these are related to the ontological and epistemological versions of the problem of mind (i.e., the *mind-body problem* and the *mind-world problem*).

In section 2.2., we'll lay out the different traditional responses to this problem in the philosophy of mind. As we'll see, the different approaches to the mind-body problem can be divided into three broad kinds of *naturalism*: a) ontologically conservative approaches, which endorse some variety of the mind-body identity thesis, i.e., the idea that mental properties are identical to non-mental ones; b) ontologically revisionary approaches, which assume that scientific research on the natural causes of behavior should eventually reshape our ontological assumptions; and c) ontologically radical approaches, which reject all versions of the identity theory and conclude that mentalistic talk is incompatible with a scientific worldview.

In section 2.3., we'll see how the different therapeutic models can be seen as implementations in clinical practice of the different approaches outlined in section 2.2. We'll see

that, while the classical debates between supporters and detractors of the medical model were generally framed by straightforward reductivist and straightforward eliminativist assumptions, respectively, the biopsychosocial model developed its theoretical synthesis from an emergentist framework. In addition, contemporary efforts to rethink this integrative project oscillate between two kinds of discourse eliminativist approaches: those that posit the brain circuitry level as the focal unit of analysis (e.g., third-wave biological psychiatry) and those that reinstate the organism-environment relation at the root of psychopathology (e.g., post-Skinnerian third-wave behavior therapies). Finally, newly-born post-cognitivist approaches to mental health (e.g., enactivist approaches) retrieve and refine the biopsychosocial's emergentist framework to advance a truly integrative approach to mental health.

In section 2.4., we'll come back to the problem of normativity, which has been often overlooked in theoretical debates about the ontological and explanatory status of mental states and processes in the field of mental health. We'll point out that none of the naturalist approaches described in section 2.2., nor thus their implementation in mental health research and practice offer an adequate solution. The problem, pointed out by Szasz and other critical thinkers, lies in the tight connection between mind and normativity; while mental-state ascriptions enable us to assess someone's doings in normative terms (i.e., in terms of correction and incorrection, merit and demerit, etc.), purely descriptive reports of facts about a living being (e.g., its neural states, its bodily constitution, its self-organizing and adaptive dynamics, its relational behaviors, etc.) lack this normative force. As we'll develop in the following chapters, the reason why the different approaches to the mind reviewed here fail to provide a proper account of this relation is that they all draw from an implicit and often-overlooked Cartesian commitment: *descriptivism*, or the idea that mental language plays a primarily descriptive or representational role.

Finally, in section 2.5. we'll summarize the main points of the chapter and sketch out what is, in our opinion, the main challenge for providing a proper philosophical framework for mental health theory and practice.

2.1. The problem of mind

The problem of mind can be characterized as the problem of the causal or epistemic *relation* between the mental properties that we typically ascribe to ourselves and other agents, and the natural[13] (also called "physical", "material", "non-mental", etc.) properties that we ascribe

[13] In this chapter, we'll use the terms "natural" or "material" to refer to any kind of non-mental property that could be describable and explainable drawing exclusively from the resources of the natural sciences. Relatedly, we will use the term "naturalism" to refer to any stance committed to monism, materialism, and the principle of causal closure (see section 2.2.2.). Although "physical" and related terms (e.g., "physicalism") are also commonly

to ourselves, other organisms and the world around us, and which are properly described and explained by the natural sciences.

Imagine that you are, for whatever reason, confined in a cozy and charming, yet strangely wallpapered, 16 m^2 flat with your partner. In this situation, examples of natural properties, as we are using the term here, would be your aberrant, caffeine-induced morning patterns of dopaminergic activity; the electromagnetic radiation of the (scarce) sunlight coming through the window and reflecting in the bizarre wallpaper behind you; your partner's patterns of behavioral interaction with the house environment and with you; the not-so-ample physical dimensions of the room; or the chemical properties of the fried chicken stripes that you both order on Friday evenings.

On the other hand, you and you partner could also be described in terms of a manifold of possible mental properties. Although our characterizations of each other's psychological attributes are both rich and vague enough to preclude any kind of clear-cut classification, here we'll draw from a common distinction between *dispositional* mental states (e.g., intellectual capacities, propositional attitudes like beliefs, desires, intentions, expectations, etc.) and *occurrent* (i.e., phenomenological, experiential) mental states (e.g., inner speech, sensory experiences, etc.) (see Nottelmann, 2013; Villanueva, 2019; see also Ryle, 1949/2009; Wittgenstein, 1953/1958; see also Chapter 4, section 4.2.1.)[14]. Examples of the former would be the continuously unmet yet steady expectative that a sufficiently large dose of caffeine will compensate your lack of inspiration in the morning; your partner's yearning desire for Mediterranean sunlight; her sharp intelligence; your belief that you have the most enriching and fulfilling relationship that you've ever had; your shared taste for unsystematic philosophical discussions, red wine, and indie videogames; or your also shared though yet unacted intention to put those loose wooden planks of the bed base back in their place at some moment. Examples of the latter would be the sudden and overwhelming warmth in your belly when you wake up next to your partner; your partner's odd yet strangely evocative dreams; the erratic and misguiding caffeine-induced sense of conviction that you feel when you think that you have finally figured out the right structure for your PhD thesis; your partner's astonishingly effective self-regulating inner monologue; your intense and jittery excitement

employed in the literature (see Stoljar, 2021), we have decided to use "natural" and "material" in order to distance ourselves from the idea that, ultimately, only the kind of events and processes studied by physicists are *natural*, properly speaking, and other related ideas (e.g., that only physical properties *exist*, that the explanatory tools and concepts of other natural sciences should be ultimately reduced to descriptions of purely physical events, and so forth). In this sense, not only things like configurations of elementary particles, but also things like neural patterns, operant and Pavlovian conditioning processes, or evolutionary phenomena are natural.

[14] In this and the following chapters, we'll mainly discuss examples of dispositional mental states, specifically propositional attitudes, e.g., beliefs, desires, intentions, etc.

when it finally comes time to order food and dine while watching that TV series together; that painful, burning, yet addictive sensation of spiciness at the tip of your tongues when you eat those delicious chicken stripes; or your intense feeling of being at home despite being two thousand kilometers away from your hometown.

In natural language, we continuously rely on our *folk psychology* (i.e., our commonsensical, mentalizing conception of one another) to make sense of each other's doings. We try to make sense of Green's disproportionate daily caffeine intake by appealing to their *intention* to submit their PhD on time; we explain Fuchsia's nervous wandering around the living room on the grounds of their *belief* that their date will arrive soon; and we expect Turquoise to visit Jaen's cathedral at some point in the future when they express their *desire* to see a good exemplar of Spanish Renaissance architecture. We generally don't find any problem when we use this kind of vocabulary: we often discuss whether others really believe, desire, or intend to do what they claim to believe, desire, or intend to do; we reflect on whether these beliefs, desires, or intentions are properly justified; we use that mental-state ascriptions as a basis for predicting their actions, feelings, and thoughts; and we are often subject to this kind of scrutiny on the part of others (and sometimes ourselves).

However, these folk-psychological interpretative practices have motivated various philosophical puzzles, at least since Descartes articulated his peculiar theory about the nature of mind and its relation to the body and the world. Nowadays, the label "Cartesian" is often used as a throwing weapon, especially among self-proclaimed "anti-Cartesians" or derivatives (e.g., "anti-representationalists") in the literature of the philosophy of mind and the philosophy of psychology (see Pinedo-García, 2020). But what does it exactly consist in, and why is it undesirable?

As we'll now see, Cartesianism is not exactly a necessarily unified set of theses; on the contrary, one might be Cartesian in more than just one sense. Thus, in this section we'll begin by establishing a distinction between the main different theses or ideas that comprise Cartesianism, and how these are related to the problematization of our folk-psychological understanding of ourselves and one another. Specifically, we will distinguish among three core commitments of Cartesianism: a) an ontological commitment to *dualism*, which introduces the *mind-body problem* and conceives of the mind as a special kind of substance or entity; b) an epistemic commitment to *representationalism*, or the idea that our epistemic access to the world is necessarily mediated by representations of it, which is related to what has been called the *mind-world problem*; and c) a semantic commitment to *descriptivism*, which underlies the other two commitments.

2.1.1. Descartes's (relatable) angst and the existential origin of his theory of mind

To understand the scope, aim, and flaws of Descartes's theory of mind, we must turn our attention to its historical and cultural roots. His work was developed during a bursting age of scientific discoveries, which laid the foundations of the modern conception of science and of the world as a material, mechanic tapestry of facts governed by the laws of nature. In fact, Descartes himself contributed with several important scientific improvements, and he stood out for his innovative empirical and mathematical thought and his intellectual curiosity. However, as one can easily conclude from the reading of his *Metaphysical Meditations* (see Descartes, 1641/2008), Descartes was also, or even first and foremost, a devout religious person. His main concern in the *Mediations* was to find certainties where his destructive methodological skepticism could not reach; yet, his somewhat confusing and sometimes contradictory arguments were clearly driven by his desire to establish an indubitable proof of the existence of God and, relatedly, of the soul or mind; one which could resist any kind of factual or conceptual counter-argument.

But why exactly was Descartes so eager to provide such unbreakable rational foundations for the existence of God? As Gilbert Ryle (1949/2009, p. 8) puts it:

> When Galileo showed that his methods of scientific discovery were competent to provide a mechanical theory which should cover every occupant of space, Descartes found in himself two conflicting motives. As a man of scientific genius he could not but endorse the claims of mechanics, yet as a religious and moral man he could not accept, as Hobbes accepted, the discouraging rider to those claims, namely that human nature differs only in degree of complexity from clockwork. The mental could not be just a variety of the mechanical. (Ryle, 1949/2009, p. 8).

In this sense, the origin of the Cartesian theory of mind can be found in a deep, longstanding, and echoing existential worry: if the world is nothing but a deterministic, mechanic[15] arrange of natural states of affairs –including human beings and their doings– related to one another by cause-and-effect relationships and governed by the laws of nature, what place is left for mentality, morality, value and meaning? If humans are not really *free*, if there's no such thing as a free will upon which humans make their own decisions, or if it merely is an *illusion* of control, then how can we make them accountable for their actions? How can we value the moral or epistemic merits and demerits of their deeds?

[15] In the contemporary scientific worldview, strict "mechanical" views of causality are often criticized in favor of more sophisticated probabilistic views. Nonetheless, the consequences for our normative attitudes are the same.

The Cartesian philosophy and theory of mind can thus be construed as an attempt to reconcile two opposing and, in principle, conflicting views of the world that started to collide in Descartes's time: in Sellars's (1963/1999) terms, the clash between the *manifest* (commonsensical) *image* and the *scientific image* of the world (see Pinedo-García, 2014, 2020; see also McDowell, 1996; Price et al., 2013; Rorty, 1979). According to the former, the world would be populated by human *agents*: rational, free, and accountable beings whose actions are explained in *personal* terms, i.e., in terms of their beliefs, desires, expectations, intentions, feelings and other mental or intentional[16] states. On the contrary, according to the image of the world characteristic of modern natural science, the world is a complex set of natural events that are causally related to each other; here the world would be populated, among many other animal creatures, by human *subjects* (i.e., primates of the species *Homo sapiens*, in our contemporary understanding of our evolutionary status), whose behavior can be fully explained in *subpersonal* terms, i.e., in terms of different natural events and processes, such as the anatomical and functional structure of their nervous systems, the contingencies of reinforcement established by the natural and social environment, etc. (see Pinedo-García & Noble, 2008; Pinedo-García, 2014, 2020).

Therefore, the problem that Descartes was facing can be construed in more contemporary terms as follows: what conceptual space is left for mentalistic and normative (i.e., personal) explanations of behavior in the modern scientific worldview? In other words: how can we accommodate the possibility to talk about the moral, epistemic, political, logical, etc. *correctness or incorrectness* of human practices within a purely subpersonal approach to natural phenomena? Note that what is at stake here is not just the possibility of maintaining a comforting narrative of the world and ourselves -namely, one which depicts us a rational, mindful, free agents, and not mere bundles of atoms at the mercy of the laws of nature; rather, what is at stake is the very possibility of making sense of *any* kind of theoretical approach to the world and ourselves, even the most nihilistic of approaches (see Chapter 3, section 3.2.2.) (see Parent, 2013; Pinedo-García, 2014). In this sense, whatever alternative conception of the mind that we propose must deal appropriately with the problem of normativity.

[16] In philosophical discussions, the term "intentional" does not refer to a volitional or "willed" feature of such states and objects (i.e., those that are the product of our will or our intentions), but has an alternative, special meaning. This special use of the term derives from Franz Brentano's nineteenth-century characterization of the mental; according to him, the signature feature of the mental is *intentionality*, that is, the capacity of mental states "to be about, to represent, or to stand for, things, properties and states of affairs" (Jacob, 2019). In this sense, not only intentions (to act in a certain way), but *all* propositional attitudes (i.e., attitudes towards a certain propositional content, like beliefs, desires, etc.) are intentional, as well as anything that can be said to have representational capacities (e.g., language).

As already advanced, Descartes's solution to this problem was far from appropriate; in fact, his philosophy created a whole new series of enduring philosophical puzzles, which together constitute what we've called the "problem of mind". For the sake of clarity, we'll here focus on two aspects of this problem: a) its ontological aspect, related to the problem of the causal relation between mind and body (i.e., the mind-body problem); and b) its epistemological aspect, related to the problem of the epistemic relation between mind and world (i.e., the mind-world problem). After that, we'll see how these they're both anchored in a particular implicit view of the semantics of folk-psychological interpretation.

2.1.2. Cartesian ontology: substance dualism and the mind-body problem

Descartes's particular solution to the problem of normativity was to establish a difference between the *causal* mechanisms involved in the production of agential behavior (i.e., rational, free, intentional or goal-directed action) and those involved in the production of automatic, unreasoned or unfree behavior. In Ryle's (1949/2009, p. 9) terms, Descartes's theory of mind takes it that

> [t]he difference between the human behaviours which we describe as intelligent and those which we describe as unintelligent must be a difference in their causation; so, while some movements of human tongues and limbs are the effects of mechanical causes, others must be the effects of non-mechanical causes, i.e. some issue from movements of particles of matter, others from workings of the mind. (Ryle, 1949/2009, p. 9)

Specifically, Descartes built his distinction between the two kinds of causation (mental and natural) onto an *ontological* distinction between two different kinds of substances: the *res extensa* and the *res cogitans*. On the one hand, the realm of matter, of the natural world, where he placed the body and the rest of extended stuff (i.e.., describable in spatial-temporal terms) and subject to the laws of nature; on the other hand, the realm of the immaterial (the mind or soul), whose essence is pure thought and therefore lacks extension (i.e., it has temporal, but no spatial properties). It is here, in the realm of mind, where free will resides, the uncaused cause of genuinely free action.

For Ryle (1949/2009), Descartes's *substance dualism* is just the most historically salient exemplar of a hackneyed conception of mind; one that he refers to as "the dogma of the Ghost in the Machine" (p. 5); in this dogma, the mental is *reified*, i.e., conceived of in terms of entities (i.e., objects, states, processes, "happenings", etc.). Furthermore, these mental events are considered to differ in metaphysical status from natural, mechanistic ones (hence their queer and ghostly nature), yet to stand in special *causal* relations to other mental states

and to bodily and behavioral processes. The former assumption corresponds to the idea of *factualism* (i.e., that minds are some kind of thing or *res*), while the latter corresponds to the idea of *mental causalism* (i.e., that minds stand in causal relations with perception, other mental states, and action). In addition, mental causalism sets the stage for what Ryle (1949/2009) called the *intellectualist legend*, historically tied to the dogma of the Ghost in the Machine. In his own words, the intellectualist legend is "the absurd assumption [...] that a performance of any sort inherits all its title to intelligence from some anterior internal operation of planning what to do" (Ryle 1949/2009, p. 20). That's what distinguishes between intentional, autonomous action and purely automatic or mechanic reactions: that the former are preceded by the entertainment of "inner rules" or "regulative propositions" (p. 19) before the mind's eye, so to speak, while the latter are not.

These are the core ontological commitments of Cartesianism: substance dualism, mental causalism, intellectualism, and factualism (see section 1.4.). Taken together, they constitute what Ryle (1949/2009, p. 9) described as the *para-mechanical hypothesis*:

> [t]he differences between the physical and the mental were thus represented as differences inside the common framework of the categories of 'thing', 'stuff', 'attribute', 'state', 'process', 'change', 'cause' and 'effect'. Minds are things, but different sorts of things from bodies; mental processes are causes and effects, but different sorts of causes and effects from bodily movements. [...] Their theory was a para-mechanical hypothesis. [...] As thus represented, minds are not merely ghosts harnessed to machines, they are themselves just spectral machines. (Ryle, 1949/2009, p. 9)

Descartes's theory of mind was subject to intense criticism and questioning from the very beginning (see Descartes, 1641/2008). Princess Elisabeth of Bohemia (1618–1680) was among the first to formulate what we now know as the *mind-body problem* (see Kim, 2011, p. 46; Shapiro, 2007, 2021). This problem essentially consists in a series of conceptual puzzles related to the specific nature of the causal relations between mind and nature, i.e., between the nonspatial, non-extended realm of mind and the spatial, extended realm of nature (see Davidson, 1970/2001, 1991; Kim, 1993, 2011; Pinedo-García, 2014, 2020; Price et al., 2013; Ramsey, 2020; Smart, 2017; Stoljar, 2021). In her correspondence with Descartes (May 6th, 1643), Princess Elisabeth pointed out that she couldn't understand "how the soul of a human being (it being only a thinking substance) can determine the bodily spirits, in order to bring about voluntary actions" (see Saphiro, 2007, p. 61). Later (June 20th, 1643), unsatisfied with Descartes's reply, she contested that she also found "very difficult to understand that a soul, (...) being able to subsist without the body, and having nothing in common with it, (...) is still so

governed by it" (see Saphiro, 2007, p. 68). In other words: how can an immaterial substance, which is not extensional nor thus subject to the same mechanical restrictions as matter, causally affect the body and in turn be affected by it? How can our intentions, combined with our beliefs, desires and other mental states, produce (or fail to produce) our behavior of putting the wooden planks of the bed base back in their place? And how come some specific activation of our optic nerves causes our belief that there is a laptop in front of us (and a PhD dissertation waiting to be written)?

If we accept substance dualism, we need to posit a whole new bizarre metaphysical realm, different from the natural one, to explain behavior. For example, to explain why you keep drinking clearly excessive amounts of coffee in the mornings, we need to posit the existence of a ghostly mental object (i.e., the one corresponding to your expectation that an excessive intake of caffeine will inspire your writing) in your inner, non-spatial theatre of consciousness; not only that, but we also need to posit the existence of an even stranger interdimensional portal between the two worlds. The case is even worse with mental health problems. For example, delusions are typically defined in terms of somehow epistemically wrong (e.g., "irrational", "fixed") beliefs (APA, 2013, p. 87; Bortolotti, 2010) (see Chapters 5 and 6). Following Descartes, now we would not only need to posit the existence of an inner phantom that causes delusional behavior, plus an interdimensional bridge serving as a causal conductor between worlds; in addition, we would also need a whole theory of the regular, normal functioning of such para-mechanical phantom, as well as some sub-theory of how and why exactly delusions deviate from that regular functioning. Given the puzzling nature of dualism, many have felt inclined to agree with Princess Elisabeth in her claim that "it would be easier [...] to concede matter and extension to the soul than to concede the capacity to move a body and to be moved by it to an immaterial thing" (see Saphiro, 2007, p. 68).

This is the most usual line of criticism against Cartesianism, both in the philosophy of mind and, perhaps more sharply, in the conceptual debates among mental health researchers and practitioners (e.g., see Bolton & Gillett, 2019; de Haan, 2020a, 2020b, 2020c; Fulford et al., 2013; Graham, 2010; Varga, 2015; Walter, 2013). As we mentioned above, "Cartesian" is currently a shameful epithet that is often wielded forth and back among contending approaches to scientific accounts of perception, cognition and action. This is somewhat ironic, since nowadays few espouse the exact kind of ontological framework that Descartes advocated for and almost everyone would identify as a "non-Cartesian" or an "anti-Cartesian" in this sense, i.e., with regard to Descartes's substance dualism. Nonetheless, when researchers call each other out on account of the alleged Cartesian nature of their theories, they generally attempt to invoke the idea that opposing theories retain *some* (although not

necessarily all) of the theoretical commitments that characterize the Cartesian ontology. In this sense, they might be implying several different things: that the opposing theory explicitly or implicitly draws from a similarly bifid ontology (i.e., substance dualism); that it doesn't, but still retains a commitment to the idea that goal-directed or intentional action is caused by a different kind of events than reactive behavior (i.e., mental causalism and intellectualism); that it doesn't either, but still retains a commitment to the idea that minds or mental events are some kind of thing or *res* (i.e., factualism).

However, Descartes didn't assume this ontological framework as an axiom, but derived it from some more fundamental *epistemological* assumptions. These have also been widely discussed in the philosophical literature, but they have received less attention from psychiatrists and clinical psychologists interested in the conceptual underpinnings of their practice.

2.1.3. Cartesian epistemology: representationalism and the mind-world problem

In his introspective investigations of the limits of doubt and knowledge, the first basic certainty at which Descartes arrives on his steadfast application of his methodological skepticism is the Cartesian *cogito*. One can doubt the senses, and the information about the external world that comes through them; ultimately, one can even doubt that there's in fact an external world, for a maleficent genie might have placed an everlasting veil of illusions and chimeras before the eyes of our mind. However, what one can never doubt is that one thinks; thus, since thinking is assumed to be an activity exclusively characteristic of existing beings, that "thinking thing" that Descartes so "clearly and distinctly" envisaged must exist.

The Cartesian *cogito* gave a rational foundation to *representationalism*, which consists of a series of still lasting epistemological assumptions regarding our capacity to gain knowledge about our own minds, the world around us, and the minds of other people. According to this representationalist account of the mind, our intellectual activity primarily consists of a series of representational operations: partially drawing on the information provided from the senses, partially from what the mind already knows before even taking a look at reality, our minds build internal representations of the outer world. Thus, in the Cartesian epistemology, we have a *mediated* epistemic access to the world (i.e., we only know about it through our representations of it), but we have an *immediate* epistemic access to our own mental states. Not only it is immediate, but also infallible; for one may doubt whether the world in fact is as it seems to us (i.e., as it is represented in our thoughts), but one can never doubt about the very content and nature of one's own representations. The following excerpts from Descartes's *Meditations* (II and V, respectively) are fine exemplars of these epistemological assumptions:

> But what, then, am I? A thinking thing, it has been said. But what is a thinking thing? It is a thing that doubts, understands, [conceives], affirms, denies, wills, refuses; that imagines also, and perceives. [...] In fine, I am the same being who perceives, that is, who apprehends certain objects as by the organs of sense, since, in truth, I see light, hear a noise, and feel heat. But it will be said that these presentations are false, and that I am dreaming. Let it be so. At all events it is certain that I seem to see light, hear a noise, and feel heat; this cannot be false, and this is what in me is properly called perceiving (sentire), which is nothing else than thinking. (Descartes, 1641/2008, pp. 19-20)

> [...] as I have discovered what must be done and what avoided to arrive at the knowledge of truth, what I have [...] to do is to essay to emerge from the state of doubt in which I have for some time been, and to discover whether anything can be known with certainty regarding material objects. But before considering whether such objects as I conceive exist without me, I must examine their ideas in so far as these are to be found in my consciousness, and discover which of them are distinct and which confused. (Descartes, 1641/2008, p. 44)

This representationalist conception of the mind has produced a series of long-lasting philosophical puzzles, which configure the epistemological problem of Cartesianism or what we'll call here the *mind-world problem*. This problem can be analyzed into three distinct though inter-related aspects, each corresponding to what Davidson (1991) called "the three varieties of knowledge": a) knowledge of our own minds, which would entail *the problem of incorrigibility* (see Almagro-Holgado, 2021; Bar-On, 2015; Borgoni, 2019; Coliva, 2016; Curry, 2020; Davidson; 1991; Ryle, 1949/2009; Schwitzgebel, 2002, 2013, 2021; Srinivasan, 2015; Wittgenstein, 1953/1958); b) knowledge of the world around us, which would entail *the problem of the external world* (see Coliva, 2016; Davidson, 1991; Hurley, 2001; McDowell, 1994, 1996; Nöe, 2001; Pinedo-García, 2014; Pinedo-García & Noble, 2008; Srinivasan, 2020); and c) knowledge of other minds, which would entail *the problem of other minds* (see Avramides, 2020; Coliva, 2016; Davidson; 1991; Fernández-Castro & Heras-Escribano, 2020; Rorty, 1979; Ryle, 1949/2009; Tanney, 2009; Wittgenstein, 1953/1958).

To begin with, the Cartesian conception of mind adopts a strong view of what in contemporary philosophy has been called the idea of *first-person authority*, i.e., the relatively commonsensical idea that one typically is in a better position than others to tell what one thinks or how one feels about a certain issue, or that doubts about one's mental self-ascriptions (i.e., one's "reports" of one's mental states) "are unreasonable and generally misplaced", as Borgoni (2019, p. 295) puts it (see also Almagro-Holgado, 2021; Bar-On, 2015; Bar-

On & Sias, 2013; Borgoni, forthcoming; Coliva, 2016; Curry, 2020; Srinivasan, 2015). In the Cartesian framework, this idea is framed within a *privileged access* conception of self-knowledge, according to which we have a privileged, immediate, and incorrigible epistemic access to our own mental states. One can thus never be wrong about one's own mental states; or, to put it differently, one is always "one's best acquaintance". Consequently, if we sincerely claim that we believe, desire, or expect that *p*, then we in fact believe, desire, or expect that *p*.

Thus construed, the idea of first-person authority is clearly problematic. It's true that we often, or even typically, confer others an authority over their own mental states, and that we take their *avowals* (i.e., sincere mental-state self-ascriptions) at face value when we are interested in knowing what they think or how they feel about certain issue. Furthermore, as we'll point out in Chapter 6 (see section 6.3.2.), we agree with those who think that this is what we *should* do in many, if not most cases (see Borgoni, 2019, forthcoming). But we can easily think of many examples where we don't (and shouldn't). For instance, we might confidently and honestly assert that, contrary to others', our experience as PhD students is being quite serene and peaceful, while being noxiously blind to the fact that we're having sudden panic attacks and prompts of hysterical laughter on a regular basis. We might also be completely honest when we say that we want people from all races to be treated equally, or that we believe that women and men are equally suited for whatever intellectual task, yet still display unnoticed racist and sexist behaviors. In these cases, others might perfectly be in a better position than us to know what's "going on in our minds", i.e., what we *really* believe, desire, expect, intend to do, etc. (see Almagro-Holgado, 2021; Curry, 2020; see also Borgoni, 2014 Coliva, 2016; Schwitzgebel, 2013, 2021; Srinivasan, 2015); and, if they care for us, they would do well to try to correct our mistakes.

Another important line of criticism against representationalism has focused on the problem of the external world. Recall that, for Descartes, we have an immediate and infallible access to our own representations of the world, but only a mediated epistemic access to the world around us. However, if our knowledge of the world is irrevocably mediated by a veil of mental representations, how can we assess their correctness or incorrectness? Whatever criteria that we could use to test their truth or falsity would necessarily involve the manipulation of *further* representations. Thus, ultimately, Cartesianism leads us to some kind of solipsistic view of the relation between mind and world, where the former would be

endlessly trapped in an infinite regress of sanctionless representations[17] and the latter would basically be unthinkable (see also Coliva, 2016; Davidson, 1991; Hurley, 2001; McDowell, 1994, 1996; Nöe, 2001; Pinedo-García, 2014; Pinedo-García & Noble, 2008). Furthermore, as we understand it, Descartes's epistemology would imply that the mind is intelligible independently of the existence of everything else, the subject's body included.

This results in a not-so-rosy picture of our capacity to gain knowledge of the world -one which might have inspired many interesting contemporary cultural products, but which is untenable for a scientific view of the world, ourselves, and our place in nature. Yet it paints an even darker picture of our common sense, folk-psychological understanding of each other in terms of mental states. In the Cartesian view, we can only gain epistemic access to other's minds by analogy with our own minds, i.e., we infer others' mental states from their observable behavior. However, if our observation of others' behavior is necessarily mediated by our mental representations of it, attaining objective knowledge of what another person believes, desires, or intends to do from a third-person perspective would not only be impossible, but *doubly impossible*; there would be a two-step inferential route to other minds, and no external criteria that we could use to determine whether our mental representations of their mental representations are true or false (see also Almagro-Holgado, 2021; Avramides, 2020; Coliva, 2016; Davidson, 1991; Fernández-Castro & Heras-Escribano, 2019; Rorty, 1979; Ryle, 1949/2009; Tanney, 2009).

The idea that we can never really know what's "out there" in the world nor what's in others' minds departs drastically from our common practical take on each other and the world around us. One *knows* for certain (or as certain as anyone can know anything), that one is sitting in front of a jerkily-working laptop and that one is writing a PhD thesis –what kind of bitter tragicomedy would this be if that was just a ghostly play in one's "theatre of consciousness". One also knows that one's partner eagerly wants to eat chicken strips tonight, that she believes that it's already ten o'clock in the evening, and that she thinks that it's one's fault that we aren't having chicken strips for dinner because one has forgotten to call the restaurant in time.

As we have seen, Cartesianism yields not only bizarre ontological puzzles, but also epistemological ones. Now, in order to explain an agent's behavior -whether its rational or irrational, moral or immoral, clinical or non-clinical, etc.- we would not only need to posit the existence of a wholly different, inapprehensible, and unearthly kind of substance,

[17] For Descartes, God's benevolent nature was the main warrant that an outer world actually existed and that our mental representations of it were typically true. However, this isn't a legitimate argument in our contemporary, secular view of nature.

together with a theory of its normal and anomalous functioning; in addition, we would have no way to test whether our theory is correct or not. For how could we even know whether there is something like other minds out there, whether others also have such a spectral machinery inside of them? And how could we then make sense of the idea that there might be something amiss in that machinery, as for instance in the case of mental health problems? At best, we could know our own minds, but not others. And not even: if, as we'll discuss in more detail in Chapter 3 (section 3.2.3.) knowledge entails the possibility of error (i.e., the possibility that we might have *failed* at knowing what we know), then we couldn't properly speak about "knowledge of our own minds", since, by principle, we could *never* be wrong about our own mental states (see Coliva, 2016; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014; Wittgenstein, 1953/1958).

2.1.4. The big Cartesian family

Thus far, we've seen that a theoretical framework might be Cartesian in more than just one sense. Here there are some examples of ontological and epistemological claims that would imply some sort of commitment to Cartesianism, and which might be adopted *en bloc* or not.

1. Ontological commitments.
    a. *Substance dualism*: Mind and body are two separate substances, pertaining to two separate ontological realms; while the mind is an immaterial, non-spatial substance, whose core essence is thought, the body is a material, extensional substance.
    b. *Mental causation*: Minds operate in a sort of mechanical way, and at least certain behaviors (typically, those that would qualify as goal-directed actions) are caused by mental activities.
    c. *Intellectualism*: To be in a mental state, and to act in accordance with it, is a matter of entertaining certain representations or "regulative propositions" in the mind and acting accordingly.
    d. *Factualism*: Minds are some kind of entity, thing, organ, or *res*.
2. Epistemological commitments:
    a. *Representationalism*: Minds essentially are representational systems, i.e., mechanical (or computational) devices that generate, store, retrieve, and manipulate representations of the world around us. We perceive the external world through representational lenses.

b. *Privileged self-knowledge*: One can never be wrong about one's own mental states, for one has an immediate and incorrigible access to one's own representations.
c. *Analogical knowledge of other minds:* We can only learn about the contents of other minds by analogy with ours; specifically, we observe other's behaviors and then infer which mental states may have caused them.

In the next section, we'll see some of the most widely discussed strategies to overcome the many problems of Cartesianism. Since the debates in the realm of mental health research have typically revolved around ontological issues, we will place a special emphasis in the discussion of the different strategies that have been proposed in philosophical research to overcome the mind-body problem.

As we'll see, mainstream approaches have typically tackled the problem of dualism, leaving most of the other Cartesian commitments untouched. By contrast, increasingly recognized alternative approaches have also tried to avoid the commitment to some other Cartesian ontological and epistemological tenets (namely, representationalism and the idea of mental causation). In any case, the main approaches to the scientific study of cognition, experience, and behavior (as well as their implementations in the field of mental health) commit to at least one of the above-mentioned Cartesian tenets (typically, to some subset of them).

In this sense, contemporary debates around the concept of mind, in both basic and applied contexts, should not be understood as disputes between Cartesian vs. non-Cartesian approaches; instead, they can be better construed as family disputes among more or less distant relatives (many of which are unaware of their common genealogical ties and their shared Cartesian heritage). This family disputes typically revolve around the ontological and epistemological tenets of Cartesianism (Pinedo-García, 2020). However, what we'll defend here is that, at the bottom of the Cartesian family tree, there's an implicit *semantic* commitment; namely, a *descriptivist* conception of language in general and of the meaning and function of mental statements in particular (see Almagro-Holgado, 2021; Austin, 1962; Chrisman, 2007; Frápolli & Villanueva, 2012, 2013; Heras-Escribano & Pinedo-García, 2018; Pérez-Navarro et al., 2019; Pinedo-García, 2014, 2020; Price et al., 2013; Rorty, 1979; Tanney, 2009; Villanueva, 2014, 2018, 2019). In a nutshell, descriptivism "encompasses a family of theories according to which the function of declarative sentences is to describe *facts* concerning worldly entities such as objects, properties, relations, events, etc." (Pérez-Navarro et al., 2019, p. 411). Applied to mental vocabulary, it implies the assumption that our mental-state ascriptions

and self-ascriptions describe or represent[18] some given facts (e.g., specters inside our heads, brain states or their functional organization, specific relations between an organism and its environment, etc.) -or, alternatively, that *either* they represent some fact *or* they lack truth conditions (see Chapter 3, section 3.1.2.).

We will delve into this problem in more detail in Chapter 3. However, in line with both early analytic philosophers and recent developments in the philosophy of mind and language, we would like to point out that this conception of language has significantly and recursively constrained the exploration of alternative conceptions of the mind that definitely overcome the Cartesian paradigm, confining the discussion within the conceptual limits of what Ryle (1949, p. 9) called the "the logical mould into which Descartes pressed his theory of the mind" (see also Almagro-Holgado, 2021; Almagro-Holgado & Fernández-Castro, 2019; Heras-Escribano, 2019; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014, 2020; Price et al., 2013; Rorty, 1979; Ryle, 1949/2009; Tanney, 2009; Wittgenstein, 1953/1958). In the upcoming sections, we'll be able to glimpse to what extent this seemingly intuitive conception of mental vocabulary has pervaded debates about the place of mind on nature.

2.2. The mind on nature

In this section, we'll discuss some of the main contemporary approaches to the mind-body problem. Specifically, we'll focus on those which have had a relatively direct impact on debates about the status of mentalistic explanations of mental health problems. We'll begin by introducing the common framework from which these responses usually draw. After that, we'll review some of the different ways in which contemporary philosophers of mind have attempted to address the mind-body problem from a naturalist point of view.

2.2.1. The standard image of folk psychology

As we have already mentioned, the truth is that few researchers embrace substance dualism nowadays (at least not overtly, and at least not those working in more basic or theoretical fields). By contrast, and notwithstanding its also problematic character, many still retain some other Cartesian commitments; namely, factualism, mental causalism, intellectualism, and the representationalist conception of mind. According to a still mainstream (though somewhat declining) understanding of the mental, minds are basically computational devices, functionally structured in more or less compartmentalized modules that store and

[18] That's why this view of meaning has also been called "representationalism" (e.g., Pinedo-García, 2020; Price, 2011; Price et al., 2013; Rorty, 1979). It thus constitutes a linguistic version of representationalism, different from (although tightly related to) epistemological representationalism (see section 2.1.3.). For the sake of clarity, we'll use "descriptivism" to refer to the former and leave "representationalism" to refer to the latter.

process information (Block, 1995; Block & Fodor, 1972; Carruthers, 2013; Fodor, 1983, 1987, 2006; Putnam, 1967/1975). On this account, having beliefs, desires, intentions, expectations, etc., is a matter of entertaining propositional contents before the eyes of the mind, of storing certain representations in certain "mental boxes", to use Schwitzgebel's (2013) phrase (e.g., the Belief box, the Desire box, the Intention box, etc.).

This mainstream approach typically draws from the assumption that the main purpose and function of folk psychology is to predict and control others' behavior by means of positing the inner mental operations that would causally explain it. This assumption reflects what McGeer (2007, p. 138) has called the *standard image* of folk psychology. We'll delve more deeply into it in Chapter 3 (section 3.1.1.). For now, we'll just briefly characterize it in order to understand what motivates the usual responses to the mind-body problem.

According to this standard image, our folk, unreflective understanding of each other in mentalistic terms would subserve some kind of nomological or causal-explanatory purpose. This is the basic assumption underlying the contemporary research on what has been called the Theory of Mind (hence, ToM), i.e., the capacity to interpret each other's behavior in terms of mental states, or to "read" other's minds on the grounds of their behavior (Premack & Woodruff, 1987; see also Carruthers & Smith, 1996; Westra & Carruthers, 2018); that's why this capacity has also been called *mindreading* (McGeer, 2007, 2015, 2021; Zawidzki, 2008; see also Almagro-Holgado & Fernández-Castro, 2019; Fernández-Castro 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019; Zawidzki, 2013).

Mainstream approaches to the study of this capacity come in either one of two possible flavors (see Carruthers & Smith, 1996). On the one hand, according to the *theory-theory approach* to mindreading, our capacity to interpret each other in folk-psychological terms would be due to us having some tacit proto-scientific theory; specifically, one that tells how mental states causally relate to behavior and other mental states. Thus, when we attribute mental states to each other, what we are doing is subsuming each other's behavior under implicit law-like generalizations, that we later use to predict future behavioral outcomes (Carruthers, 1996). In a different vein, *simulation theory* states that we do not rely on implicit knowledge about how mental states and behavioral outcomes causally relate, but on our own mental or cognitive states and processes. Thus, our capacity for mindreading is explained on the grounds of some kind of analogical reasoning, whereby we project ourselves to others' minds to model their mental activity and thus causally understand and predict their behavior (Gordon, 1996; see also Almagro-Holgado & Fernández-Castro, 2019; Fernández-Castro 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019; McGeer, 2007, 2015, 2021; Zawidzki, 2008).

As we mentioned at the beginning of section 2.1., this kind of approaches draw from relatively self-evident and unproblematic facts about our daily use of language: it is true that, in numerous occasions, we try to guess each other's intentions, beliefs, desires and feelings; that we sometimes guess right and sometimes we don't; that there are particularly perspicuous people when it comes to "reading other people's minds" or others' "true intentions", while others are more gullible and naiver. And it's true that this ability to "theorize" about each other's minds is of paramount importance when it comes to everyday communication and coordination in a manifold of different settings.

This everyday understanding of each other typically crosses over to more technical or professional contexts, such as the clinical setting. Here it is also common to hear somewhat special mentalistic concepts (e.g., "irrational beliefs", "repressed desires", "unusual perceptions", etc.) to explain the behavior of the users of mental health services. We could thus say that, in some sense, the default view for many people, including many mental health practitioners, is some kind of *ontologically non-committal approach* to the mental (i.e., one where no specific ontological commitments are endorsed). It is only when we think of mental states as *hypothetical constructs* (that is, in terms of entities that maintain certain causal relations with observable behavior and other mental states; see MacCorquodale & Meehl, 1948), that we seem forced to endorse a particular ontological stance; most likely, one that respects the naturalist ontological framework of contemporary science.

Thus, the urge to reconcile mentalism with naturalism results from the conflation of a series of self-evident remarks regarding our folk-psychological interpretative practices (e.g., that we often *explain* or *predict* our own and others' behavior in folk-psychological terms) into an intellectualist construal of them, according to which the exercise of such interpretative ability must be grounded on a theoretical representation of how mental states causally relate to other mental states and behavior; as a consequence, if we want to avoid the invocation of a second, ghostly, and spooky ontology, we must find a way to reconcile such intellectualist approach to our interpretative practices with the defining principles of the scientific image.

In the following sections, we'll see that some of the most common strategies in the philosophy of mind to solve the mind-body problem have attempted to develop this reconciliation in one way or another. Others, by contrast, assume that such reconciliation is either impossible or, at least, not necessary nor desirable for a properly scientific explanation of behavior.

2.2.2. Naturalisms and the mind-body identity theory

The different contemporary theoretical approaches to the problem of the ontological and causal-explanatory status of mind draw from a commitment to *ontological naturalism*. Ontological naturalism can be characterized by the defense of *monism* (i.e., the assumption that there is only one ontological framework, only one general kind of states of affairs), *materialism* (i.e., the assumption that every actual or potential state of affairs is of a scientifically describable nature) and *the principle of causal closure* of the natural world (i.e., the assumption that every state of affairs must be the effect of a natural cause) (for related -albeit sometimes different- characterizations of the basic commitments of naturalism, see Heras-Escribano & Pinedo García, 2018; Kim, 1993, 2011; Price et al., 2013; Stoljar, 2021).

Drawing from these common axioms, different approaches have attempted to provide different solutions to the mind-body problem. As we've seen in the previous section, mainstream conceptions of our folk psychology view it as some kind of *theory* or explanatory effort to causally account for behavior; consequently, we can distinguish different kinds of naturalism depending on the kind of *theory change* they propose, whereby a theory change implies a replacement of one explanatory framework for another in our scientific understanding of a given phenomenon (see Ramsey, 2020; Ramsey et al., 1990; Savitt, 1975). For example, Savitt (1975, p. 436) distinguishes between *ontologically conservative* and *ontologically radical* theory changes: while the former maintain the ontological framework of the replaced theory (see Block, 1995; Feigl, 1958; Kim, 1993, 2011; Lewis, 1966, 1980; Place, 1956, 1988; Putnam, 1967/1975; Smart, 1959, 2017;) (hence their identification with so-called *reductivist*[19] approaches to the mind-body problem), the latter dispose of it (hence their identification with so-called *eliminativist* approaches) (see P. M. Churchland, 1981; P. S. Churchland, 1986; Epstein et al., 1980, 1981; Ramsey, 2020; Ramsey et al., 1990; Rorty, 1965, 1970; Skinner, 1945, 1953, 1974, 1981, 1990).

However, as Bickle (1992) has pointed out, this binary classification obscures certain nuances concerning the different ways in which reductivist and eliminativist approaches can be implemented, which will become important in upcoming sections (see 2.2.2.2. and 2.3.5.). That's why we have decided to add a third kind of theory change: ontologically revisionary approaches -as in Bickle's (1992) "revisionary physicalism"- whereby the proposed theory

[19] The term "reductivism" has been commonly employed to refer to a more specific kind of ontologically conservative naturalism; namely, that which implements some kind of *type* identity theory (see section 2.2.2.1.). By contrast, ontologically conservative naturalisms that implement a *token* identity theory have often been called "non-reductivist" approaches. For the sake of clarity, we'll here use the term "straightforward reductivism" to refer to the former and "contextualist reductivism" to refer to the latter.

change neither results in the maintenance nor the removal of the old ontological framework, but in its progressive shaping.

These three kinds of theory changes provide a working classification of the different naturalist approaches to the philosophy of mind that have had a major impact in mental health research and practice. Firstly, ontologically conservative naturalisms try to implement some variety of the *mind-body identity theory*, i.e., a theory that identifies mental properties (e.g., mental states and processes) with natural or non-mental properties (e.g., brain states, sensorimotor contingencies, particular relations between the organism and the environment[20], etc.). This would provide a place for the mind within a naturalist view of the world. By contrast, ontologically revisionary approaches, although willing to accept that particular instances of mental properties could be reduced to their natural realizers, assume nonetheless that scientific research on the "real", natural causes of behavior may eventually reshape our ontological assumptions. Finally, ontologically radical naturalisms reject the very possibility of establishing any kind of identity theory and conclude that our folk-psychological interpretative practices are just a vestige of a common, yet mythical conception of human behavior.

2.2.2.1. Ontologically conservative naturalisms

Typically, the main goal of ontologically conservative naturalist approaches is two-fold: first, to preserve the idea that the mental concepts that we deploy in our folk-psychological understanding of one another have an actual causal-explanatory value, without the need to postulate the existence of queer and spooky entities; and second, to preserve the idea that our mental-state ascriptions are properly meaningful or *truth-apt*, and not just some fictional or illusive use of language. In this sense, ontologically conservative naturalisms consider that the mental terms that we employ in our folk-psychological explanations of behavior point to some kind of entity. Specifically, our mental terms are considered to be coreferential with (and thus translatable, or *reducible* to) exhaustive descriptions of natural events.

Consider the following sentences:

> (1) Citric and Emerald believe that left-wing people commit more crimes than right-wing people.

[20] Standard approaches establish an identity relation between mental and bodily states, typically brain states. However, the way in which we've characterized the identity theory here allows us to also cover approaches that establish an identity between mental and non-mental properties that are not strictly "bodily" (e.g., relational properties, such as those that characterize the relation between an organism and its environment), but that are nonetheless natural.

(2) Crimson and Ruby desire that companies have it easier to fire the workforce during the coronavirus crisis.

According to the mind-body identity theory, (1) and (2) should be translatable or reducible to sentences like the following:

(3) Citric and Emerald are in the brain state V.
(4) Crimson and Ruby emit –covertly or overtly– certain vocal sounds, such as "firing workers should now be easier for companies".

This seemingly allows us to account for the explanatory power of mental vocabulary. From this viewpoint, the reason why we can accurately predict that Crimson and Ruby will advocate for the introduction of policies to facilitate the dismissal of workers in response to the coronavirus crisis is that (2) captures some material state of affairs (e.g., a brain state, some covert or overt vocalization, etc.) that is causally linked to such behavior.

These reductivist or ontologically conservative approaches are commonly divided into two groups, depending on whether they advocate for a *type identity theory* or a *token identity theory*[21] (e.g., see Kim, 2011, p. 122). According to the former, it's possible to establish an identity relation between types of mental events and types of natural events, regardless of their particular instantiation in different people or moments of time. In this sense, type identity theorists would then maintain that sentences such as (1) would be then straightforwardly translatable to sentences such as (3) (where "V" would reference a specific type of neural state and not just a particular state of the brain at some given moment, potentially different in Citric's and Emerald's cases). This mainstream kind of ontologically conservative approach, which we'll here call *straightforward reductivism*, is historically grounded in the work of Fiegl (1958), Place (1956, 1988), and Smart (1959), and it characterizes the standard view of mind and cognition implicit in many correlational approaches to the relation between cognitive processes and brain areas or patterns of activity (see also Kim, 2011; Smart, 2017; Stoljar, 2021).

Many authors have rejected the type identity theory, together with the straightforward reductivist program. One of the main criticisms is that it doesn't account for the possibility of *multiple realization*, i.e., that different particular instances of the same mental type

[21] The "type-token" distinction comes from the distinction between word types and word tokens; the sentence "run, Forrest, run" is composed of three words (three tokens), yet only two types of words (i.e., "Forrest", which appears once, and "run", which appears twice) (see Smart, 2017).

may be due to different natural states and processes (e.g., different patterns of brain activity, different material constitution, different learning histories, etc.). Many authors have instead adopted a *token* identity theory (Block, 1995; Davidson, 1970/2001; Lewis, 1966, 1980; Putnam, 1967/1975; see also Kim, 2011; Smart, 2017; Stoljar, 2021). Token identity theory can be seen as a form of *contextualist reductivism*, according to which the identity (and thus possibility of translation and reduction) between certain mental states and certain natural states can only be established for particular instances of such mental state. For example, in (1), Citric's belief and Emerald's belief could be due to different natural realizators in each case (e.g., the brain state V in Citric's case and the brain state P in Emerald's case), thus (1) wouldn't be straightforwardly reducible (or translatable in a context-independent manner) to (3). Furthermore, even in the case of just one individual, the natural realizators of the same mental type could vary over time; thus, Citric's belief could be identical to a certain brain state V in t_1, yet identical to the brain state P in t_2. In a radical sense, the multiple realizability argument implies that a given mental token might be realized by different physical states across time, species, individuals, or even material constitutions (e.g., nervous cells and tissues in the case of human beings vs. technological components in the case of artificial intelligences) (see Block, 1995; Lewis, 1980).

The argument from multiple realization has motivated many of the self-styled "non-reductivist" approaches to the mental (in the sense that they reject straightforward reductivism, although they could be committed to a contextualist variety of the reductivist stance[22]). Many of them make use, in one way or another, of the concept of *supervenience*, which imposes some restrictions on the possible relations between the mental and the non-mental. Particularly, mental properties are said to *supervene on* physical states; what this means is that while two agents (or two agent-environment systems, depending on the unit of analysis) might have identical mental properties despite having different material properties, the reverse relation doesn't hold: no materially identical systems can have different mental states. At this point, the different approaches vary depending on which kind of states (mental or non-mental) is prioritized in the causal explanation of behavior.

[22] Davidson's (1970/2001) anomalous monism constitutes an exception to this characterization. According to anomalous monism, mental events (i.e., singular instances of events described in mental terms) are identical to physical events (i.e., singular instances of events described in purely material terms) "for events are mental only as described" (p. 141); however, mental predicates cannot be reduced or translated to physical descriptions, even if they're coextensive, for "we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions" (p. 145). Hence no psychophysical laws can be established. In any case, since the kind of naturalist approaches that have had a greater impact in the field of mental health are those that assume the possibility of establishing such psychophysical laws, here we'll leave aside the discussion of anomalous monism.

Functionalist approaches, for example, highlight the explanatory role of mental states. In the functionalist conception of mind, mental states are not primarily individuated by their internal constitution, but by their *causal profile* (i.e., their causal relations to other mental states and behavior) (Block, 1995; Block & Fodor, 1972; Fodor, 1983, 1987, 2006; Lewis, 1966, 1980; Putnam, 1967/1975). In a sense, functionalists needn't be worried about the place of mind within a naturalist ontology. Assuming a token identity theory, they can assume that any particular instance of a given mental state is always identical to a specific convergence of natural states and processes; however, since these may vary across time, species, individuals or even material constitutions, a proper causal account of behavior must always involve a reference to mental states. Therefore, mentalistic explanations are not only irreducible, but also essential for a proper causal understanding of behavior.

Throughout the second half of the twentieth century, functionalist theories of mind established the theoretical foundations for the rise of the so-called "cognitive revolution" in the behavioral sciences. Against the behaviorist paradigm that had hitherto dominated the field, the cognitive uprise brought mental states and processes back at the center of psychological research. The new cognitivist paradigm reinstated the still echoing Cartesian parallel between the working of minds and that of machines; though this time, instead of clockworks, the metaphor established a relation between minds and the then newly-developed computers. On this new metaphor, minds are the software and whatever natural states that realize them (typically, brains) are the hardware: just as the same type of computer program can be realized by a manifold of different hardware realizators, mental activities can be the result of a manifold of different natural realizators (see Block, 1995). Thus, just like we don't think of computer programs as spooky, ontologically bizarre entities, we should not worry about the non-straightforwardly-physical character of mental states and processes. Whether we choose to describe and causally explain behavior in terms of material or mental processes would just be a matter of the exact level of analysis that we are prioritizing; cognitive science and functionalism just concede an explanatory primacy to the mental (or software) level.

A somewhat related approach to the mind-body problem is *emergentism*, historically grounded in the work of the British Emergentists, such as Samuel Alexander (1920/1966) or C. D. Broad (1925) (see McLaughlin, 1992/2008; see also Bedau, 1997; Bedau & Humphreys, 2008; Corradini & O'Connor, 2010; Maturana & Varela, 1980; O'Connor, 1994, 2020; O'Connor & Wong, 2005; Silberstein & McGeever, 1999; Varela et al., 1991; Varela & Thompson, 2003; Von Bertalanffy, 1950, 1968; Zhong, 2019). Two main emergentist approaches can be distinguished: weak or epistemological emergentism and strong or ontological emergentism (see Bedau, 1997; O'Connor, 2020; O'Connor & Wong, 2005; Silberstein & McGeever, 1999).

Roughly, epistemological emergentism can be defined as an ontologically non-committal proposal, which just proposes a distinction between higher-order and lower-order scales of analysis in scientific activity. In addition, epistemological emergentists hold the relatively weak claim that at least some higher-order predicates are indispensable for a full explanatory account of a given phenomenon, thus being non-reducible to lower-order predicates[23]. This idea is commonly expressed by the maxim that the properties of a whole cannot be justly accounted for by a summative description of the properties of the composing parts, since the complexities of both the interaction among parts and the interaction of the whole with other wholes must also be taken into account. For example, when talking about tap water, properties like "liquidity", "transparency", or "having an astonishing texture and unique mineral flavor when coming out from Madrilenian taps" can only be predicated from the whole, but not from its component parts (e.g., atoms of hydrogen and oxygen); thus, a full description of Madrid's tap water in terms of its lower-order components would leave its marvelous higher-order properties unexplained (see Bedau, 1997; Broad, 1925; O'Connor, 2020; O'Connor & Wong, 2005; Silberstein & McGeever, 1999).

On the other hand, ontological emergentists advocate for the stronger claim that *reality* is hierarchically structured in increasing levels of complexity (see Alexander, 1920/1966; Humphreys, 1997; Maturana & Varela, 1980; O'Connor, 1994, 2020; O'Connor & Wong, 2005; Silberstein & McGeever, 1999; Varela et al., 1991; Varela & Thompson, 2003; Von Bertalanffy, 1950, 1968; Zhong, 2019). Applied to the philosophy of mind, mental or cognitive states and processes would thus pertain to a different level of organization of matter. From this standpoint, higher-order phenomena and higher-order properties are the emergent result of complex interactions among lower-order phenomena, whose possibilities for interaction are consequently constrained by the higher-order phenomena. The first causal process has sometimes been called "upward causation", while the second one has been called "downward causation" (O'Connor & Wong, 2005; Varela & Thompson, 2003; for a thorough description and criticism of this view, see Kim, 1992, 1993). Thus, from an ontological emergentist approach, both the lower-order and the higher-order scales of analysis are explanatorily relevant: research on the laws and principles governing lower-order natural processes provides information on the upward causation dynamics that could explain the emergence of higher-order (e.g., mental) events or properties. These, in turn, would not only be explanatorily primary at the higher-order level of scientific explanation (and thus non-reducible), but also causally informative with regard to the lower-order processes from which

[23] In this sense, functionalism, as described above, can be understood as a form of epistemological emergentism.

they emerge (since these lower-order processes would also be constrained by the higher-order ones via downward causality). In what follows, we'll use the term "emergentism" to refer to this latter, stronger kind of emergentist approach.

If functionalism nurtured the 1970's cognitive revolution, emergentism provided the theoretical grounds for some of the *postcognitivist* approaches that emerged during the 1990's and that are now shaping the landscape of contemporary cognitive science (Chapter 1, section 1.5.3.). As we saw in Chapter 1, this relatively new post-cognitivist paradigm integrates various "non-Cartesian" and "non-reductivist" research approaches to the study of perception, cognition, and action. Instead, they propose alternative conceptions of mind, which emphasize the embodied, extended, enacted, or embedded character of cognition (hence their usual grouping under the name of "4E approaches" to cognitive and behavioral science) (see Chemero, 2009; Clark & Chalmers, 1998; Hutto & Myin, 2013; Maturana & Varela, 1980; Nöe, 2001, 2004; O'Regan & Nöe, 2001; Thompson, 2007; Varela et al., 1991; see also Heras-Escribano, 2019; Newen et al., 2018; Saphiro, 2014). Although some of these approaches still draw from functionalist assumptions (e.g., classical extended mind approaches; see Clark & Chalmers, 1998; see also Newen et al., 2018), more radical kinds of postcognitivism, which put a stronger emphasis on the embodied, embedded, and enactive character of cognition, explain cognitive abilities as emergent properties of the dynamic coupling between an organism and its environment (see Maturana & Varela, 1980; Nöe, 2001, 2004; O'Regan & Nöe, 2001; Thompson, 2007; Varela et al., 1991).

2.2.2.2. Ontologically revisionary and ontologically radical naturalisms

As we've seen, both functionalism and emergentism draw from the multiple realization argument to defend the explanatory value of our folk-psychological concepts at a higher-order scale of analysis. By contrast, other authors take the argument from multiple realization in the opposite direction: if mental types don't have a clear or unitary ontology (i.e., if it's not possible to establish a 1:1 relation between different types of mental events and different types of natural events), then any kind of mentalistic explanation will be, at best, a poor explanatory device; mentalistic explanations might have some heuristic value, but are essentially unable to pick up in a precise way the relevant causal processes involved in the production and maintenance of behavior. Therefore, behavioral and/or cognitive scientists should aim to *eliminate* some or all folk-psychological vocabulary from their models and theories (see P. M. Churchland, 1981; P. S. Churchland, 1986; Cornman, 1968; Epstein et al., 1980, 1981; Feyerabend, 1963a, 1963b; Lycan & Pappas, 1972; Ramsey, 2020; Ramsey et al., 1990; Rorty, 1965, 1970; Savitt, 1975; Skinner, 1945, 1953, 1974, 1981, 1990; Stich, 1983).

Although the term *eliminativism* (or *eliminative materialism*, as coined by Cornman, 1968) first appeared during the 1960's debates around Feyerabend's (1963a, 1963b) and Rorty's (1965, 1970) early radical proposals, this variety of naturalism can be traced back to at least Broad's (1925) discussion of a "pure" version of the materialist stance (see Ramsey, 2020), as well as to Skinner's (1945, 1953) radical behaviorist approach to psychology. Despite the differences among eliminativist proposals, the core idea of eliminativism is that folk psychology is or will prove to be a radically *false* or inherently implausible *theory*. An eliminativist approach can thus be understood as any stance that accounts for the mind-body problem through the elimination of our mental vocabulary (or at least some part of it) from scientific accounts of behavior. Instead, a proper science of behavior should aim at providing lawlike explanations in strictly naturalistic terms (i.e., terms that make reference to natural facts or sets of facts that are proven to be causally linked to a given behavioral manifestation).

This eliminativist stance can be implemented in at least two different ways (see Bickle, 1992; Irvine & Sprevak, 2020; Lycan & Pappas, 1972; Ramsey, 2020; Savitt, 1975). Drawing from the argument of multiple realization, some eliminativists admit that folk-psychological explanations might partially capture some of the relevant natural events that actually explain behavior, although in an imprecise or incomplete way. According to this approach, folk-psychological concepts would thus be analogous to old and already superseded scientific concepts such as phlogiston, employed until the 18th century to account for combustion (Bickle, 1992; Ramsey, 2020; Ramsey et al., 1990). In a sense, the explanatory use of phlogiston captured different chemical reactions between different chemical compounds that actually accounted for different instances of combustion. In the same vein, an eliminativist might still hold on to the idea that a given mentalistic explanation at time t_1 captures the natural processes that actually explain behavior, while the same mentalistic explanation given at time t_2 might indeed capture a different set of causally relevant natural processes.

This obviously affects the predictive capacity of our scientific explanations of behavior. Thus, mentalistic explanations should be progressively abandoned and replaced by others that just make reference to subpersonal processes (i.e., those actually involved in the causal production of a given behavioral outcome of interest). This, in turn, would foreseeably enhance the predictive power of our explanations (Bickle, 1992; Irvine & Sprevak, 2020; Lycan & Pappas, 1972; Ramsey, 2020; Rorty, 1965; Savitt, 1975). Irvine & Sprevak (2020) have recently referred to this kind of eliminativism as *discourse eliminativism*, since its eliminativist program aims at the continuous reshaping of our scientific discourse. Note that functionalism and discourse eliminativism can be regarded as two opposite sides of the same coin. Both may admit some "token" variety of the mind-body identity theory; however, while

functionalism gives explanatory primacy to mental types, discourse eliminativism gives explanatory primacy to physical types. Emergentism, on the other hand, could be seen as an intermediate account, which doesn't prioritize any given level of explanation.

In a nutshell, discourse eliminativism sees mentalistic explanations as poor explanatory devices, which might have some initial explanatory value but that should nonetheless be progressively removed from a proper scientific account of behavior. In this sense, this kind of eliminativism constitutes an *ontologically revisionary* naturalism: mental-state ascriptions capture some natural states and processes; however, these might be explanatorily irrelevant or inaccurate, and thus science should progressively fine-tune our ontological assumptions (Bickle, 1992). The functional contextualist strand within functional analytic approaches to behavior (Chapter 1, section 1.5.2.2.; see Hayes 2021) can be viewed as an example of this kind of ontologically revisionary naturalism (see also section 2.3.5.).

On the contrary, other eliminativists have pointed out that the mental is not amenable to reduction at any given moment nor for any given species or individual. These authors advocate for what we could call a *straightforward eliminativist* approach to the mental[24], since they draw from an outright rejection of any kind of identity theory. According to this more classical variant of eliminativism, our mental concepts are individuated through a number of properties (e.g., self-causation, intentionality, privacy, normativity, etc.) that no description of physical or material events can retain; therefore, and contrary to the aforementioned approaches, straightforward eliminativism is an *ontologically radical* naturalism: strictly speaking, mental states and processes *do not exist*, since our mental vocabulary does not describe any state of affairs in our natural world (P. M. Churchland, 1981; P. S. Churchland, 1986; Feyerabend, 1963a, 1963b; Ramsey et al., 1990; Rorty, 1970; Skinner, 1945, 1953, 1974, 1981, 1990; Stich, 1983; see also Irvine & Sprevak, 2020; Ramsey, 2020; Savitt, 1975). According to this approach, our mentalistic explanations have the same explanatory value as our explanation of Saint Theresa's ecstatic behavior in terms of a mystic union with the Holy Spirit; contrary to the case of phlogiston, the kind of alleged explanatory entity at stake here (a supernatural spirit) is so far from reconcilable with the most basic assumptions of naturalism that it makes no sense to even consider this kind of spiritualistic explanation as a kind of poor or primitive scientific explanatory tool.

Some of the most historically influential approaches to the behavioral and cognitive sciences have tried to implement different versions of the straightforward eliminativist stance. One of the first attempts to formulate an eliminativist approach to psychology was

[24] Irvine & Sprevak (2020) refer to this kind of eliminativism as *entity eliminativism*. However, we've preferred to use "straightforward eliminativism" in order to stress the contrast with straightforward reductivism.

Watson's (2013) behaviorist manifesto, which aimed at removing all trace of mentalistic vocabulary from psychology and replace it with precise descriptions of the subpersonal mechanisms involved in the production of behavior (see Chapter 1, section 1.3.1.). Later on, the main tenets of Watson's behaviorist approach where subsequently developed and specified in Skinner's radical behaviorism (see Chapter 1, section 1.3.1.; see also section 2.3.2.). From this perspective, the mental terms that appear in our folk-psychological explanations of behavior lack explanatory power and must be jettisoned as mere "explanatory fictions" that don't pick up any real fact (Baum, 2011; Chiesa, 1994; Moore, 2009; Schnaitter, 1984; Skinner, 1945, 1953, 1971, 1974, 1977, 1990). In a similar vein, other contemporary positions, more akin to the basic tenets of cognitive neuroscience, also draw from the idea that our folk psychology explanatory devices are readily dismissible. According to these positions, the real causes of behavior are to be properly established via an empirical inquiry into the lower-order subpersonal mechanisms (typically, brain states and processes) that are actually responsible for behavior (see P. M. Churchland, 1981; P. S. Churchland, 1986; Rorty, 1965, 1970; Ramsey et al., 1990; Stich, 1983).

To sum up, there are several ways in which ontological naturalism has been implemented. The main contemporary approaches to the philosophy of mind outlined above adopt either one of three possible views of the ontological status of mind: a) a form of ontologically conservative naturalism, which gives room for the mental within a naturalist ontology (e.g., straightforward reductivism, emergentism); b) ontologically revisionary naturalism, according to which science should progressively sharpen our ontological commitments (e.g., discourse eliminativism); or c) ontologically radical naturalism, which outrightly rejects the existence of mental properties (e.g., straightforward eliminativism). These different approaches also vary on a continuum with regard to the issue of the explanatory status of our mental states and processes, ranging from approaches that deny any possible explanatory role for mental concepts (e.g., straightforward eliminativism) to approaches that fully vindicate their explanatory role (e.g., straightforward reductivism, functionalism, and emergentism), as well as somewhat intermediate approaches which hold that mentalistic accounts of behavior are not much more than poor and primitive explanatory tools (e.g., discourse eliminativism). In the next section, we'll see how these different naturalist approaches have been implemented in the field of mental health.

2.3. The mental in mental health

In this section, we'll go back to the therapeutic models that we viewed in Chapter 1, pointing out their underlying understanding of the place of mind (and its tribulations) on nature. As

we'll see, although classical approaches typically leaned towards two antithetical positions (straightforward reductivism and straightforward eliminativism), the biopsychosocial model took inspiration from emergentist theories in its integrationist attempts. More contemporary approaches can be seen as advancing positions that contest or refine this emergentist framework. In any case, we'll try to show that none offers a sound response to the original worries of critical thinkers, due to their inability to yield a proper account of normativity in mental health contexts.

2.3.1. Straightforward reductivism in second-wave biological psychiatry

As we said at the beginning of section 2.2., many mental health researchers and practitioners don't necessarily adopt a particular stance with regard to the ontological and causal nature of the mental. In the case of psychiatry, the minimal interpretation of the medical model might be seen as such a position regarding the ontological status of mental disorders (see Chapter 1, section 1.1.). Conceptualized as mere "diagnostic kinds" (Tabb, 2017), mental disorders might just be defined as constellations of signs and symptoms with some predictive value, leaving more strong ontological commitments "about what is really going on with the patient" (Murphy, 2013, p. 967) aside. On this view, traditional diagnostic manuals, such as the DSM-5, are merely descriptive tools that aim at the establishment of statistically deviant patterns of behavior and their organization in clinically and statistically significant clusters. This nosological enterprise, in principle, grants no explanatory value to the different mental disorders that it gathers; it merely aims to describe them (Klerman, 1978; Kupfer et al., 2002; Murphy, 2009, 2013, 2020; Spitzer et al., 1978/2018¸ Tabb, 2015, 2017).

By contrast, others have favored a strong interpretation of the medical model, according to which mental disorders essentially are the result of specific neurobiological alterations (Kallmann, 1946; Karasu, 1982; Kety et al. 1968, 1971; Lehmaan & Hanrahan, 1954; van Praag, 1972; see also Kupfer et al., 2002; Murphy, 2009, 2013, 2020; Shorter, 1998; Tabb, 2015, 2017, 2020; Walter, 2013). This stronger ontological assumption characterizes the classical biomedical model, or the therapeutic model endorsed by second-wave biological psychiatry (see Walter, 2013; see Chapter 1, section 1.1.). The overall research project of second-wave biological psychiatry was to find the alleged neural basis of the mental disorders that are described in traditional nosological tools (i.e., the deviancies in typical neural functioning allegedly responsible for the clusters of psychopathological behaviors that constitute each mental disorder). On this approach, the proper assessment and intervention strategy is to: a) establish a correct diagnosis drawing from the person's symptoms; and b) consequently decide which treatment procedure should be applied. This draws from the assumption that diagnoses have explanatory and predictive power, i.e., that they inform us about the

biological causes of clinical phenomena and that they can thus inform us about the prognosis of the problem.

Therefore, the fundamental premise of second-wave biological psychiatry was that the different *types* of disorders gathered in mainstream diagnostic manuals should somehow correspond to different *types* of neurophysiological alterations; these would constitute the actual causes of the observed symptoms. This is the reason why biomedical research has traditionally focused on the search for *biomarkers* of the different mental illnesses (namely, alterations in the functional-anatomical structure of the brain) (see Bolton, 2013). In this regard, second-wave biological psychiatry constitutes an applied version of the type identity theory; in other words, it constitutes a straightforward reductivist approach to mental health.

As we saw in Chapter 1, the classical biomedical model behind second-wave biological psychiatry has been subject to widespread and relentless criticism for decades. The term "biomedical model" itself almost conveys a pejorative meaning, and not so many practitioners straightforwardly identify themselves as its supporters (Colombo et al., 2003; see also Fulford & van Staden, 2013, p. 393-394). First the critical approaches of the 1960's (see section 2.3.2.), then Engel's early formulation of the biopsychosocial model in the late 1970's and 1980's (section 2.3.4.) and finally third-wave biological psychiatry since the 1990's (section 2.3.5.) and post-cognitivist models more recently (section 2.3.6.), all these distinct and even antagonistic approaches to mental health practice have found some common ground in their analyses of the critical problems and main misconceptions of the biomedical model: its reductivist approach to mental health problems. However, it's far from clear what "non-reductivist" means in this context, since different approaches have pointed out to different "non-reducible" facets of psychological problems.

2.3.2. Straightforward eliminativism in Szasz's critical approach and early applied behavior analysis

Critical approaches to mental health emerged during the 1960's as a reaction against the medical model of mental health problems (e.g., Laing, 1960/2010; Szasz, 1960, 1961/1974) (see Chapter 1, section 1.2.). Szasz (1960, 1961/1974, 2011), one of the most important representatives of such critical approach, considered that the medical conception of mental illnesses was based on a *myth*. In this regard, Szasz's criticisms bear some resemblance to those raised by some foundational authors within first-wave behavior therapy (Dougher & Hayes, 2004;

Hayes, 2004), specifically those within early applied behavior analysis[25] (Ayllon & Haughton, 1964; Ayllon & Michael, 1959; Ferster & DeMyer, 1962; Lindsley, 1956, 1962, 1963, 1964; see also Skinner, 1953, 1977). Despite Szasz's well-known antipathy towards behaviorism (especially Skinner's radical behaviorism; see Szasz, 1991), the truth is that his critical approach shares at least two points in common with that of some early behavior analysts: a) that the conceptualization of diagnostic labels as picking out separate entities constitutes a case of fallacious reasoning; and b) that mental disorders, *qua* mental entities, don't exist. As Szasz (1960, p. 114) states:

> [...] the notion of mental illness is used to identify or describe some feature of an individual's so-called personality. Mental illness—as a deformity of the personality, so to speak—is then regarded as the *cause* of the human disharmony. [...] This is obviously fallacious reasoning, for it makes the abstraction "mental illness" into a *cause,* even though this abstraction was created in the first place to serve only as a shorthand expression for certain types of human behavior. (Szasz, 1960, p. 114)

In the same vein, early behavior analysts claimed that explanations and definitions of mental disorders in terms of "pathological", "biased", or simply "dysfunctional" mental states and processes weren't explanatory in any relevant sense; they were, at best, mere "explanatory fictions" (e.g., Ayllon et al., 1965; Lindsley, 1964; Rachlin, 1977a, 1977b; Skinner, 1953, 1971, 1977; see also Goddard, 2014). According to these authors, since the assumption of the existence of mental disorders as presumed etiological entities is not based upon the previous discovery of a distinct source of evidence other than the very same behavior they're purported to explained, talk of mental disorders constitutes an example of circular reasoning. Furthermore, these authors considered folk-psychological explanations of behavior -whether clinical or not- as embedded in mythical or "creationist" (Skinner, 1990, p. 1209; see also Vargas, 1991, p. 1) approaches to psychology, obviously inappropriate from a scientific point of view.

Thus, Szasz's critical approach and that of early behavior analysts shares some common ground. In particular, both can be understood as straightforward eliminativist approaches to the use of folk-psychological explanations in the natural sciences, according to which psychiatric talk of "mental disorders" wouldn't pick out any real entity. However, the

[25] Early advocates of behavior therapy (e.g., Eysenck, 1959) yielded similar arguments against the medical model. However, many of them seemed to be opposing psychoanalytic theories of psychopathology in particular; their criticisms weren't directed as such to the conflation of mentalistic and medical language in the definition and explanation of mental health problems, which is the main point raised by early behavior analysts and Szasz.

radical point of departure between Szasz's approach and early behavior analysis lies precisely in what each considers that must constitute the proper kind of scientific framework for clinical practice. In the same vein as biological psychiatry, early behavior analysts thought of mental health practice as just an applied scientific field whose object of study (i.e., problem behaviors for behavior analysts; neurophysiological states and processes for biological psychiatrists) is to be causally explained using the ontological and explanatory framework of the natural sciences. By contrast, as we'll see in more detail in section 2.4., Szasz explicitly rejected this naturalist framework for psychiatry and the social or human sciences in general, and assumed that human behavior should be explained in personal or normative terms (i.e., in terms of agency, free will and autonomy).

Both in the fields of psychiatry and clinical psychology, different therapeutic approaches emerged in response to the similar challenges posed by Szasz and behaviorist approaches to psychology. On the one hand, cognitive therapy (e.g., Beck, 1963, 1964; Ellis, 1958, 1962) and its subsequent development into cognitive behavioral therapy (e.g., Mahoney & Kazdin, 1979; Meichenbaum, 1977) aimed to reintroduce mental variables as an inescapable aspect of the subject matter of clinical psychology. On the other hand, Engel's (1977) biopsychosocial model emerged to challenge both straightforward reductivist and straightforward eliminativist arguments by proposing an integrative paradigm for psychiatry and medical healthcare in general. In the next sections, we'll delve into the philosophies of mind behind these two approaches.

2.3.3. Functionalism in second-wave behavior therapy (CBT)

Partly in reaction to early behavior analysts' dismissal of mentalistic explanations as creationist-like myths and the eliminativist assumptions behind it, during the 1960's and 1970's there was a progressive revival of interest in mental states and processes as potential mediators of clinical change. In the field of clinical psychology, this eventually led to the development of the now prevailing psychological model of mental health problems: cognitive behavioral therapy (hence CBT) (see Chapter 1, section 1.3.2.). Classical CBT approaches -Hayes's (2004) "second-wave behavior therapy") provide more or less detailed descriptions of the mental structures and processes that allegedly explain psychopathological behavior (e.g., Bandura, 1969; Beck, 1979; Mahoney & Kazdin, 1979; Meichenbaum, 1977). Despite the many conceptual differences among cognitive approaches to clinical practice, the central tenet of classical CBT is that many psychological problems are the result of inner maladaptive cognitive structures that affect the way we perceive and appraise life events. The content of these maladaptive cognitive structures consists of a series of irrational beliefs that systematically produce automatic negative thoughts and utterances, which in turn cause the

emotional and behavioral disturbances that characterize different mental health problems. The alleged mechanism by which these irrational beliefs systematically produce such problematic thoughts, emotions, and behaviors is explained in terms of a series of cognitive biases or distortions that affect normal information processing dynamics, yielding negative appraisals that, in turn, come to reinforce our previous irrational beliefs (see Hyland & Boduszek, 2012).

These approaches thus employ folk-psychological concepts to explain clinical phenomena. The way they were originally formulated can easily lead us to think of them as contemporary exemplars of the eerie specter of Cartesian substance dualism. However, we must remember that CBT was the result of merging behavior therapy and cognitive therapy (see Guinther & Dougher, 2013). On the one hand, the former encompasses the different behavioral approaches to clinical practice whose underlying philosophy of psychology was methodological behaviorism (Guinther & Dougher, 2013; Madden et al., 2016; see Chiesa, 1994; Moore, 2009), which takes the observation of behavior as the only properly scientific method for psychology (see Chapter 1, section 1.3.1.); on the other hand, the latter was at least indirectly influenced by the cognitive revolution in basic psychological research, which, as we've seen, was significantly driven by functionalist conceptions of the mind (see Dobson & Dozois, 2010). In this sense, the thorough descriptions of the inner mechanistic workings of our minds characteristic of CBT could be understood as an ontologically non-committal way of pointing out predictively relevant functional states mediating perception and action. Thus, we may think of cognitive therapy and its subsequent development into CBT as an implementation of functionalist theories of mind in clinical practice. According to such position, our talk of cognitive schemata, irrational beliefs, cognitive distortions, and automatic negative thoughts might be identical to their natural realizers in particular instances, but it's the kind of fine-grained mentalistic descriptions that cognitive theory provides what must constitute the primary explanatory tools of a proper science of clinical psychology.

2.3.4. Emergentism and the biopsychosocial model

On the other hand, in the field of psychiatry, Engel's (1960, 1977, 1978, 1980, 1997) classical biopsychosocial model attempted to find a middle ground between the straightforward reductivist program of the classical biomedical model and the straightforward eliminativism entailed by critical approaches like Szasz's. According to the biopsychosocial model, both the classical biomedical model and Szasz's critical approach shared a narrow definition of illness, according to which medical illnesses would be strictly identified with bare deviations from anatomical or functional bodily patterns (Chapter 1, section 1.2.). Engel proposed instead the adoption of a broader and holistic conceptualization of illness in general, not to be

exclusively identified with some neuropathological process, but to also integrate the psychosocial dimensions of *any* kind of illness (whether mental or somatic). Critically, Engel held that psychosocial factors should be taken into account not only because of moral or humanitarian reasons, but scientific ones as well; the psychological and social dimensions of illness were conceived of as relevant *causal* factors, which a properly scientific medicine should take into account in order to provide a full explanation of any illness and thus adequately inform their intervention designs and targets (although see Awais & Nielsen, 2021). In this sense, the distinctive mark of the biopsychosocial model has been its plea for theoretical and practical *eclecticism* and its insistence on the importance of not committing to any kind of reductivist scope to address health-related issues (Engel, 1960, 1977, 1980; see also Awais & Nielsen, 2021; Bolton & Gillett, 2019; de Haan, 2020b; Ghaemi, 2009, 2010; Pilgrim, 2015; Van Oudenhove & Cuypers, 2014).

Contrary to the ontologically non-committal attitude of many functionalists, Engel's non-reductivist and integrative approach was at least originally based on relatively substantial ontological commitments. In particular, as we saw in Chapter 1, his holistic conceptualization of mental health was grounded on von Bertalanffy's (1950, 1968) General System Theory (see Engel, 1977, 1978, 1980), which draws from the idea that living creatures are *open systems* (as opposed to the closed systems studied by physics) and that "reality [is] a tremendous hierarchical order of organised entities, leading, in a superposition of many levels, from physical and chemical to biological and sociological systems" (von Bertalanffy, 1950, p. 164). Contrary to reductivists, von Bertalanffy assumed that the study of the behavior of living beings as open systems required a broader scope and an intertheoretical set of explanatory tools. Inspired by this approach, Engel viewed mental health problems as complex entities, as multifaceted products of the interplay of different causal factors that should thus be addressed from different scales of analysis: the biological, the psychological, and the social. As such, his biopsychosocial model can be understood as an emergentist approach to mind in both mental and somatic healthcare.

Despite its widespread implementation, however, the biopsychosocial model has been thoroughly criticized (Chapter 1, section 1.4.). Criticisms revolve around three interrelated issues: a) its lack of clarity regarding the conceptualization of psychological or mental phenomena and its relation to biological states and processes (which is directly related to the mind-body problem); b) its inability to yield a consistent overarching conceptual framework (i.e., the integration problem); and c) the unclear status of normativity within a naturalist -even if multi-layered- perspective (i.e., the problem of normativity) (Craddock et al., 2008; de Haan, 2020a, 2020b, 2020c; Ghaemi, 2009, 2010; Murphy, 2013; Matthews, 2013;

Pilgrim, 2015; Van Oudenhove & Cuypers, 2014; see also Bolton & Gillett, 2019). Here we'll focus on the first two concerns, and in section 2.4. we'll come back to the third one.

A primary line of concern with the biopsychosocial model has to do with its account of the mental and of the exact way in which the mental is related to other scales of analysis in Engel's holistic framework (de Haan, 2020a, 2020b; Ghaemi, 2009, 2010; Van Oudenhove & Cuypers, 2014). Emergentists and like-minded "non-reductivists" might be right in pointing out that a correct understanding of the different processes and phenomena studied by different scientific disciplines require their study in the appropriate scale of analysis. However, straightforwardly applying this line of thought to the mental and claiming that mental states and processes are just states of affairs at an unspecific higher-order scale of analysis, distinct from the biological level, doesn't solve the ontological puzzles of the mind-body problem; if anything, it reinstates them (see Kim, 1992, 1993). As we've seen, the biopsychosocial model takes it that our mental life can be considered as the emergent result of complex interactions among brain or bodily processes (i.e., upward causation), whose possibilities for interaction are in turn constrained by our mental life (i.e., downward causation). But what then is this new and higher-order mental kind of stuff? What exact properties define it? Are they of an immaterial nature? If so, how exactly does an immaterial kind of entity (no matter how higher-order) causally relate to matter? Princess Elisabeth's worries (section 2.1.2.) can now be rephrased in the emergentist's jargon: how exactly do higher-order mental processes (downwardly) "constrain the possibilities of interaction" among lower-order natural ones? And how are the former (upwardly) caused by the latter?

This leads to the second major line of criticism against the biopsychosocial model, related to the integration problem. In this line, several authors have complained that the biopsychosocial model, as proposed by Engel, is just too loose or vague to constitute a properly integrative model for scientific research and clinical practice (de Haan, 2020a, 2020b; Ghaemi, 2009, 2010; Murphy, 2013; Matthews, 2013; Van Oudenhove & Cuypers, 2014). Nowadays, almost everyone agrees with its basic tenets: that not only neurobiological, but also psychological and social factors interact in the causal production of mental health problems and that optimally effective interventions should tackle all the relevant factors at play in each scale of analysis. However, the inherent theoretical and practical eclecticism of the biopsychosocial model yields no systematic way of analyzing how exactly these different factors come to interact in the production of mental health problems (de Haan, 2020a, 2020b; Van Oudenhove & Cuypers, 2014).

Two contemporary approaches to mental health practice have arised partially in response to these problems: third-wave biological psychiatry, which reestablishes the brain as

the locus of integration among different scales of analysis (section 2.3.5.), and the more recent postcognitivist approach to mental health (section 2.3.6.), which instead advocates for a refinement of Engel's holistic framework.

2.3.5. Discourse eliminativism: third-wave biological psychiatry vs. post-Skinnerian third-wave behavior therapy

As we saw in Chapter 1 (section 1.5.1.), third-wave biological psychiatry arose during the 1990's in the so-called "decade of the brain", and it was revindicated at the beginning of the last decade as a "new paradigm" for mental health care; specifically, one which should replace and overcome the problems of the previous generation of biomedical research (Walter, 2013; see also Andreasen, 1997; 2001; Cuthbert & Insel, 2013; David & Halligan, 1996, 2000; Insel et al., 2010; Insel & Cuthbert, 2015; Murphy, 2013, 2020; Kotov et al., 2017, 2018, 2020; Tabb, 2020). The validity and reliability issues of the traditional nosological tools (e.g., DSM), together with the hitherto fruitless research on the biomarkers of mental disorders, fostered the rise of critical voices within and without the field of mental health, with some of them coming from the highest mental health institutions in the Western world, e.g., the National Institute of Mental Health (hence NIHM) (see Cuthbert & Insel, 2013; Insel et al., 2010; Insel & Cuthbert, 2015; see also Kotov et al., 2017, 2018; Tabb, 2020; Walter, 2013).

The Research Domain Criteria (hence RDoC) initiative (e.g., Insel et al., 2010) emerged amidst the crisis of traditional nosologies to provide a new framework for research in the field of mental health (see Cuthbert & Insel, 2013; Insel & Cuthbert, 2015). This initiative constitutes the institutional endorsement of the main commitments behind third-wave biological psychiatry. We already saw in Chapter 1 (section 1.5.1.) that a core characteristic of the RDoC initiative is its emphasis on the multi-level and dimensional character of psychopathology. The central assumption here is that the validity and reliability problems of traditional nosologies are fundamentally due to their categorical character. Traditional classification systems conceive mental health problems as an on/off phenomenon, and include relatively arbitrary cut-off criteria to distinguish between those who get a specific diagnosis and those who don't (or those who get another one). Presumably, this fosters the apparition of common comorbidity, diagnosis instability, and other related problems. Instead, third-wave biological psychiatry advocates for the open investigation on the actual structure of mental health phenomena, which are assumed to be widely distributed among the population, cutting across traditional clinical/non-clinical distinctions. This, in turn, should allow for a more precise investigation of the multi-level etiology of mental health problems (Cuthbert & Insel, 2013; Insel et al., 2010; Insel & Cuthbert, 2015; see also Kotov et al., 2017, 2018).

In this sense, third-wave biological psychiatry inherits the multi-level perspective of the biopsychosocial model; as we saw, the RDoC initiative promotes research at various scales of analysis, in order to understand the multi-level "genotype" or causal structure of mental health problems. But how are all these scales of analysis integrated? And how does the mental fit within this multi-level approach to the etiological "genotype" of mental disorders? Contrary to the eclectic premises of the biopsychosocial model, supporters of third-wave biological psychiatry give a straightforward answer: mental disorders essentially are *brain disorders* (Insel et al., 2010, p. 749; Insel & Cuthbert, 2015), and the (brain) "circuit-level (…) is the focal element of the RDoC organization" (Insel et al., 2010, p. 749). Furthermore, even if they include cognitivist constructs within their research domains, their "units of analysis" make strict reference to biological and behavioral dimensions; hence the definition of RDoC in the NIHM's website as "an ongoing initiative to explore psychopathology based on dimensions of observable behavior and neurobiological measures" as well as the requirement "that [RDOC's] concepts meet a set of equally weighted twin criteria: evidence for a functional dimension of behavior or cognition, and evidence for a specific neural system involved in this functional dimension" (NIHM, 2019). Eventually, it's assumed, the RDoC initiative will provide mental health research and practice with sufficient evidential basis to clarify how the different relevant dimensions relate to brain structure and functioning. In this sense, third-wave biological psychiatry can be seen as some kind of discourse eliminativist project; one where research on the cognitive processes involved in mental health will progressively depart from folk-psychological assumptions and adopt instead the language and explanatory tools of the more mature neurocognitive and behavioral sciences.

Contrary to this overemphasis on the brain circuitry level, some contemporary psychological approaches to mental health have highlighted the central role that the individual's context play in the maintenance of psychological distress. Drawing from behavior analytic origins, some approaches within third-wave behavior therapy (Hayes, 2004) reinstate the psychological scale of analysis (i.e., the scale of the individual's interaction with their social and natural environment) as the primary locus of psychopathology (see Chapter 1, section 1.5.2.). In particular, we'll focus here on Acceptance and Commitment Therapy (hence ACT; Hayes et al., 1999), and its underlying post-Skinnerian approach to human cognition and language, Relational Frame Theory (hence RFT) (Barnes-Holmes et al., 2001; Hayes et al., 2001).

As we saw in Chapter 1 (section 1.5.2.), ACT departs from more "traditional" versions of behavior analysis in several important respects. One important line of divergence resides in its approach to the mediational processes that cognitive approaches posit as the proximal

causes of both clinical and non-clinical behaviors and experiences. Early behavior analysts dismissed cognitivist attempts to explain behavior via intermediate "mental" or "cognitive" processes as remnants of a mythical or creationist-like approach to psychology; by contrast, some of the most prominent authors within post-Skinnerian third-wave behavior therapy encourage a different approach to cognition and cognitive mediation (see Chapters 1 and 8, sections 1.5.2.2. and 8.3.). The hallmark of these post-Skinnerian third-wave approaches is their attempt to reinterpret core cognitive concepts like that of "belief" or "mental representation" in terms of arbitrarily applicable relational responding. Many instances of psychological suffering are thus construed as the result of the individual's inflexible behavioral rules, which embed the arbitrary relations that the person establishes between events in their environment (Hayes et al., 1999).

However, given the highly technical nature of RFT terms and concepts, ACT proponents are comfortable with taking "topographically mentalistic terms seriously if they turn out through functional contextual analysis to orient applied and basic work that is behaviorally sensible" (Hayes, 2021, p. 239). In this sense, these authors adopt a pragmatic attitude towards what they call "middle-level terms" (i.e., non-technical, yet theoretically-specific terms that lie between technical and folk-psychological terms; Barnes-Holmes et al., 2016, p. 367). Rather than straightforwardly dismissing them as "explanatory fictions", let's put them to test: insofar as hypothesized middle-level terms work for predicting and influencing an individual's behavior, let's use them; as soon as they hinder progress in these areas, let's dismiss them (Barnes-Holmes et al., 2016; Hayes, 2021; Hayes et al., 2012). In this sense, post-Skinnerian strands within third-wave behavior therapy share with third-wave biological psychiatry their discourse eliminativist assumptions; however, instead of the "brain circuitry level", these approaches reinstate the relation between an individual and their environment as the proper scale of analysis from which explanatory progress should be assessed.

This conceptual tension between third-wave biological psychiatry and post-Skinnerian third-wave behavior therapy rekindles the integrationist worry: even if we take the explanatory complexity of psychopathology at face value and adopt a multi-level framework, how should we spell out the details of the links among different scales of analysis? Moreover, which should we adopt as the one against to which assess explanatory progress? These questions ultimately come down to our concern about the relation between the mental and the non-mental: which science is to provide us with the most fundamental language to explain (or explain away) mental states and processes; the science of brain circuitry, or the science of behavior?

2.3.6. Emergentism revisited: the enactive approach to psychiatry

Some may take the recurrence of these questions as indicative of the decay of the biopsychosocial model and the futility of its integrationist aspirations. Much to the contrary, as we saw in Chapter 1 (section 1.5.3.), the spirit of the biopsychosocial model is alive and well in the field of mental health. In this regard, post-cognitivism (see section 2.2.2.1.) has inspired a number of contemporary approaches to mental health that aim to provide a more appropriate framework for the conceptualization of psychological problems (e.g., de Haan, 2020a, 2020b, 2020c, 2021; de Jaegher, 2013; Glackin et al., 2021; Drayson, 2009; Fuchs, 2007, 2009; Nielsen, 2021; Nielsen & Ward, 2018, 2020; Roberts et al., 2019; Sneddon, 2002). Among them, the recently developed enactive approaches to mental health (see de Haan, 2020a, 2020b, 2020c, 2021; Nielsen, 2021; Nielsen & Ward, 2018, 2020) have explicitly undertaken the task of providing a sounder and properly integrative version of the biopsychosocial model.

In Chapter 1, we illustrated the enactive approach to mental health with de Haan's version of it (2020a, 2020b, 2021; for a related, yet different approach, see Nielsen, 2021; Nielsen & Ward, 2018, 2020). As we saw, de Haan's goal is to overcome the pitfalls of dualist approaches, the classical version of the biopsychosocial model and, in her terms, "neuro-reductionist" accounts of mental health. As she views it, dualism is an obviously inadequate ontological stance; however, (neuro-)reductionist (or eliminativist) approaches like the one adopted by third-wave biological psychiatry aren't satisfactory accounts of the mental either. In her words, "what is problematic about neuro-reductionism is [...] that it *a priori* assumes the brain's causal primacy. When it comes to complex processes, however, reductionist strategies are unlikely to be adequate" (de Haan, 2020a, p. 5). In addition, its primary focus on brain states and processes as the node of integration among the different levels of scientific description and explanation yields a purely passive view of people with mental health problems, where their agential status is diminished and endangered. Finally, de Haan (2020a) praises Engel's biopsychosocial model for its attempt to provide an integrative approach to mental health, but criticizes its incapacity to provide a unitary and coherent conceptual framework to such multi-level account. Thus, she concludes that "both for reasons of adequacy as well as for ethical reasons it is worthwhile aiming for a model that is integrative and coherent without being reductionist" (de Haan, 2020a, p. 5; see also Nielsen, 2021; Nielsen & Aftab, 2021; Nielsen & Ward, 2018, 2020).

The alternative integrative and multi-level model that she proposes draws from enactivism (see section 2.2.2.1.); in particular, from autopoietic enactivism (e.g., Varela et al., 1991), which emphasizes the continuity between life and cognition. The idea behind the "life-mind continuity thesis" is that the "mental" and the "living" character of certain biological

systems essentially are one and the same thing (de Haan, 2020a). Enactivists of this sort reject the functionalist and cognitivist construal of mind as a separate, mechanical or computational entity; instead, their core idea is that what characterizes living beings is that they display a particular organizational structure, namely, a self-organizing one; living organisms are autonomous, self-organizing biological entities, who radically depend on a continuous exchange of matter and energy with the environment to subsist and maintain their internal organization. Thus, in order to survive, living organisms need to continuously engage in at least a minimal kind of "sense-making"; they need to be able to discriminate between potentially beneficial and potentially harmful resources in their environment, and thus between "correct" and "incorrect" courses of action. This, in turn, is the distinctive mark of the mental: the value-laden or *normative* character of such interactions. Therefore, living beings essentially are mental beings (de Haan, 2020a).

This characterization of living (i.e., mental) beings draws from a multi-level approach to material reality; reality is hierarchically structured in increasing levels of complexity, and relations of emergence mediate between the different levels. In this sense, the therapeutic model proposed by enactivist psychiatry also is an example of an emergentist approach to mental health practice. However, in contrast with the biopsychosocial model, the enactivist approach to psychiatry provides a specific account of how the biological and the mental relate to each other. In the enactivist view, our mental vocabulary does not point to some ontologically extraneous and spooky entities; on the contrary, it ultimately refers to the complex, dynamic, non-mechanical self-organizing processes that characterize living beings and the meaningful way in which they relate to their environment to preserve such self-organization. In de Haan's words:

> The life-mind continuity thesis thus adopts emergence in that the properties of matter also depend on their organizational structure. Once matter is organized in such a way as to be living matter, it will engage in sense-making. There is no need then to assume that matter and cognition refer to two wildly separate, incomprehensibly connected, realms: matter in specific (i.e., self-organizing) patterns is minded (de Haan, 2020a, p. 7).

As we already advanced in Chapter 1 (section 1.5.3.), we think that this approach has a number of virtues. Firstly, it provides a more nuanced view of how the different realms of analysis integrate, thus promoting a richer and more comprehensive understanding of mental health problems. In addition, it avoids the commitment to many of the Cartesian ontological and epistemological tenets that we saw at the end of section 2.1.: precisely, the reason

why enactivist approaches like de Haan's reject functionalism about the mind is that these typically retain a commitment to the Cartesian framework -at least, to intellectualism, mental causalism, intellectualism, and representationalism. Instead, it endorses a *non-cognitivist* approach to mental health that has some interesting similarities with both new and old behavior analytic approaches, such as the establishment of the organism-environment system as the proper unit of analysis. As we'll discuss in Chapters 7 and 8, this kind of approach might have some important benefits for the intervention with people with mental health problems.

Finally, and perhaps more importantly, de Haan's approach explicitly addresses the problem of normativity, which has been often neglected, and explicitly establishes a close connection between it and the problem of mind. Here, however, is where we think that problems start for de Haan's and other post-cognitivist approaches to the mind: in the way they conceptualize normativity and the tight link between the mental and the normative (see Heras-Escribano et al., 2015; Heras-Escribano & Pinedo-García, 2018). In the next and final section, we'll revisit this problem in the field of mental health.

2.4. Mind and normativity in mental health care

As we've seen, most of the above-mentioned approaches to mental health revolve around the problem of the "mental" aspect of mental disorders, placing a special emphasis on the mind-body problem. However, the problem of normativity, which in the field of mental health amounts to the problem of the "disorder" aspect of mental disorders, has received far less attention -at least in mental health research and practice.

Recall that Descartes's whole theory of mind revolved around this problem. His substance dualism was designed to keep the mental away from the jaws of modern science. Why? Because Descartes already observed the intimate conceptual connection between our capacity to describe ourselves and other beings in normative or personal terms, (i.e., in terms of agency, freedom, responsibility, epistemic and moral merit or demerit, etc.) and our capacity to ascribe mental states to one another. In mental health practice, this tight connection between mind and normativity is especially visible. Mental health problems are often defined in terms of deviations from certain, often interrelated norms: the norm that one should act in accordance with one's own and other's well-being; the norm that one should find meaning in one's relationships and doings; the norm that one should feel integrated in one's environment; the norm that one should be able to make rational decisions about one's life; the norm that one's behavior and cognition should be intelligible to oneself and to others; and myriad other norms that we sometimes find hard to make explicit. And mental health researchers,

practitioners, and users typically spell out the deviations from these norms in terms of mental states (i.e., irrational beliefs and desires that make one or others suffer; unusual perceptual experiences that set us "apart from reality"; sustained failures to act on one's intentions; contradicting values that immobilize us; etc.).

Szasz (1960, 1961/1974, 2011), Laing (1960), and other critical thinkers already pointed this out (see Chapter 1, section 1.2.). In particular, on Szasz's view, the subject matter of mental health practice is not (nor should be) a human primate's statistically deviant behaviors, but a human being's decisions and actions; hence his *theory of personal conduct* (Szasz, 1961/1974). As we saw, Szasz's theory makes use of normative notions such as "free will" and "responsibility" to map the *meaning* of the patients' courses of action, both for themselves and for their social context; the goal is, ultimately, to spell out the *norms* that patients follow when they act, think, and feel the way they do. Mental health problems thus are *problems in living*, i.e., contradictions between what a person does and what they should be doing according to *their own* norms and values; likewise, the therapist's role -the only legitimate one according to Szasz- is to help the person explore, recognize, and solve these potential contradictions.

That's why, apart from the circular reasoning argument that we saw in section 2.3.2., Szasz rejected the medical understanding of mental health problems; for him, the term "medical" necessarily conveyed the commitment to the describability of the subject matter of medical inquiry in purely materialist terms. However, this cannot be done in mental health contexts. As we'll further develop in Chapters 3 and 4, the core idea here is that mental vocabulary has an irreducible and ineliminable *normative or prescriptive force*; it allows us to rationalize or make each other's behavior intelligible in a manner that no purely descriptive statement can do (see Heras-Escribano et al., 2016; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014, 2020). If psychiatry were to be thought of as an applied field of the natural sciences, the mental should thus be eliminated from psychiatric speech. However, Szasz stated that it's impossible to make sense of mental health practice without such mentalistic notions; it is only through our talk of mental states and processes that we're able to spell out the norms and values at play in clinical practice. For him, a truly scientific psychiatry could not dispense with the use of mental, personal vocabulary; on the contrary, it ought to reject any naturalist (e.g., medical) framework and begin to understand itself as a social or human science rather than a natural one.

Engel's attempt to bridge this gap between the realm of biology (or, more broadly, the realm of natural laws) and the realm of mentalistic, normative explanations (i.e., the realm of reason and normativity) thus entirely mistook Szasz's point; while the latter was

revendicating a *personal* approach, i.e., one based on normative concepts such as free will, meaning, autonomy, etc., to mental health research and practice, the former's attempted solution was to fit the normative (i.e., the mental) back into a multi-level, yet *subpersonal* nomological framework. Enactivist approaches fall prey of the same problem (see Heras-Escribano et al., 2016; Pinedo-García, 2020). Normativity is here viewed as a matter of causal complexity: being a free, autonomous, and accountable agent is conceived of as a matter of merely having a certain kind of biological system; one whose interactions with the environment could not be accounted for in terms of simple, mechanical, or linear causation, but in terms of non-linear or complex causation. (Paradoxically, this account seems to assume that the more causal factors influencing one's behavior, the more "autonomous" or agential it is).

Szasz's theory of personal conduct poses some problems of its own though. Despite the merit of highlighting the irreducibility and ineliminability of personal vocabulary in mental health contexts -often forgotten in biomedical thinkers' most ardent dreams-, his steadfast libertarian approach to mental health leaves much to be desired. Other than its political fancifulness and its ethical questionability, it poses some conceptual puzzles when it comes to answering questions about why people behave the way they do and why they change (for example, through therapy). When a person's behavior changes in a certain way, is it due to their deciding that they're going to behave that way, due to a number of historically antecedent events (e.g., neurophysiological processes, past experiences, etc.) or due to a combination of both? According to this last option, the mental states and processes involved in decision-making would be part of the causal events that determine or "influence" what a person does. Szasz would surely reject this, since such idea would imply that the concepts of "free will" and "choice" would be again just a part of a larger causal-deterministic chain, thus no longer being apt to qualify someone's behavior in personal (i.e., voluntaristic, normative) terms. However, this posits another problem: either personal explanations appeal to laws of a different nature than natural laws (thus reinstating the Cartesian para-mechanical hypothesis), or voluntaristic explanations are not causal in a relevant sense. The latter seems to be what Szasz is claiming when he states the following:

> What, then, can *we* say about the relationship between psychosocial laws and physical laws? We can assert that the two are dissimilar. Psychosocial antecedents do not cause human sign-using behavior in the same way as physical antecedents cause their effects. Indeed, the use of terms such as "cause" and "law" in connection with human affairs ought to be recognized as metaphorical rather than literal. (Szasz, 1961/1974, p. 8)

That's why Szasz overtly rejected all causally deterministic accounts of human behavior, regardless of their psychoanalytical, behavioristic, or neurobiological flavor. Though recognizing the "effects, which are indeed significant, of past personal experiences" on behavior, he aimed "to maximize the scope of voluntaristic explanations—in other words, to reintroduce freedom, choice, and responsibility into the conceptual framework and vocabulary of psychiatry" (Szasz, 1961/1974, p. 6).

Szasz's outright rejection of every attempt to address human behavior as the product of natural causes is clearly problematic. There are two different ways to flesh out this rejection: we can either understand Szasz's attack on subpersonal accounts of behavior as implying that a) we *cannot* actually account for behavior in deterministic terms (i.e., that subpersonal, deterministic explanations are mere self-inflating metaphors or fictions employed by members of different mental health institutions to secure their professional status); or b) that we *should not* employ a subpersonal framework to account for problems in living of human beings *qua* persons. The former interpretation is just false, at least for anyone familiarized with the explanatory, predictive, and intervention power of certain nomological frameworks and their related clinical procedures; the second, from our point of view, raises some serious ethical problems. Firstly, Szasz's unrealistic view of the relation between the individual and their environment yields an overwhelming blame culture where the individual is primarily responsible for all and every aspect of their "problems in living". Secondly, it gives us no clue as to what should inform our intervention designs, nor how should we evaluate them. Ideally, intervention designs should draw from a given conception of the etiological factors that at least have some significant causal influence on people's mental health issues. Why wouldn't mental health practice be also concerned with the discovery of such subpersonal etiological factors? After all, if therapy must be limited to accompanying people in the discovery and resolution of their problems in living, why shouldn't mental health aim to offer the most well-supported and evidence-based available methods to introduce such desired changes in people lives?

Now we seem to stand at a crossroads. Should we follow Szasz and reject the possibility of analyzing the natural causes of mental health problems? Or should we instead pursue the reduction (or elimination) of personal and normative vocabulary in our explanations? As we'll see in Chapter 3, these two "exclusivist" positions are the particular expression in the field of mental health of a commonplace, yet disputable general assumption about the relation between mental vocabulary and the vocabulary of the natural sciences. Specifically, what Szasz and his opponents seem to have in common is the idea that either mental vocabulary is *reducible* to mere descriptions of causally relevant states of affairs or our mentalistic,

folk-psychological interpretative practices are incompatible with a naturalist worldview. In the former case, we seem to lose track of what's specifically "mental" about mental health problems, as Szasz pointed out; in the latter case, we're left with two options: to expurgate the mental from our ontology, or to reject ontological naturalism.

In the following chapters, we'll see in more detail why neither of these options provides a satisfactory account of the problems of mind and normativity. We'll also see that the key to escape this dilemma lies in rejecting the implicit semantic commitment of Cartesianism; that is, descriptivism about mental-state ascriptions.

2.5. Conclusion

In this chapter, we've seen how the problems of mind and normativity traverse general contemporary debates in the philosophy of mind, the behavioral and cognitive sciences, and, finally, the applied field of mental health research and practice. We've traced these problems back to Descartes's para-mechanical theory of mind, lying out its core theoretical commitments. As we've seen, the most often discussed Cartesian commitment has been substance dualism, or the idea that mind and matter are two different kinds of substances; in this sense, competing approaches to the mental in both theoretical and applied research fields tend to play what Pinedo-García (2020) has called the "you are more dualist than I am" game, which he takes to be a sign of "an unavoidable, though unfortunate, consequence of the still felt Cartesian influence" (p. 7).

Following Pinedo-García (2020), we've argued that this influence goes well beyond the common use of "Cartesian" or "dualist" as shameful epithets to throw against opposing theories. Cartesianism doesn't only comprehend substance dualism, but a wide array of other ontological and epistemological theses (see section 2.1.4.). Apart from substance dualism, other important ontological commitments comprise mental causalism (i.e., the idea that mental states stand in causal relations with the individual's body, behavior, or other mental states), intellectualism (i.e., the idea that being in a mental state or acting upon it is a matter of entertaining certain regulative propositions and acting accordingly) and factualism (i.e., the idea that minds are some kind of *res*, whether material or not). On the other hand, Cartesian epistemology is characterized by representationalism (i.e., the idea that minds are representational devices and that we don't have a direct epistemic access to the world) and two closely related ideas: the "privileged access" conception of self-knowledge, linked to the idea that, from a first-person perspective, one can never be wrong about one's own mental states, and the "analogical" conception of knowledge of other minds, i.e., the idea that, from

a third-person perspective, we can only "read" others' minds by drawing an analogy with ours.

These other ontological and epistemological theses are differentially shared by many of the naturalist approaches to the mental that we've seen in section 2.2. Most of them draw from a standard image of folk psychology; specifically, one according to which our mentalistic interpretative practices respond to a theoretical, proto-scientific effort to causally explain each other's behavior. Drawing from this mindreading conception of our folk-psychological interpretative practices, the different naturalist approaches to the mental either try to reconcile this capacity with the basic tenets of naturalism or reject the very possibility of doing so.

These approaches can be divided into three different kinds. Ontologically conservative naturalisms typically implement some variety of the mind-body identity theory. Straightforward reductivist approaches advocate for a type identity theory, whereby types of mental events are equated to types of material events; on the contrary, functionalists and emergentists can be understood as implementing some kind of contextualist reductivism, whereby particular instances of mental events are taken to be identical to particular instances of natural events. All these approaches retain the idea that our mental vocabulary refers to some kind of "thing" (therefore committing to factualism) and that the mind is somehow causally related to at least certain kinds of behavior (thus committing to the idea of mental causalism). Typically, straightforward reductivists and functionalists also maintain a commitment to the representationalist and computationalist conception of mind. By contrast, contemporary emergentisms (e.g., some post-cognitivist approaches) typically adopt a non-representationalist view of the mental, which emphasizes the embodied, embedded, and enactive character of perception, cognition, and action.

In line with functionalists and emergentists, discourse eliminativists admit that particular instances of mental events might be identical to their material realizers (thus maintaining a residual commitment to factualism and the idea of mental causation). However, this approach prioritizes research on the natural causes of behavior. In this sense, this kind of ontologically revisionary naturalism maintains that scientific research should progressively reshape our ontological assumptions.

Finally, ontologically radical naturalisms reject the very possibility of establishing any kind of identity theory, for the defining properties of the mental are nowhere to be found in a purely naturalist account of the world and of living beings. These straightforward eliminativist approaches take it that our interpretative folk-psychological practices constitute some

kind of fictive or non-literal use of language that has no scientific value. Thus, in principle, these approaches reject most of the core commitments of Cartesianism.

All these approaches to the mind-body problem have been (explicitly or implicitly) implemented in the field of mental health. The different therapeutic models can thus be seen as particular instances of more general philosophical approaches to the conceptualization of the place of mind -and its tribulations- on nature.

To begin with, supporters of the minimal interpretation of the medical model typically endorse an ontologically non-committal attitude; by contrast, stronger versions of the medical model (e.g., second-wave biological psychiatry), have classically drawn from straightforward reductivist assumptions regarding the nature and etiology of mental disorders. On the other hand, despite their opposing character, Szasz and early behavior analysts converged on their diagnosis of the conceptual flaws of this classical biomedical model; namely, a) that defining mental disorders in terms of certain clusters of behaviors and then attempting to explain these on the grounds of their associated diagnostic label constitutes a case of vicious circular reasoning; and b) that mental disorders, *qua* mental entities, don't exist. Thus, both constitute exemplars of a straightforwardly eliminativist approach to the role of mental concepts in the natural sciences, although each drew radically opposing conclusions from this assumption.

Subsequent approaches have tried to offer a middle ground position between the straightforward reductivist approach of second-wave biological psychiatry and the straightforward eliminativist approach underlying Szasz's critical approach and that of early behavior analysis. Tackling the latter's general dismissal of mentalistic explanations of behavior, cognitive approaches to psychotherapy, which later developed into cognitive behavioral therapy, reinstated mental states and processes as the core unit of analysis and treatment. These approaches can be seen as applied implementations of the functionalist view of mind that provided the framework for the more general "cognitive revolution" during the 1960's and 1970's.

On the other hand, in response to the "priority wars" in the fields of psychiatry and clinical psychology, Engel's biopsychosocial model implemented an emergentist approach, which understood mental health as comprising several inter-related scales of analysis. However, due to certain problems regarding its professed theoretical and practical eclecticism, contemporary approaches to mental health have raised a series of important criticisms, regarding: a) its vague account of how mental and non-mental properties causally relate; b) its inability to provide a properly integrative framework; and c) its inadequate account of the normative aspect of mental health practice.

Regarding the first problem, third-wave biological psychiatry and post-Skinnerian approaches within third-wave behavior therapy (namely, Acceptance and Commitment Therapy) yield two possible responses, both framed within a discourse eliminativist strategy. Third-wave biological psychiatry, although retaining the multi-level framework of the biopsychosocial model, aims to integrate different scales of analysis at the "brain circuitry level", and assumes that mental disorders essentially come down to brain disorders. According to this approach, talk of mental or cognitive functions and processes should be progressively refined by neurobiological research. By contrast, ACT proponents focus on a higher-order scale of analysis: the one corresponding to the relation between an organism and their natural and social environment. Contrary to the early behavior analysts' dismissal of cognitive explanations as mere "explanatory fictions", ACT doesn't dismiss them; rather, they redefine core cognitive concepts in terms of verbal or relational behavioral process and let middle-level terms into their scientific models as long as they prove to be predictive.

Finally, recent post-cognitivist approaches to mental health aim to provide a truly integrative framework. In this line, de Haan's (2020a, 2020b) recent enactive approach also draws from emergentism to overcome the tensions between dualist and "neuro-reductionist" approaches to mental health; however, her proposal differs from Engel's biopsychosocial model in that it advances a more developed view of the relation between mind, body, and action. In particular, de Haan's enactive approach takes it that mental properties are simply intrinsic to self-organizing biological systems (i.e., living beings). Her proposal directly tackles all the above-mentioned problems of competing approaches: a) the concept of "emergence" provides the key to giving an integrative account of how all scales of analysis relate without prioritizing any given level; b) it yields an attractive non-cognitivist approach to mental health; and c) it explicitly addresses the problem of normativity.

In section 2.4., however, we've seen why invoking a multi-level framework to account for this latter problem won't do the trick. The problem, pointed out by Szasz and other critical thinkers (see Laing, 1960/2010), lies in the tight connection between mind and normativity; mental vocabulary is primarily characterized by its normative force, or its capacity to rationalize behavior. By contrast, purely descriptive reports of an agent's material properties (e.g., their neural states, but also their self-organizational structure or their patterns of interaction with the environment) lack this normative force, no matter how complex they are nor how many scales of analysis they involve. In other words: conflating a personal-level approach to mental health within a multi-level subpersonal explanatory framework doesn't provide a proper account of the problem of normativity (Heras-Escribano et al., 2016; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014, 2020).

However, we've seen that Szasz's approach yields some problems of its own, such as, for example, its rejection of the possibility (or adequacy) of addressing human behavior in properly causal terms. Here we seem forced to choose between Szasz's exclusively personal approach to mental health and some of his opponents' exclusively subpersonal approaches. Both, however, yield ethical and conceptual problems: while the former precludes the scientific investigation of the causal determinants of people's "problems in living" and glorifies a steadfast libertarian, individualist, and blaming view of human beings and their misfortunes, the latter ultimately wipes off normativity and, with it, our capacity to account for what's exactly "mental" about mental health problems. As we view it, none of these options can provide a sound philosophical framework for mental health theory and practice. Here's the challenge: to find a way to reconcile normativity and our folk-psychological interpretative practices with the defining commitments of naturalism; or, as some authors put it, to provide a truly *having-it-both-ways* approach (see Varga, 2015; see also Fulford & van Staden, 2013; Thornton, 2007; Varga, 2015; Graham, 2010b).

In Chapters 3 and 4, we'll address this challenge. To do so, we'll argue, we must contest an assumption that both Szasz's and his opponents seem to rely on: that either we can reduce mental language to mere descriptions of material states of affairs, or our folk-psychological interpretative practices are incompatible with scientific explanations of behavior. In Chapter 3, we'll show how this assumption is anchored in the descriptivist view of mental language that we hinted at in section 2.1.4., which has significantly constrained the range of options considered when attempting to find a place for mind (and hence for mental health problems) on nature (Pinedo-García, 2020). In particular, we'll argue that it forces us to choose between reductivist or incompatibilist kinds of naturalism, on the one hand, and non-naturalism, on the other, none of which are satisfactory approaches to the problems of mind and normativity. In Chapter 4, we'll defend that, once we abandon the commitment to descriptivism, it's possible to make mind and normativity compatible with ontological naturalism without endorsing reductivisms or eliminativisms of any sort. We'll see that a proper answer to the problems of mind and normativity doesn't lie in conflating the normative character of the mental into some kind of multi-level causal account of behavior; in fact, we'll argue that it doesn't lie in either inflated nor deflated metaphysics. Rather, to reconcile mind and normativity with naturalism, all it takes is that we "accept the existence of a plurality of ineliminable explanatory approaches, some mechanistic, some agential, intentional and normative" (Pinedo García, 2020, p. 6.). In other words: we must turn our attention from ontology to semantics; from discussions about mental objects, relations, and properties, to

discussions about the variety of *language games* that we play when we try to account for each other's behavior.

Chapter 3
Descriptivism and the puzzle of translatability

In Chapter 2 we've seen which are the main approaches to the philosophy of mind that underlie the different therapeutic models in mental health research and practice. These constitute diverse attempts at challenging the Cartesian view of the mind and its characteristic ontological and epistemological commitments. Both in the philosophy of mind and in the philosophy of mental health, most debates have tended to focus on the ontological puzzles of Cartesianism, mainly related to the mind-body problem, i.e., the problem of the ontological status of mind and of the causal relation between mind and body -or, more broadly, between the mental and the non-mental or material.

However, as we pointed out, the Cartesian view of mind relies on a more or less implicit semantic commitment: descriptivism, or the idea that "the function of declarative sentences is to describe *facts* concerning worldly entities such as objects, properties, relations, events, etc." (Pérez-Navarro et al., 2019, p. 411). With regard to our folk-psychological interpretative practices, descriptivism amounts to the idea that declarative sentences that contain mental terms describe or represent some state of affairs (e.g., ontologically queer entities, brain states, self-organizing structures, relational responding patterns, etc.). In this chapter, we'll see how this implicit semantic commitment significantly constraints the range of plausible responses to the mind-body problem, which in turn leads to unsatisfactory answers to Descartes's main angst: the problem of normativity. We've already seen that, in the field of mental health, this leaves us unable to spell out what's specifically "mental" about mental health problems (Szasz, 1961/1974). Here, we'll point out what tragic consequences this mode or reasoning has for our conception of naturalism *itself*. The main goal of this chapter will thus be to lay out the main drives behind the commitment to "the dogma of descriptivism", as well as to identify possible ways out of its puzzles.

The structure of the chapter will be as follows. In section 3.1., we'll delve into the main assumptions that characterize descriptivism and what picture it renders of the meaning of mental-state ascriptions. We'll see how descriptivism about mental-state ascriptions is related to the standard or mindreading image of folk psychology (McGeer, 2007) and how a naturalized version of this commitment is at the heart of the mind-body identity thesis. In addition, we'll see how this descriptivist framework leaves only two possible ways out of substance dualism and the mind-body problem: a) reductive compatibilism (e.g., ontologically conservative and revisionary approaches) and non-reductive incompatibilism (e.g., ontologically radical approaches).

In section 3.2., we'll see why both kinds of naturalism yield unsatisfactory answers to the problem of normativity, due to their respective commitments to reductivism and incompatibilism. In consequence, both lead to a *self-defeating* kind of naturalism, i.e., one which defeats its own logical axioms. We'll also consider whether non-naturalism can provide a better approach to mind and normativity. However, we'll conclude that such kind of approach not only entails a return to the mind-body problem, but also implies a commitment to the idea of *private rule-following* (Wittgenstein, 1953/1958; see also Kripke, 1982), which leads to a flawed view of normativity and hence to a self-defeating kind of normativism. This is what we'll call "the puzzle of translatability", whereby naturalists are forced to choose between two competing, yet equally unappealing varieties of self-defeating naturalism. Under the descriptivist's dogma, however, their only way out of this puzzle is to endorse some kind of non-naturalism about the mental, not only off-putting in scientific terms, but also untenable from a normativist perspective. After exposing the argumentative rationale that forces naturalists into the puzzle of translatability, we'll see that the way out of this dilemma lies in rejecting the underlying commitment to descriptivism about mental language.

Finally, in section 3.3., we'll summarize the contents of this chapter and set the challenge to be accomplished in the following one: to develop a non-reductive, yet compatibilist account of the relation between mind and nature.

3.1. The dogma of descriptivism

As we saw in Chapter 2, most approaches to the philosophy of mind draw from a common rejection of substance dualism and a subsequent commitment to ontological naturalism, which entails monism (i.e., the assumption that there is only one general kind of states of affairs), materialism (i.e., the assumption that every actual or potential state of the world is of a scientifically describable nature), and the principle of causal closure (i.e., the assumption that every state of affairs must be the effect of a natural cause). However, they differ widely

as to whether the mind and mental properties are compatible or not with a defense of naturalism. Ontologically conservative approaches, on the one hand, assume that there's room for mentality within a naturalistic worldview; mental-state ascriptions capture relevant facts for explaining certain kinds of behavioral and cognitive phenomena (e.g., intentional or goal-directed behavior, inferential abilities, etc.). Ontologically radical approaches assume instead that the mental is individuated by a series of necessarily non-natural properties (privacy, self-causation, intentionality, etc.); hence mentalistic explanations and descriptions must be purged from a properly natural science of behavior. Finally, ontological revisionary approaches take a somewhat intermediate approach, assuming that mental-state ascriptions may capture some of the relevant facts for explaining behavior, albeit in a crude and imprecise way; thus, they would just constitute poor explanatory tools that a more developed cognitive or behavioral science may dispose of.

Despite their many differences, we saw that all these approaches stem from a common conception of our folk-psychological interpretative practices; what McGeer (2007) refers to as the *standard image* of folk psychology (see Chapter 2, section 2.2.1.). In this section, we'll first see how this standard or *mindreading* conception of folk psychology implies a descriptivist view of mental-state ascriptions. After that, we'll clarify what descriptivism amounts to, drawing a distinction between two fundamental kinds of descriptivism about mental-state ascriptions: internalist descriptivism and externalist descriptivism (Almagro-Holgado, 2021; Villanueva, 2014). Finally, we'll see how a naturalized version of descriptivism lies at the core of the mind-body identity thesis (see Chapter 2, section 2.2.2.).

3.1.1. Mindreading

According to the standard image of folk psychology, mental-state ascriptions play a fundamentally nomological or causal-explanatory role; they are theoretical devices that competent speakers deploy when they try to causally explain, predict, and control others' or one's own behavior. As McGeer (2007, 2015, 2021) and Zawidzki (2008; see also Zawidzki, 2013) have put it, the standard image construes our folk-psychological interpretative practices as attempts at *mindreading*: daily psychological interpretation is explained in terms of the formulation of hypothetical explanations of what might be "going on" in others' or one's own mind, in order to causally explain their past behavior and successfully predict future courses of action (see also Almagro-Holgado & Fernández-Castro, 2019; Fernández-Castro 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019).

To be sure, there are wide differences among mindreading approaches to folk psychology. Firstly, they differ on *how* mindreading is supposed to occur (e.g., whether inferentially or non-inferentially). As we saw in Chapter 2 (section 2.2.1.), traditional approaches like

theory-theory or simulation theory (see Carruthers & Smith, 1996) usually take it to be inferential (i.e., the result of inferring others' mental states from their behavior), although they differ on the kind of inferential mechanism that is supposed to be exploited in folk-psychological explanations: while the former assume that we rely on some kind of proto-scientific theory that causally relates mental events with observable behavior, the latter assumes that we use our own mental goings-on as a model to explain and predict others' doings (see Coliva, 2016; Fernández-Castro 2017a, 2017b; McGeer, 2007, 2015, 2021; Zawidzki, 2008). By contrast, postcognitivist approaches to mindreading like the Direct Social Perception model of social cognition (e.g., Gallagher, 2008; see Fernández-Castro & Heras-Escribano, 2019) take knowledge of other minds to involve non-inferential capacities; we directly perceive others' states of mind in their actions.

Although mindreading literature typically targets knowledge of other minds, approaches like the simulation theory bring up the issue of self-knowledge as well, i.e., the issue of how we "read" our own minds. On the one hand, some "first-personal" accounts of self-knowledge (see Coliva, 2016), specifically what have been called "observational" or "detectivist" models (see also Borgoni, forthcoming), commonly assume that we enjoy some kind of "special" or "privileged" self-knowledge abilities. According to these approaches, to determine whether one is in a certain mental state or not -i.e., to "read one's own mind"-, all one's got to do is to "look inside" or "turn the mind's eye inwards". As we saw in Chapter 2 (section 2.1.3.), this characterizes Descartes's epistemology, as well as many other approaches that ground the notion of "first-person authority" on such privileged self-knowledge capacities (see Borgoni, 2019, forthcoming; Srinivasan, 2015). By contrast, some "third-personal" accounts of self-knowledge (see Coliva, 2016) take it that we don't have any special or privileged epistemic access to our own minds; rather, to really "know ourselves", we need to read our minds from our behavior, our tendencies and dispositions, etc., just like others do when they attempt to "mindread" us (e.g., Schwitzgebel, 2002, 2013, 2021). As we'll see in Chapters 5 and 6, some of these competing self-knowledge models underlie the usual approaches to the proper conceptualization of delusions.

Secondly, mindreading approaches differ on *what* exactly we "read" when we mindread correctly. Some understand mental states in terms of *occurrent* events (e.g., particular instances of inner or manifest speech, mental imagery, sensorimotor contingencies, neural firings, etc.), while others view them as *dispositions*, i.e., organismic set-ups (whether cognitive, neural, sensorimotor, behavioral, phenomenal, etc.) that are individuated by the way an agent acts and reacts in certain circumstances (see Chapters 2 and 4, sections 2.1. and 4.2.1.). In addition, different approaches conceptualize the mental from different scales of analysis;

while some pay more attention to patterns of neural activity or hypothetical functional states and processes, others conceptualize mental states in terms of whole-body sensorimotor contingencies or the overall patterns of interaction between an organism and the environment. Finally, some approaches seem to assume that we read *nothing* at all when we mindread (or, on a milder version, that we read fictional stories), i.e., that there are *no facts* to be read off others' or one's own mind because there are no such things as minds (see Chapter 2, section 2.2.2.).

What we want to claim here is that, regardless of the large differences between approaches to folk psychology, all of them share the mindreading view of folk psychology, i.e., that our folk-psychological interpretative practices are primarily *nomological* or proto-scientific practices, aimed at the description, causal explanation, prediction, or control of behavior -at least, of intentional or goal-directed behavior. Different approaches implement this assumption in varying degrees of commitment to the ontological tenets of Cartesianism (i.e., substance dualism, factualism, mental causalism, and intellectualism). Some reject substance dualism, but maintain the other three. Others go as far as rejecting mental causalism and intellectualism as well, but still maintain a factualist conception of mind. Yet others reject all four, hence dismissing folk psychology as a mythical conception of human and non-human behavior. But the assumption that folk-psychological interpretation constitutes some kind of pre-scientific theoretical exercise remains. At root, this assumption entails the idea that mental-state ascriptions are first and foremost *descriptive* or *representational* devices, whose primary function is to describe or represent possible combinations of objects, properties, or relations among them (e.g., an agent's brain, their internal structure, their relationship with the environment, etc.). In other words: the mindreading conception of folk psychology is rooted in the commitment to descriptivism about mental vocabulary. In this sense, all mindreading approaches are somehow conceptually tied to the "logical mold" of Cartesianism (Ryle, 1949/2009, p. 9); and this, as we'll see, systematically leads to untenable forms of naturalism. Let's now see what descriptivism is about.

3.1.2. Descriptivism

First of all, we can make a rough distinction between two different -yet deeply entangled- descriptivist theses: one regarding the *pragmatic* aspect of language (i.e., regarding, roughly, what we *do* with words), and another one regarding the *semantic* aspect (i.e., regarding, roughly, what we *say* with words, the information communicated by means of words)[26].

[26] As we'll see in Chapter 4, we don't think that pragmatics and semantics, as we've defined them, can be neatly distinguished. However, we'll maintain this rough distinction here, for it helps to illustrate different senses in which one may assume a descriptivist view of language, which are often conflated.

Regarding pragmatics, we'll identify descriptivism with what Austin (1946/1961, p. 71, 1961, p. 221, 1962, p. 3) referred to as the *descriptive fallacy*, or, as we understand it, the idea that the only or primary function of language, the primary thing that we do with words, is to make assertions about what there is or what there's not in the world. Regarding semantics, we'll identify descriptivism with what Chrisman (2007) calls the *dogma of descriptivism*, which entails the assumption "that since semantic content of indicative sentences is standardly given in terms of their truth-conditions, the characteristic function of all indicative sentences is to describe worldly objects, properties, and relations" (p. 227)[27]. For now, we'll primarily delve into the implications of descriptivism at the semantic level of analysis, although the discussion of the pragmatic aspect will be of relevance in Chapter 4 (see section 4.1.).

Chrisman's (2007) definition of the dogma of descriptivism needs some unpacking though. As we understand it, the dogma of descriptivism entails the idea that the meaning or content of declarative sentences -i.e., those that make a statement or affirm or deny that something is the case- lies in a *description* or *representation* of some state of affairs -i.e., some possible combination of objects, relations, events, properties, etc. (Austin, 1962; Chrisman, 2007; Frápolli & Villanueva, 2012, 2013; Heras-Escribano & Pinedo-García, 2018; Pérez-Navarro et al., 2019; Pinedo-García, 2014, 2020; Price et al., 2013; Rorty, 1979; Villanueva, 2019; see also Almagro-Holgado, 2021). Accordingly, these possible states of the world establish the *truth-conditions* of the sentence, i.e., the conditions that have to be met for the sentence to be true.

There are two possible readings of this dogma. On the one hand, descriptivism may amount to an *affirmative* statement; namely, that declarative sentences always describe some given state of affairs. Thus, it's assumed that any possible expression that has the form of a declarative sentence is always representing some particular combination of events, objects, properties, and relations among them that may or may not obtain. We might call this a *shallow* version of descriptivism. As we'll see in this and the following chapter, this assumption is ill-founded; the meaning of at least some declarative sentences (mental-state ascriptions among them), in at least some occasions, is not exhausted by a description of any particular state of affairs (see section 3.2.2.; see also Chapter 4, section 4.2.1.). However, here we're interested in discussing a *deeper* version of the descriptivist dogma; one that implies a *conditional* statement about what kind of declarative sentences shall count as *meaningful*

[27] It is somewhat anachronistic to attribute such full-blown portrayal of this characteristic commitment to Descartes. However, several passages from his *Meditations on First Philosophy* and the replies to some of the objections suggest a similar conception of language, or at least of mental language; see, for example, the definition of "idea" in his reply to the Second Objections (Descartes, 2008, p. 102) or his reply to the second of Hobbes' Third Objections (Descartes, 1641/2008, p. 109).

proper. Such conditional can be stated as follows: only if a declarative sentence in fact represents some given state of affairs, then it has meaning proper or, more precisely, *content* or *cognitive meaning*. The notion of "content" or "cognitive meaning" here refers to the truth-evaluable information conveyed by a sentence when asserted by a speaker (i.e., *what* the sentence says), which in principle can be analyzed independently from how the sentence is formulated, or in what voice, tone, or attitude (i.e., *how* it's said) (Blackburn, 2008, p. 65)[28]. So, thus understood, the dogma of descriptivism not only implies the affirmative statement that declarative sentences always describe possible states of the world, but also the conditional statement that only if a declarative sentence *successfully* represents possible states of the world, then it's truth-apt or truth-evaluable, i.e., can be assessed in terms of its truth or falsity. Importantly, "truth-apt" here only applies to those sentences whose truth or falsity is a *contingent* matter -i.e., those that represent states of affairs that might be the case or not - but not those whose truth or falsity is necessary (Wittgenstein, 1921/2001; see also Villanueva, 2019).

As we'll see below (section 3.1.3.), this "deeper version" of descriptivism would correspond to the main tenet behind the early Wittgenstein's so-called *picture theory of language*, which he developed in *Tractatus Logico-Philosophicus* (Wittgenstein, 1921/2001). Some authors take "descriptivism" proper to amount to what we've called "the shallow version of descriptivism", and *non-cognitivism* to what we've called the deeper version, i.e., the idea that those declarative sentences that fail to represent some particular state of affairs (e.g., those containing ethical, epistemic, or logical expressions, among others) do not have cognitive meaning, or do not express something truth-evaluable (see Frápolli & Villanueva, 2012). However, there're several reasons why we've preferred to consider both commitments as two varieties of the descriptivist dogma, instead of distinguishing between "descriptivism" and "non-cognitivism". On the one hand, we've done so to distinguish this sense of the term "cognitivism" from its more popular variety in psychology and philosophy of mind, where "non-cognitivism" refers to the rejection of the traditional approaches to cognitive science (see Chapters 1, 2, and 8). On the other hand, many traditional approaches to the analysis of certain regions of language (e.g., moral and mental vocabulary) have taken the shallow and deep versions of cognitivism to go hand in hand in (see Frápolli & Villanueva, 2012). In what

[28] For example, assuming that Aquamarine and Mustard are a couple, the sentences "Aquamarine gave Mustard an instant camera as a present", "Dammit, Aquamarine gave Mustard an instant camera as a present", "Mustard was given an instant camera as a present by Aquamarine", and "Mustard's partner gave Mustard an instant camera as a present" have the same cognitive meaning (i.e., they "say the same thing"). "Cognitive" is used here as "informative" or "truth-evaluable", in order to distinguish this from other possible modalities of meaning (e.g., "expressive meaning", which has to do with the speaker's attitude towards what is said) (Blackburn, 2008, p. 65).

follows, unless specified, we will use "descriptivism" as a synonym of "representationalism about language" (see Pinedo-García, 2020; Price, 2011; Price et al., 2013; Ramberg, 2000, Rorty, 1979; see footnote 18); that is, to refer to its deeper version.

Once we accept the dogma of descriptivism, there might be different criteria for declarative sentences to count as "properly descriptive", "successfully representing some state of affairs", "expressing a proposition", or as "cognitively meaningful". Syntactical correction, for example, might be taken to be an example of such criterion. Different implementations of descriptivism can thus be characterized in terms of the criteria that they impose on sentences to constitute an appropriate description of the world. As we'll see in section 3.1.3., ontological naturalism can be reconstructed as the imposition of certain restrictions on what counts as *possible* states of the world, and thus on what claims count as truth-apt.

Thus explained, this conception of meaning and language may seem somewhat bizarre and complex. However, it's relatively intuitive. Take, for example, the following two sentences:

(1) Fuchsia is preparing a lentil salad
(2) There's a square circle on the table

It would be relatively intuitive to say that the sentence (1) "stands for", "represents" or "describes" certain state of affairs, i.e., a possible situation where a certain relation (e.g., the action depicted by the present continuous tense of "prepare") holds between two objects (e.g., Fuchsia and the lentil salad). To put it another way, the words that form (1) are symbols that, in the particular combination in which they appear in the sentence, describe or represent a possible state of the world, in virtue of which the sentence can prove to be true or false. Critically, it's a *contingent* matter whether such possible state of the world is the case or not; if, in fact, Fuchsia is preparing a lentil salad, then (1) will be true; if not, it will be false.

On the contrary, it also seems natural to assume that (2) expresses no cognitive content at all nor provides any kind of information about the world whatsoever. Since a "square circle" is a logical impossibility, (2) simply cannot be assessed in terms of its truth or falsity; we can't even start to figure out what kind of object or relation among objects would "satisfy" (2) or make it true. Therefore, (2) is *necessarily* false.

Things get messy when we try to apply this view of language to mental-state ascriptions. Applied to the kind of expressions that we use when we engage in folk-psychological interpretation, descriptivism entails the view that mental-state ascriptions either describe

some possible state of the world or have no cognitive content at all. Take for example the following sentence:

(3) Saffron *believes that* Fuchsia is preparing a lentil salad[29]

For now, let's assume that (3) actually describes some state of affairs. What does it represent then? Almagro-Holgado (2021) has recently distinguished between two possible ways of implementing the descriptive stance to the analysis of the meaning of sentences like (3): *internalist descriptivism* and *externalist descriptivism* (see also Villanueva, 2014). Keeping in line with the previous analysis of the meaning of (1), *internalist* descriptivists take (3) to also represent a relation between two objects; although, this time, the represented relation is established between a material object (i.e., Saffron) and a mental, internal, private object: the proposition expressed by (1), i.e., «Fuchsia is preparing a lentil salad». Specifically, the mental process verb in (3) would indicate the kind of relation that Saffron bears to the proposition: that characteristic of *believing*, which would in principle be different from the one that characterizes desire, expectation, intention, and other propositional attitudes.

Russell's (1913/1992) relational theory of the meaning of psychological predicates (see Almagro-Holgado, 2021; Thornton, 2007; Villanueva, 2019) provides an early explicit formulation of this kind of descriptivist analysis, but it's main tenets are inherent to traditional cognitivist approaches to mind. Compare our analysis with Fodor's (1987; see also Thornton, 2007) explanation of the first of the two central claims that he views as characteristic of the representationalist view of mind:

> Claim 1 (the nature of propositional attitudes): For any organism O, and any attitude A toward the proposition P, there is a ('computational'/'functional') relation R and a mental representation MP such that MP means that P, and O has A iff O bears R to MP. (Fodor, 1987, p. 17)

So, according to this view, (3) would represent Saffron as entertaining a mental representation of Fuchsia preparing a lentil salad. Often -although not necessarily-, internalist descriptivism goes in hand with "observational" or "detectivist" models of self-knowledge, like the one endorsed by Cartesianism (section 3.1.1.; see also Chapter 2, section 2.1.3.). By contrast, other descriptivist accounts of the meaning of mental-state ascriptions reject that

[29] Note that (2) contains (1) (i.e., "Fuchsia is preparing a lentil salad"). Since (1) is taken itself to express a proposition, the mental states that commonly appear in sentences with "that-clauses" -e.g., sentences containing mental process verbs like (2)- are typically referred to as *propositional attitudes*, since they indicate a certain attitude (e.g., the attitude of belief, of desire, etc.) towards a certain proposition.

what sentences like (3) represent is some relation between an agent and some private, internal object. *Externalist* descriptivists take it instead that what (3) represents is some external, *public* state of affairs, e.g., typically, the agent's behavior or some aspect of their relation with their natural or social environment -although it could also be the agent's neural states, since these are also subject to public scrutiny. According to this view, (3) would represent Saffron as, for instance, behaving in certain ways (e.g., saying "Fuchsia is preparing a lentil salad", searching for Fuchsia in the kitchen, etc.). In this sense, externalist descriptivism often goes in hand with those "third-person" approaches that understand self-knowledge as a matter of knowing how one would act or react in certain circumstances (see Almagro-Holgado, 2021; Villanueva, 2014).

So far, we've seen what different descriptivist analyses of the meaning of mental-state ascriptions may look like. As we'll see, a fundamental line of divergence among the different approaches to the mind-body problem turns on whether these analyses are actually possible, or whether they render something useful for the behavior sciences.

3.1.3. Naturalizing description

We've seen that descriptivism, in its deepest version, entails that only *descriptions* (i.e., declarative sentences that successfully describe or represent some state of affairs) are truth-apt (i.e., might be declared true or false), and that only truth-apt sentences are cognitively meaningful (i.e., informative about the world). This much is shared, we argue, by both naturalists and non-naturalists about the mind.

Their differences, however, turn on what each count as *possible* states of the world, and hence what may count as a "successful" description. Non-naturalists about the mind (e.g., Cartesianism) don't impose any significant restriction on the range of possible objects, properties, events, or relations, that might be described by mental-state ascriptions; if anything, they adopt an expansive attitude in this matter. Substance dualism, for instance, allows for two possible kinds of states of affairs: those corresponding to the extense or natural world, and those corresponding to the cognitive, non-natural world. The natural sciences can only tell us everything there is about the former; the world of consciousness, on the other hand, can only be properly described by turning "the mind's eye inwards". Naturalists, on the other hand, assume that the range of possible objects, events, properties, and relations among them is exhausted by what the natural sciences might tell us about the world. In this sense, the principles that define ontological naturalism (i.e., monism, materialism, and the principle of causal closure) can be recast as restrictive criteria on what counts as possible states of the world and, consequently, what may sensibly count as a "description" proper.

Let's see this in more detail. To begin with, we may call many different things a "description". Novelists describe their characters' physiognomy, their actions, the landscapes where those actions take place; but they also describe their characters' mentality and personality, their intentions, beliefs and values, as well as the enchanting or terrifying character of the landscapes that these characters inhabit. However, the sense of "description" that interests us here is one that may apply only to the former kind of things, but -as some argue- not to the latter. To illustrate what sense of description is at stake here, we must turn to Wittgenstein, whom is commonly credited both for one of the most thorough elaborations of a descriptivist conception of language and, at the same time, for some of the most influential and illuminating efforts to debunk such conception. Villanueva (2019) has argued that there's a particular notion of description that seems to remain constant in Wittgenstein's thought, from the formulation of his "picture theory of language" in the *Tractatus* (1921/2001) -what is commonly referred to as "the first" or "the early Wittgenstein"- through the later stages of his philosophical production -what is typically referred to as "the second" or "the later Wittgenstein" (Wittgenstein, 1953/1958, 1969). For Wittgenstein, descriptions are "empirical propositions (e.g., ones which describe a visible distribution of objects in space and could be replaced by a representational drawing)" (Wittgenstein, 1974, § 82). Villanueva (2019) elaborates on Wittgenstein's view of descriptions as follows:

> Descriptions are empirical propositions (...), and we cannot describe or *say* what cannot be either the case or not the case. What cannot be described, can only be *shown*[30] (...). Wittgenstein claims in [*Philosophical Grammar*, §7,] that to understand a description is to form a picture of what is described, and descriptions are characterized by being subject to empirical, causal, spatial-temporal restrictions. (Villanueva, 2019, p. 157, author's translation)

We can find two core commitments of the naturalized version of descriptivism in Wittgenstein's picture theory of language. The first amounts to the idea that only those sentences which are contingently true or false are truth-apt or truth-evaluable sentences, and hence cognitively informative or meaningful. The second and most important one for our purposes here is that only descriptions (i.e., sentences that describe possible arrangements of empirical, spatially and temporally distributed states of affairs) are truth-apt. Thus, the characteristic criterion that this naturalized version of descriptivism imposes on sentences to be cognitively meaningful or "contentful" (i.e., to successfully represent some state of

[30] "Say" and "show" are the terms that Wittgenstein employs in the *Tractatus* to distinguish between what can be assessed in terms of truth or falsity and what can only be shown in action, but not "said" with meaning proper.

affairs) is that they are not only syntactically well-formed, but also apt to some conceivable kind of empirical investigation.

On this view, only sentences like "It's already March and we haven't yet removed the Christmas tree" would count as cognitively meaningful, since they represent a possible combination of spatially and temporally distributed objects, properties, events, and relations which may be the case or not. It might be true or false that it's still March and we haven't yet removed the Christmas tree; in any case, the issue is subject to empirical investigation. On the contrary, sentences like (2) above (i.e., "There's a square circle on the table") wouldn't count as cognitively meaningful, since "square circles" are a logical impossibility; we wouldn't even know what we're searching for if we tried to see if (2) is true or false. In this sense, sentences that express necessarily true or false propositions wouldn't count as descriptions either (nor thus as properly contentful or meaningful sentences): syntactically well-formed declarative sentences like "I love doing research, but I don't love doing research" or "I am either writing my PhD dissertation or else" -*contradictions* and *tautologies*, in Wittgenstein's terms- wouldn't count as having cognitive meaning. But neither would those sentences that, when properly analyzed, seem to aim at saying (i.e., describing) what cannot be said, i.e., what doesn't count as empirical, spatially and temporally distributed states of affairs. For the early Wittgenstein (1921/2001), these comprised metaphysical, esthetical, moral, and -essentially for our present purposes- psychological predicates. He called these *pseudopropositions*, since they only seem to be saying (i.e., describing) something, but don't really say anything: they just point to the limits of language, of "what can be said" with sense. Critically, for the early Wittgenstein, "what can be said" amounted to the "propositions of natural science" (Wittgenstein, 1921/2001, § 6.53, p. 89).

In sum, this naturalized form of descriptivism amounts to the idea that only "propositions of natural science" or "empirical propositions", which describe spatially and temporally distributed states of affair, count as descriptions proper, and that only descriptions proper are evaluable in terms of their truth or falsity (i.e., are truth-apt), hence being contentful or carrying "cognitive meaning".

We can now see more clearly how the mindreading conception of folk psychology and its core descriptivist assumption have shaped contemporary debates on the ontological status of mind. In a nutshell, all mindreading approaches are committed to the central idea that mental-state ascriptions pursue a descriptive goal: representing possible states of the world that may be causally relevant for explaining each other's behavior. Against this background, the characteristic commitments of ontological naturalism (i.e., monism, materialism, and the principle of causal closure) introduce specific *constraints* on the states of affairs that are

logically permissible within a naturalistic worldview: mental-state ascriptions must represent empirical, spatially and temporally distributed states of affairs (as dictated by the principles of monism and materialism), which must be circumscribed to the laws of causality (as dictated by the principle of causal closure). Otherwise, their truth -and hence their truth-aptness, if we follow Wittgenstein (1921/2001)- is incompatible with a defense of ontological naturalism.

Ultimately, these assumptions lie at the core of the mind-body identity thesis that we saw in Chapter 2. Descriptivism about the mental, together with the defining maxims of ontological naturalism, yield the logical necessity for establishing an identity relation between mind and body (or, more generally, between mind and nature). We might call this the *translatability assumption*: if and only if mental-state ascriptions are *semantically identical*, and thus *translatable* or *reducible* to explicit descriptions of material events (i.e., if they describe or represent the same spatially, temporally and causally circumscribed state of affairs), then folk psychology is compatible with ontological naturalism. Compatibility, thus, stands or falls with the mind-body identity thesis.

This much is shared by most ontological naturalists; what they seem to differ on is on two fundamental points: a) the degree of applicability of the mind-body identity thesis (and thus the *compatibility* of folk psychology with behavioral science); and b) the degree of scientific legitimacy that folk psychology is endowed with.

On the one hand, ontologically conservative and revisionary approaches uphold some kind of *reductive compatibilism*, according to which mental-state ascriptions are in some sense identical to descriptions of material events and thus compatible with a naturalistic worldview. Reductive compatibilists differ on whether this identity relation holds in a context-free way, as type identity supporters assume (i.e., straightforward reductivists), or rather in a context-relative way, as token identity supporters defend (i.e., functionalists, emergentists, and discourse eliminativists) (see Chapter 2, section 2.2.2.1.). Be that as it may, while ontological conservatives endow folk-psychological interpretation with full scientific legitimacy -and hence their proposed theory changes entail the preservation of folk-psychological concepts-, ontological revisionists endow it with limited or merely provisional scientific legitimacy -and hence their proposed theory changes rather entail the progressive shaping of our folk-psychological assumptions.

On the other hand, ontologically radical approaches assume a *non-reductive incompatibilist* view of the mental, according to which no kind of identity relation can be established between the mental and the non-mental; hence, mental talk is incompatible with a naturalistic worldview. These approaches assume that mental facts can *never* obtain (i.e., can

never be the case), since their core defining properties (e.g., self-causation, intentionality, normativity, privacy, etc.) are nowhere to be found in nature. Take intentionality, for example, or the idea that "the mark of the mental" is that mental states are *about* something (Jacob, 2019; see footnote 16). Regarding (3) above, one may describe Saffron's neural states or their Pavlovian or operant responses, but none of these "stand for" (i.e., are symbols of) Fuchsia preparing a lentil salad; no description of some spatially, temporally, and causally restricted state of affairs can capture the "aboutness" of Saffron's belief. Ontological radicals thus endow folk psychology with no scientific respectability at all; for them, a proper theory change involves banishing folk-psychological talk from the explanatory apparatus of the behavioral sciences.

This approach, however, needs closer inspection. Straightforward eliminativist approaches like the ones we saw in Chapter 2 often draw from similar observations to claim that mental-state ascriptions are just *false*; at best, these approaches see folk-psychological interpretation as a "convenient fiction", i.e., a non-literal, socially convenient linguistic practice, which may be useful for certain purposes, but which is nonetheless false in literal terms[31]. Note, however, that if they take their underlying commitments seriously, they must say more than just that. If the core defining properties of the mental (e.g., self-causation) are, by principle, *impossible* within a naturalistic worldview, then mental-state ascriptions wouldn't just "happen to be always false" (i.e., they wouldn't be just *contingently* false); rather, they would be *necessarily* false. Now, if we take Wittgenstein's naturalized version of description seriously, ontological radicals are forced to go one step further: it's not just that mental-state ascriptions fail to represent something *true*; rather, they fail at *representing* whatsoever. Since mental entities, given their defining properties, are impossible, mental-state ascriptions would constitute attempts to describe what cannot be described, and hence would be non-truth-apt, cognitively meaningless sentences. Just like the verses from Lewis Carroll's poem *Jabberwocky*, mental-state ascriptions like (3) would simply *seem* to describe or say something, but they wouldn't; (3) would thus be analogous to "All mimsy were the borogoves, and the mome raths outgrabe".

So far, we've seen how all the approaches to the philosophy of mind that we saw in Chapter 2 (and hence their respective applications to the field of mental health) can be unitarily characterized by a series of shared commitments; namely, their view of folk

[31] Thus stated, this amounts to what has been called *fictionalism* about a certain region of discourse (folk psychology in this case; see Demeter, 2013; Parent, 2013; see also Price et al., 2013). Since the consequences of such position are identical to those of straightforward eliminativism for our present discussion, we won't differentiate between these two approaches.

psychology as a pre-scientific kind of theory and, most importantly, their underlying commitment to descriptivism in its deeper version. This leads to two major kinds of naturalism: reductive compatibilism, which reconciles mind and nature through some version of the mind-body identity thesis, and non-reductive incompatibilism, which rejects the latter and hence compatibilism altogether. In the next section, we'll see that neither are conceptually tenable, for both lead to some *self-defeating* kind of naturalism.

3.2. The puzzle of translatability

In Chapter 2 (section 2.4.) we already glimpsed some of the practical consequences of not taking normativity seriously in the field of mental health. The basic problem, as we saw, revolves around the fact that mentality and normativity seem to go hand in hand: eliminate or reduce the former, and you lose the latter. In this section, we'll first focus on how this problem affects the very conceptual plausibility of both reductive compatibilism and non-reductive incompatibilism about the mental. As we'll see, the problem with the former lies in its reductivism, while the problem of the latter resides in its incompatibilism. After that, we'll consider adopting a non-naturalist approach, only to conclude that, besides its scientifically off-putting character, it doesn't render a proper account of the relation between mind and normativity either. Finally, we'll lay out the basic argumentative structure that forces naturalists to choose between reductivism and incompatibilism, hence paving the way to out of this dilemma.

Before we move on, it's important to note that the argument exposed here has been developed by many authors within different philosophical traditions (see Caro & Macarthur, 2004, 2022; Giladi, 2019). However, rather than following one or another version of this general critique to more standard or prevailing versions of naturalism -"scientific naturalism", as Hutto (2022) calls it, or "object naturalism", as Price (2004) does- we've preferred to elaborate a unified argument. Specifically, we've mainly drawn from sources that, to some extent, share in common the pragmatist understanding of language that characterizes Wittgenstein's later work (McDowell, 2004; Price, 2004; Price et al., 2013; Rorty, 1979; see also Thornton, 2007), which we'll expose in Chapter 4.

3.2.1. The virtues of non-reductivism and compatibilism

First of all, we might wonder what are the key attractive features of non-reductive incompatibilism and reductive compatibilism. What makes these naturalist alternatives so appealing for many?

On the one hand, as we view it, the attractiveness of the former lies in their ability to free scientific inquiry from pre-experimental constraints on what might be the actual causes

of whatever behavioral, bodily, or neural patterns of interest; in doing so, they contribute to advance scientific research and enhance our explanatory and intervention powers. Let's call this *Nomological Power*, which can only be accommodated if we assume a *non-reductivist* approach to the mental, i.e., one that rejects the mind-body identity theory.

The key realization here is that, as many have pointed out, folk psychology is just too *vague* to be of use as a roadmap for the behavioral sciences: we often disagree over whether someone really has certain beliefs, desires or intentions (see Chapter 5; see Curry, 2020) and, as it should be expected, laypeople are more prone to fail when they attempt to causally explain, predict, or control other's behavior (e.g., see Zawidzki, 2008). Behavior analysts, for example, have given overwhelming amounts of evidence that topographically identical behaviors, which we typically explain in natural language in terms of the same mental-state ascriptions, often are under control of completely different environmental contingencies (e.g., see Epstein et al., 1980, 1981; Skinner, 1953, 1974; see also Chapter 8). The argument from multiple realizability that we saw in Chapter 2 (section 2.2.2.1.) also points in the same direction. All in all, to restrain our scientific theories within the vague mould of folk-psychological concepts amounts to setting the bar relatively low for the explanatory power, predictability, and controllability standards of contemporary science. Non-reductivism, by contrast, frees scientific inquiry from folk-psychological presumptions and leaves us in a better position when it comes to causally explain, predict, and control human affairs.

On the other hand, the attractiveness of reductive compatibilism is that it seems to be able to capture an intuitively true claim: that our folk-psychological interpretative practices are *literally* truth-evaluable, i.e., that mental-state ascriptions can be assessed in terms of their truth or falsity. Let's call this attractive feature *Truth-Aptness*. "Truth-Aptness" can only be properly accommodated if we defend a *compatibilist* view of the place of mind on nature, i.e., if we find a way to remain committed to the idea that mental-state ascriptions are truth-apt without populating our ontology with queer, spooky entities. Reductivism is the usual way to go in this sense.

To see the intuitive grip of "Truth-Aptness" more clearly, consider the following two examples:

> PHILOSOPHER'S STONE: Harry, Ron, and Hermione are three young wizards, first-year students in Hogwarts. After a year full of strange events, the three of them come to believe that Prof. Snape wants to steal the philosopher's stone. However, the three wizards are mistaken: it's Prof. Quirrell who wants to steal it; in fact, Prof. Snape, who suspected this for a long time, is trying to stop him. Prof. Quirrell knows that the three kids believe that Snape is the one who wants to steal the philosopher's stone, which helps him keep his true identity and evil

intentions hidden. He also knows that Snape knows that he intends to steal the stone, so he must exercise caution if he is to get away with it.

POLAROID LOVE: Aquamarine and Mustard are a couple which have recently decided to move together. Next month it's their anniversary, and Mustard intends to give Aquamarine an instant camera as a present. After a night out, due to some comments that Aquamarine has made, Mustard begins to suspect that she's planning to buy the exact same thing for her. Since she wants to be the one who gets the camera, Mustard now plans to fool Aquamarine by making her believe that she finds it absurd to buy an instant camera. Whenever she has a chance, Mustard starts making negative comments about instant cameras, such as "those cameras are just too hipster for us", or "their photo paper is absurdly expensive". At some point, however, Aquamarine realizes what Mustard is trying to do, and she decides to play the same game. Now, every time that Mustard makes a negative comment about the cameras, Aquamarine plays along; she effusively agrees with everything Mustard says and even highlights other negative aspects. Eventually, Mustard begins to have doubts: has Aquamarine discovered her game? Or has she truly convinced her of the absurdity of buying an instant camera?

These examples depict two cases –the former fictional, the latter based on a true story– of what we might call *mind games*: in these mind games, the different "players" attempt to correctly guess which are the mental states of the other players involved (their beliefs, desires, intentions, etc.) or influence them for diverse practical purposes (e.g., procuring or impeding the return of Lord Voldemort, surprising one's partner, etc.). These mind games essentially are interpretative games, based on our ability to correctly *interpret* and *explain* each other's behavior in terms of mental states, i.e., to correctly ascribe mental states and to correctly assess the truth or falsity of other's mental-state ascriptions and self-ascriptions.

The intuitive grip of this explanation is however lost when we assume an incompatibilist position regarding the mental: if mental-state ascriptions are just mere fictions, or not even truth-apt, then none of these mind games would make sense. In fact, we wouldn't even be able to distinguish the former from the latter: both would count as "fictional" or plainly "senseless". This seems untenable: no sensible account of the mental should lead to a conflation of young fictional narratives with real-life examples of our interpretative practices. What reductive compatibilism offers us, as we've seen, is a way to remain faithful to ontological naturalism without leading us too far astray from the intuitive idea that folk-psychological interpretation can be *literally* true or false. The solution is the following: to preserve

"Truth-Aptness" via the assumption that the truth-marker of our mental-state ascriptions is some given natural fact (e.g., some given pattern of behavioral or neural activity).

However, we're not just interested in preserving the truth-aptness of mental-state ascriptions in and of itself. Rather, one of the main motivations behind compatibilist approaches to the mental lies in the tight connection between mind and normativity that we sketched out in Chapter 2 (section 2.4.). In other words: if we want to preserve the truth-aptness of our folk-psychological interpretative practices, it's first and foremost because we want to understand how mental-state ascriptions *rationally explain* or *rationalize* each other's doings. Let's call *Normative Force* to the idea that mental-state ascriptions have such rationalizing properties. This, as we'll now see, is where the problems of reductivism begin.

3.2.2. The perils of reductivism and incompatibilism: self-defeating naturalisms

Another example will be useful to highlight the tight connection between mind and normativity that we've been talking about so far. Consider the following continuation of the POLAROID LOVE example above:

> In a desperate attempt to get ahead of Aquamarine, Mustard has designed a futuristic *teleanalyzing* device that allows her to track, on a moment-by-moment basis, both Aquamarine's brain activity and her patterns of interaction with the environment. On Friday morning, Aquamarine tells her friend Emerald, a foreign student that she's met at university, that she intends to buy the instant camera that very afternoon, because she believes that it's already available for purchase and she wants to buy it already. Consequently, Emerald expects that Aquamarine will go to the shop to buy the camera that same afternoon. Luckily, they were talking in English, because Emerald doesn't know a word of Spanish; thus Mustard, who was observing the whole scene through her teleanalyzing device, and who isn't exactly fluent in English, has not understood a single word. Nonetheless, around the same time, and after spending the whole week analyzing both Aquamarine's behavioral and neural patterns of activity, Mustard has come to the same conclusion as Emerald: that Aquamarine will go to the shop to buy the instant camera that same afternoon. To everyone's surprise, however, Aquamarine spends the afternoon at home and doesn't go to the shop.

Would Mustard and Emerald react in the same way? Probably not. The main difference between Emerald's and Mustard's prediction is that the former is based on what would be *rational* to expect from Aquamarine, while the latter merely constitutes an *empirical* prediction. The link between Aquamarine's self-ascribed beliefs, desires, and intentions is fundamentally *logical*: given their mental self-ascriptions, *if Aquamarine is rational*, then she should go to the shop and buy the camera. If she doesn't, then Emerald would be entitled to

say that either Aquamarine's mental-state self-ascriptions were false (i.e., she didn't *really* want to buy the camera that afternoon, she didn't really believe that the camera was available for purchase, or she never intended to go to the shop that afternoon), or that her behavior was in some sense *irrational*. By contrast, this presumption of rationality is not a part of Mustard's scientific prediction; the relation between Aquamarine's patterns of brain activity or interactions with environmental contingencies and her actual behavior is not logical, but merely empirical or causal. It wouldn't make sense of Mustard to say that Aquamarine is irrational; at best, what she can say is that Aquamarine's behavior is "statistically unexpected", or that her scientific prediction was wrong because she missed some potentially relevant variable.

The key point is this: mental-state ascriptions *rationalize* action (i.e., make it intelligible) in a way that mere descriptions of one's neural or behavioral patterns cannot -hence the former cannot be reduced to the later. This is precisely why mind and normativity are so intimately connected, and it's this intimate connection what Descartes was trying to preserve from the ever-expanding scope of the natural sciences. Mental beings are autonomous, responsible agents, whose deeds are evaluable in terms of merit or demerit. By contrast, on the "disenchanted" view of nature (McDowell, 1996) that characterizes much of contemporary science, agents are reduced to mere *arational* creatures, whose "nature differs only in degree of complexity from clockwork" (Ryle, 1949/2009, p. 8), and whom cannot be made responsible for their actions. For an individual can only be made responsible for their actions if they can *err*, i.e., if their actions depart from some *norm* (whether it can be made explicit or not) (see Heras-Escribano et al., 2016; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014, 2020); in other words, it must be possible to distinguish between their *correct* and *incorrect* courses of action. In a sense, there's no such thing as error in nature, nor hence success; there's just variability.

Mental language, by contrast, allows us to do just that. When Aquamarine says that she intends to buy the instant camera that same afternoon and that she believes that it's already available, she's *rationally compelled* to behave in certain ways and not in others: she should go buy the camera that afternoon, and not another day; she should buy the instant camera, and not another thing; she should answer "at the shop!" if she were asked "where are you are you going to buy the camera?" -and not, say, "at the space station", etc. By contrast, purely descriptive reports of her behavioral tendencies or neural states don't carry with them this normative or prescriptive force. Establishing an empirical connection between certain brain states and certain behaviors just allows us to predict and control what an individual will *in fact* do, but remains silent on whether they *should* or should not behave

in that way (see Heras-Escribano et al., 2016; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014, 2020). Attempting to reduce the latter to the former constitutes an example of the *is-ought fallacy* (Heras-Escribano & Pinedo-García, 2018), or a species of Moore's *naturalistic fallacy*, in Sellars' (1956) terms[32].

As we already mentioned, non-reductive incompatibilists come this far: they recognize that this normative force is of the essence of mental-state ascriptions, i.e., that *genuine* mental-state ascriptions have this prescriptive role, and thus they cannot be reduced to mere descriptions of spatially, temporally, and causally-bound states of affairs. However, they draw from this to assume an incompatibilist approach to mental-state ascriptions: if they're not reducible, then they are either literally false or express no truly informative content at all. For them, there's no room within a naturalistic worldview for "rational explanations"; all we might do is to causally explain, predict, and control.

Thus, in order to remain faithful to a naturalist conception of the world, non-reductive incompatibilists have to swallow the bitterest pill: no talk of freedom, agency, rationality, or intentionality, etc. is *really* or literally truth-apt; they're merely "pseudopropositions", "explanatory fictions" at best. This has gloomy -not to say disastrous- consequences for many of our normative practices, some of which we already advanced in Chapters 1 and 2: in the disenchanted, arational picture of human affairs, none of the normative considerations and distinctions that constitute the very core of many human practices (law, politics, ethics, etc.) would be "cognitively meaningful". There wouldn't be anything "literally" true or false to say about the *rightness* or *wrongness* of slavery vs. liberation, genocide vs. peace activism, imprisonment of political dissidents vs. democratic legal systems, pathologizing non-normative gender identities vs. diversity recognition, etc.

At this point, the most steadfast defenders of non-reductive incompatibilism might want to stick to their guns -after all, "science doesn't care about our feelings", right? If so, they might "bite the bullet", as it's said, and buy into what we might call *the tragic stance*: if the price of letting go all talk of mentality is to let go all talk of normativity, then so be it: let's forego all talk of "rationality", "intelligibility", "morality", "meaning" -even "truth"; let's forego all talk about the merit or demerit, correctness or incorrectness of human affairs.

[32] The modern formulation of the is-ought problem is typically attributed to Hume (1711-1776) and his *A Treatise on Human Nature* (1738 – 1740; see Hume, 1739/1896). In a nutshell, this problem consists in the impossibility to deduce normative conclusions (i.e., which prescribe how things should be, or that establish a distinction between "correct" and "incorrect" states of affairs), from exclusively descriptive premises (i.e., which simply establish how things in fact are). In meta-ethics, Moore (1903/1922) employed a similar reasoning in his "open question argument" against what the "naturalistic fallacy", or the idea that it's possible to reduce or translate moral sentences to sentences where moral terms are replaced by mere descriptions of some state of affairs (see also Heras-Escribano & Pinedo-García, 2018).

Let's just focus on their explicability, predictability, and control. We just fancy our manifest image of the world and ourselves because we're so accustomed to it, because our learning histories have led us to speak in such terms, or because our brains create such *illusion*; let's, however, forego that pre-scientific, self-righteous, almost religious conception of ourselves as mindful, free, autonomous, and responsible agents; let's embrace the ultimate contingency and meaninglessness of it all, our merely organismic, arational condition, as much bound to the merciless laws of causality as any other natural phenomenon; let's forego the *personal* and embrace the *subpersonal*.

Tragic enough. The problem here is that, as many authors have pointed out, this kind of savage, nihilistic scientism leads us directly to a form of *self-defeating naturalism* (see Caro & Macarthur, 2004, 2022; Hutto, 2022; Pérez-Álvarez, 2011; Pinedo-García, 2014; Price, 2004). It is self-defeating because it undermines its very foundational premise: ontological naturalism. After all, the principles of monism, materialism, and causal closure *are not* empirical propositions (see section 3.1.3.); they do not represent any spatially and temporally distributed states of affairs that might be or not the case. They're *axioms*: we don't conclude them after a thorough empirical research; we just assume them as our "pre-analytic" premises (Barnes-Holmes et al., 2001; Hayes et al., 2001). They rather seem to be *grammatical* or *hinge propositions* (see Wittgenstein, 1969), i.e., propositions whose role is to set the framework of "bedrock" assumptions and presuppositions that define what moves are possible within a given practice. But if all that it makes sense to say about our scientific practices is exhausted by what can be described about them, then why accept such axioms in the first place?

Note that returning now to reductivism is not an option here, since it would lead us to the same results all over again: reducing mental-state ascriptions to plain descriptions of behavior or neural patterns, thus depriving them from their normative force, won't do the trick. What both reductive compatibilism and non-reductive incompatibilism have in common is that common is that none retain a viable account of the relation between mind, normativity, and nature. Both seem to draw from the assumption that all it makes sense to say about our human practices -science included- is how they are *in fact* conducted. But if we cannot sensibly speak of how they *should* be conducted… well, then *why* commit to naturalism in the first place? In fact, why even buy the idea that "the only thing we can sensibly say about our human practices is how they are in fact conducted"[33]? Note that our "why" here is

[33] Wittgenstein's own conclusion in the *Tractatus* that "my propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical when he has used them -as steps- to climb up beyond them" and hence that "he must, so to speak, throw away the ladder after he has climbed up it" can be read along these lines (see Wittgenstein, 1921/2001, § 6.54, p. 89).

hungry for *reasons*, not *causes*; we're not, at this point, interested in a naturalist account of our naturalist preferences -such as the one that many radical behaviorists and functional contextualists aim for (see Skinner, 1945; Hayes et al., 2001)- but in a rationalist, irreducibly and ineliminably normative account of the *logic* of such preferences. Insisting in reducing such logic to -or substituting it by- mere descriptions of our "deceitful" brains or the complexities of verbal behavior is to already buy into a particular conception of naturalism; the one whose viability is being put in question.

Some have made the point that a way to escape the self-defeating argument is to adopt a particular understanding of *pragmatism* and the so-called "pragmatic truth-criterion" in relation to science and philosophy, i.e., one according to which "truth" is equated to "successful working", and "whereby an analysis is said to be true or valid insofar as it leads to effective action, or achievement of some goal" (Hayes, 2021, p. 246). "Truth as correspondence with the world", "truth as intra-theoretical coherence", and the like are set aside as working criteria for determining the truth of scientific theories. Functional-analytic approaches (Chapters 1, 2, and 8), and functional contextualists more explicitly, endorse such "pragmatic truth-criterion" (see Barnes-Holmes, 2000; Baum, 2017; Hayes, 2021; Hayes et al. 2012; Moore, 2008; Skinner, 1945, 1953, 1957). These approaches reject strong ontological assumptions (see Barnes-Holmes, 2000); rather, they focus on the practical import and consequences of our theories and explanations, as measured against their own "pre-analytic" goals: namely, prediction and control of behavior. Minds, as folks define them, don't serve well these purposes; thorough analyses of verbal (or relational) behavior fare better.

We agree with these approaches in their pragmatist emphasis on "successful working" in practice as a source of truth for our theories -philosophical, scientific, or otherwise- as well as in their assumption that our pre-analytic, bedrock assumptions are something which cannot themselves be justified, but just acted upon (Wittgenstein, 1969). In fact, we think that these approaches offer a good model of how science, as a social practice at root, should be understood. However, insofar as these approaches are still committed to the idea that such things as "logical validity", "coherence", or "truth" itself can be reduced to or replaced by some thorough analysis of our own verbal behavior as philosophers and scientists (see Barnes-Holmes, 2000; Hayes et al., 2001, 2003), they still express adherence to the same kind of self-defeating naturalism that is at stake here. As we view it, the shift from "truth-as-correspondence" or "truth-as-coherence" to "truth-as-effective-action" analyses of the validity of our scientific and philosophical frameworks just amounts to a shift from one *normative* criterion to another. In either case, what is at stake is one way or another of assessing the *correctness* or *incorrectness* of our social practices -and not just how they're in fact

conducted; in either case, what makes one's analysis "good" or "bad" is given in terms of an irreducibly and ineliminably *logical* (not causal) relation between pre-analytic goals and human practices. What counts as "successful working", "effective action", and the like, is not *given*; rather, it depends on an evaluative framework, which determines what human goals might be valuable, and thus what scientific practices are useful to help us achieve such purposes. As we'll see in Chapter 4 (section 4.2.2.), the "prediction and control" of behavior might be just one among other guiding values that determine the "effectiveness" of different procedures and explanatory schemes.

If our analysis stands, at this point naturalists and non-naturalists seem to stand on an equal footing. None is right nor wrong; after all, a plain description of the neural or behavioral activity patterns of scientists and philosophers of science, no matter how complex, tells us nothing about the correctness or incorrectness of their philosophical and scientific frameworks. Ultimately, if we take reductivism or incompatibilism to the extreme, we cannot even sensibly or genuinely talk about such *conceptual* frameworks; all there is to them would be some primates vocalizing the sounds "ontological naturalism is the way to go", others vocalizing "no, it's not", and yet others vocalizing "All mimsy were the borogoves, and the mome raths outgrabe". In a nutshell, the problem comes down to the following: if, dazzled by the wonders of science, we choose to follow the reductivist or the incompatibilist paths, we end up being unable to answer *why should* we fancy naturalism over non-naturalism in the first place. We thus first need some kind of true compatibilism (i.e., one which can accommodate "Truth-Aptness" and "Normative Force") in order to be able to *justify* why "Nomological Power" is a powerful reason to prefer naturalism over non-naturalism, or an independent, self-correcting science of behavior rather than one constrained by pre-scientific assumptions. In sum, we need some kind of *non-reductive, yet compatibilist* kind of naturalism about the mental.

The problem is that such position is just not available for naturalists operating under the logic of descriptivism. If for mental-state ascriptions to be true-apt they must describe possible states of the world, and the range of possible states of the world is exhausted by what fits the criteria imposed by ontological naturalism, then naturalists must either choose between reductivism -which yields truth-apt, yet normatively inert mental-state ascriptions- or incompatibilism -which rejects their truth-aptness altogether. Under such descriptivist framework, the only viable alternative to reductivism and incompatibilism seems to be the rejection of naturalism itself, i.e., to embrace non-naturalism about the mental. In a nutshell, non-naturalism rejects naturalism and its restrictive notion of description; instead, it assumes that mental-state ascriptions describe private, inner facts which only

oneself has access to. Despite the scientifically off-putting character of such position, let's first consider it before we leave the descriptivist's sinking ship for good.

3.2.3. Non-naturalism, a self-defeating normativism

Securing the truth-aptness and normative force of mental talk at the cost of embracing non-naturalism about the mental doesn't sound like a sensible option from the beginning. Avoiding the problem of normativity by populating our ontology with "nomological danglers" (i.e., causally unbound entities; Smart, 1959) or non-spatial, yet causally-efficacious spooky entities, as Cartesianism does, seems to be a dead-end.

Or is it? To be sure, some might be more than willing to pay that price. Descartes is a case in point here, but other contemporary thinkers have also endorsed similar views. Szasz's rejection of subpersonal accounts of mental health problems and its exclusivist focus on the realm of the personal epitomizes this kind of approach in the field of mental health (see Chapter 2, section 2.4.). Scientifically off-putting, yes; but, to be fair, in the face of naturalists' apparent inability to sustain their own approach, why should we embrace it?

In a sense, non-naturalism seems to be the specular image of incompatibilist naturalism. Both reject reductivism on the grounds that the mental is individuated by a series of non-reducible normative properties. And both assume the incompatibility between folk-psychology and ontological naturalism. But while the latter chooses nature over mind, the former chooses mind over nature. In doing so, non-naturalists reject the "naturalized" notion of description (see section 3.1.3.), liberalizing the range of objects, properties, events and relations that may be picked out by our descriptions of each other's mental states. Instead, non-naturalists like Descartes assume an internalist kind of descriptivism (section 3.1.2.), whereby mental-state ascriptions describe some internal, private state of affairs, the unassailable fortress of individuality. The rationale behind this seems to be something like the exact opposite of the incompatibilist's tragic stance: if the price of letting our naturalist aspirations go is to reinstate meaningfulness and normativity within our worldview, then so be it: let's preserve our precious, glassy essence -to use Rorty's (1979) expression- untouched by the grim claws of nature; when it comes to human affairs, let's forego the enhanced ability to causally explain, predict, and control natural phenomena afforded by the natural sciences.

We might call this the *wonderful stance*. Most of us would feel uncomfortable, to say the least, positing self-creating souls along the way of behavior science. But the problem that we want to stress here is another one: besides the spookiness of the non-naturalist ontological framework, the fundamental problem with this kind of approach is that it doesn't render a satisfactory view of normativity either, nor is thus able to accommodate the "Normative Force" claim; it leads us, as we might say, to some kind of *self-defeating normativism*.

Wittgenstein's (1953/1958) arguments against the possibility of following a rule *privately* offer one of the best diagnoses of this basic problem (see also Kripke, 1982; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014; Thornton, 2007). The argument goes more or less as follows. Cartesians invoke an alternative world of internal, private facts (i.e., those to which only the self has direct access) in order to allow us to distinguish between voluntary, intentional, goal-directed behavior and involuntary, reactive, non-intentional behavior; that is, between normatively evaluable and non-evaluable behavior. In this para-mechanical, intellectualist view of the mind, what makes a behavior normatively evaluable -i.e., an instance of *rule-following behavior*[34] (see Kripke, 1982; Price, 2013; Heras-Escribano & Pinedo-García, 2018; see also Thornton, 2007)- is that it is preceded by "some anterior internal operation of planning what to do" (Ryle 1949/2009, p. 20).

Wittgenstein (1953/1958, § 185-186) -and Kripke (1982), in his interpretation of Wittgenstein- illustrate the problematic nature of this conception of rule-following with the example of someone that is learning to add numbers, i.e., that is learning, as we might say, to follow the rule of adding. Up to now, the person has only learned to add numbers up to 57. Now we ask them to add 68 + 7. One would expect that, if the person has "grasped" or "understood" the rule of adding, then they will give "75" as an answer; but instead, the person gives "5". Shocked, we wonder what might have happened. The person tells us that they thought that "adding" amounted to the following prescription: "for any numbers x and y smaller than 57, add $x + y$ (i.e., add as regular); for any numbers larger than 57, the result is always 5". Now it seems that, in the past, the person was never really *adding*; they were just behaving "as if adding", but they really were following a different, wrong rule, not the right one. This example may seem odd, but the truth is that not many of us has ever added numbers with, say, a million figures. The question is: how do we know that we've been adding correctly (i.e., following the correct rule) so far? Can we be confident that we know how to add correctly?

Note that this question may also apply to any conceivable normatively evaluable behavior, i.e., to anything that we may describe in terms of "following a rule"; from adding numbers to acting "in accordance" with our beliefs, desires, intentions, or even speaking a language. The worry at stake here is the following: what grounds our normative evaluations? How do we know if someone is acting in accordance with some norm or another, or if someone is *correctly* following a rule? The descriptivist's answer to this problem, once again, is

[34] This notion should not be confounded with what behavior analysts call "rule-governed behavior" (i.e., behavior under control of verbal rules; see Chapter 1, section 1.5.2.2.). As we'll see in Chapter 8 (section 8.4.), confounding these two notions is itself a form of intellectualism that might have pernicious clinical consequences.

that there must be some further *fact* that makes our rule-following ascriptions true or false. Here is where non-naturalists invoke the internalist kind of descriptivism: there must be some inner, private fact, that allows us to distinguish between true and false rule-following ascriptions. Knowing how to follow a rule correctly (e.g., knowing how to add numbers, but also knowing what your words mean, or knowing how to act in accordance with your beliefs, desires, intentions, etc.) thus amounts, again, to having formed a certain mental representation; i.e., to grasp or understand a rule is to acquire some inner representation of the rule.

But here's the problem: understanding a rule cannot amount to forming *whatever* representation of the rule; it needs to be a *correct* representation. However, if grasping a rule is to form a *correct* interpretation of it, then how do you distinguish a correct from an incorrect representation of the rule? You'd need to grasp a *second* rule (i.e., one that determines whether your interpretation of the first rule is correct or not); and then a *third* rule (i.e., one that determined whether your interpretation of the interpretation of the first rule is correct or not). In the end, this leads us to an *infinite regress of representations of the rule* (Wittgenstein, 1953; see also Heras-Escribano & Pinedo-García, 2018; Kripke, 1982; Pinedo-García, 2020; Tanney, 2009; Thornton, 2007).

This conception of normativity is self-defeating because it conflates *understanding or following a rule* with *thinking that one is following a rule*. In other words, it leads us to some kind of normative solipsism: if we're ourselves the only ones who have "direct access" to our own representations (and to the representations of the representations of the representations…) of the rules we follow, then we can potentially see *whatever action* we take as an instance of following *whatever rule* we think we're following. No one but us has the authority to sanction our rule-following self-ascriptions as either correct or incorrect. In the end, we end up with no sensible notion of normativity to talk about: if we can't err, we can't be right either; there's no correctness nor incorrectness (Wittgenstein, 1953/1958; see also Heras-Escribano & Pinedo-García, 2018; Kripke, 1982; Pinedo-García, 2020; Tanney, 2009; Thornton, 2007). This leads us to true, ultimate *madness*: from this point of view, nothing "logically follows" from nothing: one might claim to be a naturalist while advocating for the existence of bizarre supranatural entities; or claim to be an exemplar citizen while savagely plundering state coffers; or claim to be a LGBT+ rights supporter and at the same time support LGBT-phobic movements. Wittgenstein thus concludes:

> 201. This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so

> there would be neither accord nor conflict here. (...). What this shews is that there is a way of grasping a rule which is *not* an *interpretation,* but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases. (...).
> 202. And hence also 'obeying a rule' is a practice. And to *think* one is obeying a rule is not to obey a rule. Hence it is not possible to obey a rule 'privately': otherwise thinking one was obeying a rule would be the same thing as obeying it. (Wittgenstein, 1953/1958, p. §§201, 202, p. 81)

These two paragraphs already point in the direction that we'll follow in Chapter 4: that to be in a certain mental state amounts to act in accordance with some rules -i.e., those that determine what courses of action are rationally linked with what mental-state ascriptions- and that to follow such rules is just a matter of being sufficiently trained in certain social practices; it is, as we'll see, a matter of *know how* rather that a matter of social *know that*. However, let's first recap what we've said so far.

3.2.4. Paving the way out of the descriptivist's "fly-bottle"

Thus far, we've seen that both reductive compatibilism and non-reductive incompatibilism have certain attractive features, but also serious conceptual problems. Reductive compatibilism can accommodate "Truth-Aptness" (i.e., the idea that mental-state ascriptions can be contingently true or false), while non-reductive incompatibilism best accommodates "Nomological Power" (i.e., the idea that the explanatory, predictive, and controlling power of the natural sciences should not be encroached by folk-psychological assumptions). However, none of them can accommodate "Normative Force" (i.e., the idea that mental-state ascriptions *rationalize* our doings, render them intelligible), since both erase all trace of normativity from our naturalistic worldview: while the problem of the former lies in its reductivism, the problem of the latter lies in its incompatibilism. Hence neither offers an appropriate kind of naturalism, because both lead us to some kind of self-defeating naturalism, where we're left without any *reasons* whatsoever why we should prefer naturalism over non-naturalism in the first place -among other, more important practical problems.

If those are our only naturalist options, we seem to be caught in an impossible dilemma, whereby we seem forced to choose between two unloving partners: a) sticking to some self-defeating variety of naturalism, which promises us to solve the mind-body problem at the cost of obliterating all talk of truth, rationality, meaning, and so on; or, alternatively, b) embracing non-naturalism, which promises to solve the normativity problem at the cost of populating our ontology with strange creatures. While the former leads us to the normativity problem, the latter reinstates the mind-body problem. In fact, as we've seen, the

dilemma is even grimmer, since non-naturalism is not only off-putting from a scientific point of view; it also renders an untenable view of normativity as well. In a nutshell, when properly analyzed, it leads us to some kind of self-defeating normativism, where anything can be made to be in accordance with any norm (and, consequently, nothing really accords nor conflicts with any norm whatsoever).

We might call this the *puzzle of translatability*, since it seems to arise from the commitment to traducibility, i.e., to the idea that if and only if mental-state ascriptions can be successfully translated to descriptions of spatially and temporally distributed, causally-bound states of affairs, then our folk-psychological interpretative practices are cognitively informative or truth-evaluable and compatible with a defense of ontological naturalism. Under the traducibility constraint, the only tenable way of reconciling the mental within a naturalistic worldview involves the commitment to reductivism, i.e., to some variety of the mind-body identity thesis. Compatibility and reductivism thus stand or fall together -with the tragic upshot that, if reductivism stands, the kind of compatibilism thus achieved can no longer provide what we wanted it to provide: a viable account of the normative force of mental-state ascriptions.

Here is the underlying argument that we've been trying to expose in this chapter, which leads naturalists from the commitment to descriptivism (in its deeper version) to the dead-ends of reductive compatibilism and non-reductive incompatibilism:

Premise 1 (*Descriptivism*): A declarative sentence has cognitive meaning or is truth-evaluable (i.e., it can be declared as true or false) if and only if it describes a state of affairs (i.e., a possible state of the world, or combination of objects, properties, events, or relations among them).

Premise 2 (*Ontological naturalism*): A state of the world is *possible* if and only if it's compatible with the principles of monism, materialism, and the principle of causal closure, i.e., if and only if it's *material* (i.e., a spatial-temporally distributed and causally-bound combination of objects, properties, and relations among them).

Conclusion (*Translatability*): A declarative sentence is truth-evaluable and compatible with a naturalistic worldview if and only if it's *translatable* to a description of material states of affairs.

Drawing from this common argument, reductive compatibilists and non-reductive incompatibilists differ on whether *Translatability* is satisfied or not in the case of mental-state ascriptions.

Reductive compatibilism

Premise 1 (*Translatability*): A declarative sentence is truth-evaluable and compatible with a naturalistic worldview if and only if it's *translatable* to a description of material states of affairs.

Premise 2 (*Reductivism*): Mental-state ascriptions are translatable to descriptions of material states of affairs, either on a context-free or context-relative manner.

Conclusion (*Reductive compatibilism*): Mental-state ascriptions are truth-evaluable and compatible with a naturalistic worldview.

Non-reductive incompatibilism

Premise 1 (*Translatability*): A declarative sentence is truth-evaluable and compatible with a naturalistic worldview if and only if it's *translatable* to a description of material states of affairs.

Premise 2 (*Non-reductivism*): Mental-state ascriptions are not translatable to descriptions of material states of affairs (neither on a context-free nor context-relative manner).

Conclusion (*Non-reductive incompatibilism*): Mental-state ascriptions are necessarily false (hence non-truth-evaluable) and incompatible with a naturalistic worldview.

In sum, if we accept the condition of Translatability, the only way of being a compatibilist about the mental is to embrace some kind of reductivism. On the other hand, if we reject it, then our only available options are to either embrace incompatibilism, which disenchants nature to its self-annihilation, or non-naturalism, which enchants it at the cost of an inflated ontology and a self-defeating solipsism. None, in sum, deal in a satisfactory way with both the mind-body problem and the problem of normativity at once.

However, now that the the underlying argument has been laid out, we can start paving our way out of the "fly-bottle" (Wittgenstein, 1953/1958, §309) that kept us naturalists trapped, i.e., the puzzle of translatability. As we said at the end of section 3.2.2., what naturalists need to provide a unified solution to the mind-body problem and the problem of normativity at once is some kind of *non-reductive, yet compatibilist* account of mind and

normativity: one that is able to accommodate "Truth-Aptness", "Normative Force", and "Nomological Power". To do so, we must first reject *Translatability*, since it's the thesis that ties compatibilism to reductivism. Non-naturalists of the kind we've seen in section 3.2.3. also reject Translatability, but they reject it via the rejection of its second underlying premise: the commitment to *Ontological naturalism.* As naturalists, we cannot reject this premise; hence, our only choice is to reject the first one, the one underlying both Cartesianism and all the naturalist approaches that we saw in Chapter 2: the premise of descriptivism.

3.3. Conclusions

In this chapter, we've seen what are the underlying assumptions of the various approaches to the mental that we saw in Chapter 2. In particular, we've seen how descriptivism yields "the logical mould into which Descartes pressed his theory of the mind" (Ryle, 1949/2009, p. 9), which the main naturalist approaches to the mental retain. We've begun by laying out the mindreading conception of folk psychology that lies at the core of these approaches, according to which folk-psychological interpretation is primarily understood as a pre-scientific attempt to causally explain each other's doings. Different kinds of naturalism disagree on whether folk theories of psychology are correct or not, but they nonetheless retain the assumption that mental-state ascriptions are attempts to describe or represent certain facts about an individual.

We've then distinguished several varieties of this descriptivist commitment. Firstly, we've introduced a rough distinction between two different, yet often entangled descriptivist theses: a) descriptivism at the level of pragmatics, which entails the idea that the only or primary function of language is to describe how the world is or is not; and b) descriptivism at the level of semantics, which entails the idea that, when we're in fact using language to make a statement -i.e., to affirm or deny that something is the case-, the meaning of those statements lies in how they describe or represent some state of affairs (i.e., some possible state of the world or combination of objects, properties, events, or relations among them). Focusing on the latter, we've further distinguished two possible versions of it: a) an affirmative, shallower version of descriptivism, which amounts to saying that declarative sentences always describe or represent certain state of affairs; and b) a conditional, deeper version of descriptivism, which amounts to saying that declarative sentences have meaning proper (in particular, cognitive or truth-evaluable meaning) if and only if they describe or represent some state of affairs. Applied to mental language, these two varieties of descriptivism amount to the claim that mental state-ascriptions in fact describe or represent some state of affairs (e.g., a relation between an individual and a private mental object, some neural or

behavioral pattern of activity, etc.), or that, if they failed to represent such states of affairs, mental-state ascriptions would be contentless. While internalist descriptivists assume that what mental-state ascriptions represent is some relation between an individual and a private internal object, externalist descriptivists assume that what is represented is a public state of affairs (e.g., a person's patterns of actions and reactions).

We've seen how the defining principles of ontological naturalism (i.e., monism, materialism, and the principle of causal closure) can be seen as specific constraints in the range of *possible* states of the world, i.e., those which are spatially and temporally distributed, as well as bound to the laws of nature. In turn, these constraints entail a more restrictive view of what may count as a "successful description" (i.e., one which can be evaluable in terms of its truth or falsity). In the "naturalized" version of descriptivism, which we've traced back to Wittgenstein's (1921/2001) picture theory of language, a declarative sentence is truth-evaluable ("expresses a proposition", "carries cognitive or informational content", etc.) if and only if it represents a spatial-temporally, causally-bound combination of objects, properties, events or relations among them). We've then claimed that this naturalized version of descriptivism lies at the core of the different naturalist approaches to the mental. In particular, we've argued that descriptivism and ontological naturalism, taken together, yield the logical necessity for establishing an identity relation between mind and body, giving rise to the translatability thesis, i.e., the idea that the truth-evaluability of mental-state ascriptions is compatible with ontological naturalism if and only if they are translatable or reducible to descriptions of material states of affairs. Thus, the translatability thesis tights reductivism and compatibilism together, so that the only way in which a naturalist can reconcile folk psychology and science is by assuming the reducibility of mind to matter. This, in turn, leaves naturalists with just two options to account for the place of mind on nature: reductive compatibilism and non-reductive incompatibilism.

After that, we've delved into the main virtues and vices of these two forms of naturalism. On the one hand, the most attractive feature of non-reductive incompatibilism is that it "frees" science from the burden of folk psychology, understood as a vague account of why we behave the way we do. However, it leads to an untenable view of our folk-psychological interpretative practices, whereby these are viewed as fictional at best, senseless at worst. The attractiveness of reductive compatibilism lies in that it allows us to retain the idea that mental-state ascriptions carry truth-evaluable information about an agent.

However, we've argued, neither reductive compatibilism non-reductive incompatibilism provide a nice account of the normative force of mental-state ascriptions. This ultimately leads to self-defeating forms of naturalism, i.e., varieties of naturalism which

themselves undermine their own pre-analytical assumptions; in a nutshell, if for a claim to be truth-evaluable it must describe some material state of affairs, then the core axioms of ontological naturalism are neither truth-evaluable nor rationally justifiable -in fact, once we dispose of the mental or it's normative properties, there's no "rational justification" left to talk about. We've also considered an attempt to avoid this objection, which involves a shift from "truth-as-correspondence" or "truth-as-coherence" criteria to the more pragmatic "truth-as-successful working" criterion in the assessment of the truth or validity of one's conceptual framework. This "pragmatic maneuver", however, cannot avoid the self-defeating objection either, at least as long as it's bound to the idea that normative talk (including talk of "truth", "logic", etc.) can be reduced to or replaced by mere descriptions of material states of affairs (e.g., neural firings, relational responding patterns, etc.).

In the end, unable to provide a proper account of normativity, naturalism thus conceived seems to stand on an equal footing with non-naturalism. Despite its scientifically off-putting character, we've nonetheless considered the possible strengths of non-naturalist accounts of mind like Cartesianism. We've seen that these approaches reject the naturalized version of descriptivism that leads to the translatability assumption. Instead, they typically endorse some kind of internalist descriptivism, whereby the truth value of mental-state ascriptions is given in terms of a relation between the agent and some internal object. The basic problem with this kind of approach is that it leads to an intellectualist construal of rule-following behavior (i.e., "acting in accordance with a certain norm") according to which "grasping" a norm or "following it correctly" (e.g., having a belief or acting upon it) amounts to entertaining a representation or interpretation of the norm "before the eye's mind". This leads to an infinite regress of interpretations of the norm one's following, with the upshot that, in the end, any action can be understood as an instance of following the norm we think we're following. Thus, non-naturalism leads to a self-defeating kind of normativism.

Finally, we've laid out the basic argumentative structure underneath what we've called "the puzzle of translatability". The puzzling aspect of this dilemma resides in that, under the descriptivist's dogma, naturalists seem forced to choose between reductivism and incompatibilism; no real non-reductive incompatibilist approach to the mental is available for them. The non-naturalist alternative, on the other hand, not only leads to the mind-body problem, but also, as we've seen, to an untenable account of normativity. After laying out the line of thought that goes from descriptivism to the two dead-ends of reductive compatibilism and non-reductive incompatibilism, we've finally pointed out where the solution lies: if we must reject the translatability assumption while remaining committed to ontological naturalism, we must reject the descriptivist commitment.

In Chapter 4, we'll expose our *non-descriptivist* approach to the meaning and function of mental-state ascriptions. This will allow us to reconcile two ideas: a) that mental-state ascriptions do not describe any spatial-temporally, causally-bounded state of affairs -nor any state of affairs whatsoever; and b) that they can nonetheless be assessed in terms of their truth or falsity. In doing so, our non-descriptivism will allow us to meet the challenge that we exposed at the end of Chapter 2: that of providing a conceptual framework for mental health which yields a viable account of the relation between nature, on the one hand, and mind and normativity, on the other. Then, in Chapters 5 to 8, we'll see some of the conceptual and empirical implications of this non-descriptivist framework to for the intervention with people with delusions.

Chapter 4
Non-descriptivism and the post-ontological account of mind

In the previous chapters, we've seen how the problems of mind and normativity are deeply inter-related and entrenched in the various theories of mind that underlie the different therapeutic models. We've also seen how descriptivism, the implicit semantic commitment at the heart of the Cartesian theory of mind, is at the root of such problems, and how it has shaped and restrained the range of possible solutions. Specifically, we've seen how descriptivism leads us to what we called "the puzzle of translatability", whereby we're seemingly forced to either choose between some kind of self-defeating naturalism (i.e., one that may allow us to avoid the mind-body problem at the cost of erasing normativity from our worldview) or some kind of non-naturalist normativism, which populates our ontology with strange creatures -and in vain, for it doesn't even succeed at providing a proper account of normativity. We ended up the previous chapter with a challenge: to find a *non-reductive, yet compatibilist* kind of naturalism to account for the place of mind on nature; in other words, one which enables us to overcome the puzzles of descriptivism, avoiding both the mind-body problem and the problem of normativity at once. The main goal of the present chapter will be to discuss a possible way out of such dilemma, via the rejection of descriptivism and the adoption of a non-descriptivist approach to the meaning and function of mental-state ascriptions.

The structure of the chapter will be as follows. In section 4.1., we'll first draw a rough map of the different kinds of non-descriptivism at our disposal, in order to establish the specific commitments that we'll endorse, and which we'll apply to the analysis of the meaning and function of mental-state ascriptions. As we'll see, the kind of non-descriptivism that we'll advocate for is rooted in Wittgenstein's meaning-as-use conception of language, which can be described as a pragmatist kind of non-descriptivism. In section 4.2., we'll apply this

pragmatist non-descriptivist framework to the analysis of folk-psychological interpretation. Drawing from both Wittgenstein's and Ryle's view of mind and language, we'll first offer some further reasons why mental-state ascriptions are best viewed as moves within evaluative and regulative, rather than descriptive language-games. As we'll see, their work also offers a way to understand the truth-aptness and normative force of mental-state ascriptions that is not in tension with our defense of ontological naturalism, offering instead a *post-ontological* account of the place of mind on nature (see Ramberg, 2000; see also Pinedo-García, 2020). This approach affords the kind of naturalist, non-reductivist, yet compatibilist approach to the mind that we were left aiming for at the end of Chapter 3, i.e., one that preserves the truth-aptness and normative force of mental-state ascriptions while avoiding both the commitment to non-naturalism about the mind as well as to reductivism or incompatibilism; in other words: one that avoids both the mind-body problem and the problem of normativity. Finally, in section 4.3., we'll summarize the main conclusions from this and the previous chapters.

Before we move on any further, however, we owe an apology -and a promise of redress. We've already gone far astray from the field of mental health and, for a while, we'll continue to do so here. Nonetheless, we'll come back to it in the following chapters, where we'll discuss some interesting implications of these philosophical discussions to mental health research and practice.

4.1. Non-descriptivism

As we saw in Chapter 3, the descriptivist assumption that lies at the core of both non-naturalism and classical forms of naturalism leads us to the puzzle of translatability, whereby we seem forced to choose between two possible kinds of self-defeating naturalism (reductive compatibilism or non-reductive incompatibilism), or rather embrace non-naturalism, which itself leads to a self-defeating kind of normativism. The key to find a way out of the dilemma was to reject the *translatability* assumption -which ties reductivism and compatibilism together- and, specifically, the commitment to descriptivism that gives rise to it (see the argument in Chapter 3, section 3.2.4.). In other words: if we are to find some non-reductive, yet compatibilist variety of naturalism (i.e., one that is able to accommodate "Truth-Aptness", "Normative Force", and "Nomological Power"; see Chapter 3, section 3.2.1.), then we must endorse some kind of *non-descriptivist* approach to the meaning of mental-state ascriptions.

In contemporary philosophy of language, two of the most important non-descriptivist approaches are expressivism (Ayer, 1936; Bar-On, 2015; Bar-On & Sias, 2013; Blackburn,

2006; Chrisman, 2007; Field, 2009; Frápolli & Villanueva, 2012; Gibbard, 1986; Pérez-Navarro et al., 2019; Price, 2011; Price et al., 2013) and inferentialism (Brandom, 2000), although the latter is typically understood as a variety of the former, e.g., as a global kind of expressivism (Price 2011; Price et al., 2013) (see also Almagro-Holgado, 2021; Frápolli & Villanueva, 2013; Villanueva, 2018). These approaches and their many varieties differ in many important respects (see Bar-On & Sias, 2013; Frapolli & Villanueva, 2012, 2013; Villanueva, 2018) and, although most of the arguments employed here have been best articulated by expressivist or inferentialist thinkers, a full-fledged development of these approaches is well beyond the scope of the aims of this chapter (as well as beyond the author's expertise).

Thus, we've chosen to speak of "non-descriptivism" in general and to try to map out some general lines of divergence among non-descriptivist approaches. This will be useful to better appreciate the core commitments of the particular kind of non-descriptivist approach to the mind that we'll endorse here. As we'll see below, different non-descriptivist approaches can be differentiated in terms of three aspects: a) their core *negative* commitments (i.e., whether they reject descriptivism at the level of pragmatics, semantics, or both); b) their *positive* commitments (i.e., what alternative conception of the meaning of the analyzed sentences they endorse); and c) their *global* or *local*[35] character (i.e., whether their analysis is taken to apply to all kinds of claims or only to those containing certain expressions, e.g., logical, epistemic, doxastic, moral, etc.).

4.1.1. Mapping non-descriptivism

Non-descriptivism can be understood as a family of approaches to the philosophy of language that share a common *negative* commitment: the rejection of descriptivism. Now, as we said in Chapter 3 (section 3.1.2.), we can roughly distinguish between two possible levels of analysis at which descriptivism may be implemented: a) descriptivism at the level of *pragmatics* (i.e., roughly, at the level of what we *do* with words), and descriptivism about *semantics* (i.e., roughly, at the level of what we *say* with words, the truth-evaluable information that we communicate by means of words). We've identified the former with Austin's (1962) *descriptive fallacy*, understood as the idea that the only or primary function of language is to describe the world, to state how things are, and thus that the meaning of any utterance is always to be determined in terms of its truth-conditions (i.e., in terms of what would make such sentence true or false). By contrast, we've identified descriptivism at the level of semantics with the *dogma of descriptivism* (see Chrisman, 2007), which may be understood in two possible

[35] The distinction between "global" and "local" approaches is typically draws between different varieties of expressivism (see Price, 2011; Price et al., 2013).

ways: a) as an affirmative statement, i.e., that the meaning of declarative sentences is in fact constituted by a description or representation of some possible state of affairs; or b) as a conditional statement, i.e., that only if a sentence successfully describes or represents some possible combination of objects, properties, events or relations among them, then it has content or cognitive meaning (or, in other words, it is "truth-apt" i.e., can be evaluated in terms of its truth or falsity). The former amounts to what we've called the "shallow" version of descriptivism, while the latter amounts to what we've called its "deep" version.

On the other hand, different non-descriptivist approaches can be distinguished in terms of the *positive* theses they endorse regarding the meaning of whatever set of utterances of interest, i.e., in terms of the proposal that they articulate as an alternative to descriptivism (see Frápolli & Villanueva, 2012, 2013; Price et al., 2013; Villanueva, 2018). At the level of pragmatics, non-descriptivist approaches typically draw from a shared commitment to some variety of *functional pluralism*, i.e., the idea that language serves myriad practical purposes (e.g., Austin, 1962; Price, 2011); making statements about the world might be one such possible use, sure, but also making promises, issuing orders, issuing declarations, etc. Critically, many of these uses of language don't primarily involve the assertion of something that can be true or false. When we utter things like "please, eat something other than a ham and cheese sandwich, grim shadow of what used to be a person" or "stop being so stupidly cute, you little cat!", we're not primarily saying something whose truth or falsity is being endorsed or denied; rather, we're issuing commands or expressing our affective states. However, according to functional pluralism, this doesn't mean that such utterances lack meaning; rather, their meaning must be spelled out in terms of the effects, broadly construed, that these utterances have upon the speaker and the audience.

Note, however, that one may admit the plurality of uses of language, and yet remain committed to descriptivism at the level of semantics, i.e., to the idea that, when we're in fact using language to make a statement, the meaning of our utterance can (or must) always be given in terms of a possible combination of objects, properties, events, or relations that our utterance represents or "stands for". Instead, non-descriptivist approaches at the level of semantics draw from a shared commitment to the idea that at least some declarative sentences (typically, those containing ethical terms, mental terms, logical terms, etc.), when used in at least some contexts, do not represent any particular state of affairs (see Bar-On & Sias, 2013; Frápolli & Villanueva, 2012, 2013). What these approaches thus have in common is that they deny the affirmative version of the descriptivist claim, i.e., the assumption that all declarative sentences describe or represent some possible state of the world. For Frápolli &

Villanueva (2012, p. 471), this is the characteristic commitment of non-descriptivist approaches (see their description of the commitment to "Non-descriptivism").

By contrast, different non-descriptivist approaches differ as to what exact kind of meaning these declarative sentences have -if any at all. Some retain the commitment to the deep version of descriptivism, i.e., that either a sentence describes or represents some state of affairs or it has no content or cognitive meaning (i.e., it is not truth-apt). Approaches that reject the affirmative, but not the conditional version of descriptivism are typically committed to what Frápolli & Villanueva (2012) characterize as the "Truth-Conditional Status" commitment, which amounts to the idea that those declarative sentences that contain non-descriptive expressions "lack truth conditions, even though they are syntactically correct – they are not 'truth-apt'" (p. 471). Classical expressivisms, like Ayer's (1936; see also Frápolli & Villanueva, 2012, 2013) emotivism, constitute an example of this kind of approach. Framed by a factualist view of the mind, classical expressivists understood the difference between claims with cognitive content (i.e., truth-apt claims) and those without it in terms of the kind of *inner* states these claims express; on this view, truth-apt claims express cognitive states (e.g., beliefs, mental representations of the world), while other claims express conative or affective states (e.g., feelings of approval or disapproval towards certain states of affairs). In a nutshell, this kind of non-descriptivism assumes that non-descriptive sentences, at best, might have some *expressive* or *prescriptive* meaning, but no cognitive (i.e., truth-evaluable) meaning proper (see Frápolli & Villanueva, 2012). On this view, expressions such as "slavery is wrong" would be as (non-)truth-evaluable as "please, eat something other than a sandwich".

Instead, the kind of non-descriptivist approaches that we're interested in here draw from a more profound rejection of descriptivism at the semantic level and assume that declarative sentences containing non-descriptive expressions can be as truth-apt as "pure" descriptions of some state of affairs; that is, they reject what we've called the "conditional version" of descriptivism[36]. Some of these approaches endorse some kind of *pluralism about truth*, which roughly amounts to the idea that the truth-evaluability of different declarative sentences does not depend -or at least does not *always* depend- on their representing some possible state of the world, but on a variety of possible factors (see sections 4.2.3. and 4.2.4.) (Price 2011; Price et al., 2013; see also Pedersen & Wright, 2018).

Finally, non-descriptivist approaches also differ as to whether they adopt a *global* or *local* kind of non-descriptivism, i.e., whether they advance these negative and positive theses

[36] For instance, as we understand them, Frápolli & Villanueva's (2012) minimal expressivism would constitute such an approach, as well as Price's (2013) global expressivism or Brandom's (2000) inferentialism.

with regard to the meaning of *all* kinds of utterances or just to those that include some particular subset of expressions (e.g., logical, epistemic, psychological, ethical, etc.). Local non-descriptivists might thus hold a non-descriptivist approach towards some kind of sentences (say, ethical claims) but not others (e.g., mental-state ascriptions), while global non-descriptivists typically reject descriptivism as a general theory of meaning (see Price, 2011; Price et al., 2013; see also Almagro-Holgado, 2021; Brandom, 2000; Frápolli & Villanueva, 2012, 2013; Pérez-Navarro et al., 2019; Villanueva, 2018).

Now that we've sketched out a rough map of the different possible kinds of non-descriptivism, we're in a better position to understand the core commitments of our own approach.

To begin with, our approach assumes non-descriptivism at *both* the pragmatic and the semantic level: neither what we do with language nor the truth-evaluability of what we say with it is necessarily exhausted by assertion nor description. In fact, our approach assumes that both levels of analysis are deeply inter-related, and cannot be neatly divided: roughly, it's a *pragmatist* kind of non-descriptivism (e.g., Price, 2011; Price et al., 2011), since it assumes the primacy of *use* over *content*; it's the use of our declarative sentences in particular contexts, the moves that they constitute when they're used in particular social practices, what determines what criteria must be considered when assessing their truth or falsity -that's why this approach has sometimes been referred to as the *meaning-as-use* conception of language (Wittgenstein, 1953/1958; Price, 2011; Price et al., 2013; see also Almagro-Holgado, 2021; Frápolli & Villanueva, 2012, 2013; Heras-Escribano et al., 2015; Heras-Escribano & Pinedo-García, 2018; Pérez-Navarro et al., 2019; Pinedo-García, 2014, 2020; Villanueva, 2019). In a nutshell, the defining negative and positive theses of this approach, to be developed in the following sections, are the following:

Negative theses:

a. Language is not an exclusively descriptive tool.

b. The truth-evaluability of at least certain declarative sentences is not necessarily linked to their representational capacity, however this capacity might be construed.

Positive theses:

a. Language is a multi-functional tool.

b. The truth-evaluability of a declarative sentence depends on *what logically follows* from it (and what it follows from) when it's used in particular contexts for making such statement, i.e., the logical justifiability relations that are established between

> what is said and other propositions and courses of action; these justifiability relations, in turn, depend on the particular social practices that characterize a certain community.

Finally, although we're sympathetic to global non-descriptivist approaches, here we'll basically focus on the analysis of the meaning of mental-state ascriptions. Specifically, we'll focus on the analysis of ascriptions of propositional attitudes -and mainly belief ascriptions, due to their relevance for the upcoming chapters- although some of the consequences of this analysis also apply to the case of ascriptions of "occurrent" states (e.g., pain, inner speech, mental imagery, etc.) (see Chapter 2, section 2.1.).

The kind of non-descriptivist approach to mental vocabulary that we'll advocate for mainly draws from a pragmatist reading of Wittgenstein's (1953/1958; see also Wittgenstein, 1969, 1974, 1980a, 1980b, 1982, 1992) and Ryle's (1949/2009) work. Similar arguments to the ones endorsed here can also be found in the work of other early analytic philosophers (Sellars, 1956, 1963/1999) as well as contemporary thinkers working within a pragmatist, post-Rortyan criticism of representationalism about language (e.g., Brandom, 2000; Price, 2011; Price et al., 2013). In particular, our reading of Wittgenstein's and Ryle's work is itself primarily informed by the work of analytic philosophers working in or associated to the University of Granada (e.g., Acero & Villanueva, 2012; Almagro-Holgado, 2021; Fernández-Castro, 2017a; Frápolli & Villanueva, 2012, 2013; Heras-Escribano et al., 2015; Heras-Escribano & Pinedo-García, 2018; Pérez-Navarro et al., 2019; Pinedo-García, 2014, 2020; Villanueva, 2014, 2018, 2019)[37]. In recent years, these authors have drawn from similar varieties of non-descriptivism to address several conceptual and practical issues related to the use of mental language, such as: a) the problem of intentionality (Acero & Villanueva, 2012; Villanueva, 2019); b) the analysis of knowledge and self-knowledge ascriptions (Pérez-Navarro et al., 2019; Villanueva, 2014); c) the irreducible and ineliminable character of normative explanations of behavior in the cognitive sciences (Heras-Escribano et al., 2015; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2020); d) the regulative function of folk psychology (Almagro-Holgado & Fernández-Castro, 2019; Fernández-Castro, 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019); e) the conceptual analysis of key notions in cognitive science, e.g., affordance (Heras-Escribano, 2019; see also Almagro-Holgado, 2020); and f) the analysis of political phenomena such as epistemic injustice or political polarization and the

[37] Other colleagues not mentioned here but whose views have significantly informed the arguments exposed in this and the previous chapter include Daniel Galdeano Manzano, Amalia Haro Marchal, Alba Moreno Zurita, Llanos Navarro Laespada, Javier Osorio Mancilla, or José Ramón Torices Vidal, among others.

different possible strategies to measure and intervene on them (e.g., Almagro-Holgado, 2021; Almagro-Holgado et al., 2021; Almagro-Holgado & Moreno-Zurita, 2022; Frápolli & Navarro-Laespada, 2021; see also Bordonaba et al., 2022).

In the following section, we'll delve into the core characteristics of the Wittgensteinian (and Rylean) conception of language. Later on, in section 4.2., we'll see how this pragmatist kind of non-descriptivism, when applied to the analysis of the meaning of mental-state ascriptions, provides us with a way out of the puzzle of translatability and addresses both the mind-body problem and the problem of normativity.

4.1.2. Meaning-as-use. A Wittgensteinian, pragmatist kind of non-descriptivism

The Wittgensteinian "meaning-as-use" conception of language that we'll favor here can be roughly defined by the following assumption: that the meaning of a certain expression hinges on the norms that constrain its possible *uses* in different *language-games*. The concept of "game" here illuminates well what the meaning-as-use conception of language amounts to. It is deployed by Wittgenstein (1953/1958) to highlight three important features of his view of language: a) that language must be understood as a normative system of logical relations among concepts, whereby the meaning of any linguistic element is given by the logical or inferential relations that it keeps with other linguistic elements; b) that such normative system is not contained in some abstract, internal ability with which we're magically -or genetically- endowed at birth, but grounded in a loose set of norm-governed, situated, and radically social communicative practices *to which we are born*; and c) that there needn't be any particular necessary nor sufficient condition that is common to all our possible communicative practices (see also Acero, 2019; Almagro-Holgado, 2021; Brandom, 2000; Frápolli & Villanueva, 2013; Heras-Escribano et al., 2015; Heras-Escribano & Pinedo-García, 2018; Kripke, 1982; Pinedo-García, 2014, 2020; Price et al., 2013; Rorty, 1979; Villanueva, 2019).

These three features are well captured by other key Wittgensteinian notions. The first one is the Wittgensteinian distinction between the *surface grammar* and the *depth grammar* of an expression (Wittgenstein, 1953/1958, §664, p. 168). Roughly, the former refers to the syntactic structure of the expression, i.e., what kind of nouns, adjectives, verbs, etc. it has, whether it's a simple or complex sentence, and so on. By contrast, the depth grammar of an expression -what Ryle (1949/2009) called its *logical geography*- refers to the *logical, inferential or justificatory connections* that can be established between it and other possible expressions and courses of action, i.e., to the set of things that can justify having used it and the set of things that can be justified by using it. For Wittgenstein (1953/1958) and Ryle (1949/2009), the depth grammar or logical geography of a given expression is what gives it its meaning (and its truth-conditions, when it has them).

Already in the *Tractatus,* Wittgenstein (1921/2001) claimed that "*the limits of my language* mean the limits of my world" (§5.6, p. 68). The key idea here is that the criteria to assess the truth or falsity of any given claim are not out there, *given* in the world, external to our conceptual system, but rather depend on language itself, on an already available system of logical relations among possible states of the world; *language is what is given*, what determines what is thinkable or conceivable, and thus what is evaluable in terms of its truth or falsity. This antedates Sellars' (1956) attack on what he called the *myth of the given*, or the idea that there are "brute" or raw facts, i.e., facts that are not conceptually-articulated via language (e.g., sense data, like sounds or patches of color), to which we may have some immediate and incorrigible access and which have a foundational role in our epistemic practices (i.e., which are the ultimate tribunal upon which to determine the truth or falsity of our claims about the world). Both Wittgenstein and Sellars rejected this foundationalist view of knowledge; there are no such brute facts, no final tribunal where truth ultimately resides (see Bensusan & Pinedo, 2007; Pinedo, 2014). By contrast, the truth or falsity, and, relatedly, the meaning of a claim, is determined by its depth grammar or logical geography, i.e., by what it may justify and what may justify it (Frápolli & Villanueva, 2012; see also Brandom, 2000).

The later Wittgenstein did not abandon this idea; instead, he situated language -and thus the concepts of meaning, truth, falsity, etc.- in the context of the different human habits and customs. Here we arrive to the second characteristic feature of the kind of Wittgensteinian non-descriptivism that we're advocating for, and which gives its entitlement to "pragmatism", i.e., the idea that the norm-governed character of language is grounded on our social, communicative practices[38]. This second feature is best captured by the Wittgensteinian notion of a *form of life.* With this notion, Wittgenstein stresses the radically social character of language and the different norms and criteria that competent language speakers follow when determining the meaning and truth value of different claims. The arguments that he uses for establishing this radically social view of language are the ones that led us in Chapter 3 (section 3.2.3.) to reject non-naturalism *à la* Descartes and the kind of self-defeating normativism to which it leads. The problem with this approach was that it allowed for the possibility of following a norm "privately", i.e., to follow a rule that only oneself "grasps" and which only oneself knows when is being correctly applied or followed. The idea of "private rule-following" thus leads to a paradox, which Wittgenstein (1953/1958) explains as follows:

[38] For a full-fledged development of the deep connections between this late Wittgensteinian view of language and the work of classical and contemporary pragmatist thinkers, see Bernstein (2010).

> 201. This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here. (...). (Wittgenstein, 1953/1958, §201, p. 81)

Ryle's distinction between *knowledge-that* and *knowledge-how* is relevant here. For Ryle, "following a rule" or acting in accordance with a norm, is not a matter of theoretical knowledge or *know-that*, i.e., of representing "in one's head" a maxim or regulative proposition, and then acting accordingly (again, this would lead us straight to the problem of private rule-following); instead, it's a matter of practical knowledge or *know-how*, i.e., a matter of practical skill and habit. For the author, the conflation of knowledge-how with knowledge-that is just another consequence of the official doctrine and its commitment to the "intellectualist legend" (see Chapter 2, section 2.1.2.).

> Champions of this legend are apt to try to reassimilate knowing *how* to knowing *that* by arguing that intelligent performance involves the observance of rules, or the application of criteria. It follows that the operation which is characterised as intelligent must be preceded by an intellectual acknowledgment of these rules or criteria; that is, the agent must first go through the internal process of avowing to himself certain propositions about what is to be done ('maxims', 'imperatives' or 'regulative propositions' as they are sometimes called); only the can he execute his performance in accordance with those dictates [...] To do something thinking what one is doing is, according to this legend, always to do two things; namely, to consider certain appropriate propositions, or prescriptions, and to put into practice what these propositions or prescriptions enjoin. It is to do a bit of theory and then to do a bit of practice. (Ryle, 1949/2009, p. 18)

The absurdity of the intellectualist legend is revealed when we realize that "planning what to do" is something that can also be done quite stupidly (e.g., a PhD student sketching the structure for their dissertation and then realizing, while writing, that it doesn't make sense), or that verbalizing something to oneself (either covertly or overtly) before doing it may be as much an exhibition of intelligence or knowledge as an exhibition of foolishness or ignorance (e.g., a person who is learning how to speak another language and needs to constantly translate each sentence in the foreign language to a sentence in their own language). In fact, as Ryle (1949/2009) points out, that "following a rule" or knowing how to act in accordance with a rule is not assimilable to "contemplating or verbalizing a regulative

proposition" is shown by the fact that many of our most ordinary practices can be normatively evaluated, and yet we cannot even spell out what are the norms we're following. For example, when we enter an elevator, we can't spell out the regulative proposition that tells the exact minimum distance that we have to leave between ourselves and others -we don't have any possible "know that"-; however, it would be unwise to conclude from this that "we don't really know" that standing at barely two centimeters from some unknown folk when the rest of the elevator is free is clearly wrong (Heras-Escribano, 2019).

According to Wittgenstein (1953/1958), what these examples and the aforementioned paradox show, "is that there is a way of grasping a rule which is *not* an *interpretation,* but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases (§201, p. 81). He continues:

> 202. And hence also 'obeying a rule' is a practice. And to *think* one is obeying a rule is not to obey a rule. Hence it is not possible to obey a rule 'privately': otherwise thinking one was obeying a rule would be the same thing as obeying it. (Wittgenstein, 1953/1958, §202, p. 81)

Like the early Wittgenstein, the later Wittgenstein assumes that what is evaluable in terms of truth or falsity, and the criteria that we use to establish the meaning of different claims, depends on the rules of language. However, language is no longer conceived as some ungrounded, close, and abstract conceptual system that is set once and for all, about which one can effortfully think in order to spell out its ultimate and fundamental rules; instead, he rather conceives it as a loose set of myriad communicative practices in which we're *trained* by our linguistic community -the community to which we are born. Using and understanding language correctly amount to following certain rules, but "obeying" a rule is a matter of practice, not theoretical reflection. In other words: knowing the rules that determine the meaning and truth-value of different claims amounts to being a competent participant in different social practices, i.e., to knowing *how* to participate in such practices, and not (necessarily) to knowing *that* these or that are the norms that govern it. And it is through a constant social training, via social sanctions and corrections, that we progressively become competent participants in the many language-games (and the social practices of which they form part) that characterize the form of life of our community (see also Almagro-Holgado, 2021; Bernstein, 2010; Heras-Escribano et al., 2015; Heras-Escribano & Pinedo-García, 2018; Kripke, 1982; Pinedo-García, 2020; Price, 2011; Price et al., 2013). As Wittgenstein puts it at the beginning of the *Investigations*:

> 19. It is easy to imagine a language consisting only of orders and reports in battle. –Or a language consisting only of questions and expressions for answering yes and no. And innumerable others–. And *to imagine a language means to imagine a form of life* [emphasis added]. (Wittgenstein, 1953/1958, §19, p. 8)

> 23. But how many kinds of sentence are there? Say assertion, question, and command? –There are *countless* kinds: countless different kinds of use of what we call "symbols", "words", "sentences". And this multiplicity is not something fixed, given once for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten (...).
> Here the term *"language-game"* is meant to bring into prominence the fact that the *speaking* of language is part of an activity, or of a form of life. (Wittgenstein, 1953/1958, §23, p. 11)

Wittgenstein's remarks on the countless multiplicity of language-games already point to the third characteristic feature of the meaning-as-use conception of language, i.e., that there need not be any necessary or sufficient condition that is common to all our possible communicative practices. This idea is best captured by the Wittgensteinian notion of *family resemblance*, used by Wittgenstein to highlight the absence of some commonality that may be viewed as essential to every possible communicative practice. The category of "game" applies to many different sorts of things; chess, rugby, hide and seek, solitaire, athletics, rap battles, etc. If we straightforwardly *look* at -rather than *think* about- this games, as Wittgenstein (1953/1958, §66, p. 33) recommends, we'll see that there's no underlying "essence" nor "general form" that is "picked out" by the "game" category. What we see is "a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail" (Wittgenstein, 1953/1958, §66, p. 33). Wittgenstein's point is that the same holds for language-games. On the one hand, there are multiple actions that we may carry out by means of language, (e.g., asking, commanding, declaring, etc.), with making assertions about the world just being one of them -this is what we referred to above as "functional pluralism". But even when we affirm or deny that something is the case, we needn't be always playing the same language-game; in this sense, Wittgenstein's pluralism also concerns the truth-conditions of the different declarative sentences that we make use of in different communicative interactions. So, in a nutshell, in different language-games, the norms that determine the permissibility and correctness of certain moves may be different. What makes all these possible interchanges to fall under the common category of

"language" is not some essential, core feature, but their family resemblances[39]. Wittgenstein (1953/1958, §65, p. 30) states this as follows:

> 65. Here we come up against the great question that lies behind all these considerations –For someone might object against me: "You take the easy way out! You talk about all sorts of language-games, but have nowhere said what the essence of a language-game, and hence of language, is: what is common to all these activities, and what makes them into language or parts of language. So you let yourself off the very part of the investigation that once gave you yourself most headache, the part about the *general form of propositions* and of language."
> And this is true-. Instead of producing something common to all that we call language, I am saying that these phenomena have no one thing in common which makes us use the same word for all,- but that they are *related* to one another in many different ways. And it is because of this relationship, or these relationships, that we call them all "language". (...). Wittgenstein (1953/1958, §65, p. 31)

Other than Wittgenstein's endorsement of the multi-functionality of language, this paragraph also comprises one of the most important conceptual shifts in his thought, which partially explains why scholars normally talk of a "first" and a "second" Wittgenstein. This is a shift in his view of the proper *method* of philosophy, which underlies the aforementioned overall pragmatic turn in his view of language. The search for the "general form of propositions and of language" was the task that the early Wittgenstein set himself in the *Tractatus*; his conclusion was, as we saw in Chapter 3 (section 3.1.3.), that the only things that can be *said* -i.e., the only things that are actually truth-apt- are empirical propositions, which describe possible states of affairs. He sums up the result of such endeavor in the preface to the book, where he states his famous maxim: "what can be said at all can be said clearly, and what we cannot talk about we must pass over in silence" (Wittgenstein, 1921/2001, p. 3). The early Wittgenstein thus conceived the proper method of philosophy as *prescriptive*: to think about the ultimate nature of language and logic, so we may establish what can be said once and for all. By contrast, the later Wittgenstein is characterized by a radical change in what he views as the proper method of philosophy: it is to describe, rather than prescribe. On this

[39] Some may note an apparent contradiction here: on the one hand, we've said that the meaning of an expression always depends on the system of inferential connections between such expression and others, as well as certain courses of action; on the other hand, we've said that there's no common feature to every "language-game". However, we think that this is not a contradiction whatsoever; as Price (2011) puts it, admitting "that various of the different language games all avail themselves of the same inferential machinery [...] is thoroughly compatible with underlying pluralism, so long as we also maintain that the various different kinds of commitments answer to different needs and purposes" (p. 310).

view, philosophical analysis fundamentally errs when it comes to the conclusion that competent language speakers must somehow be "wrong" in their use and understanding of language (i.e., when it pretends to establish some kind of *error theory* about the normal use of certain expressions, e.g., ethical, epistemic, doxastic, logical, etc.). Instead, philosophical analysis must start from what is already given: our forms of life, our different linguistic practices. We must describe these various practices, and try to establish what norms speakers *in fact* follow when they assess the truth or falsity of different claims, instead of trying to impose some restrictive, preconceived golden rule that may apply to any possible language-game. This overtly pragmatist methodological shift is summarized in another famous Wittgensteinian maxim: "Don't think, but look!" (Wittgenstein, 1953/1958, §66, p. 30). Specifically, what we're supposed to look at is at the myriad language-games that we ordinarily engage in, and what they tell us about the logical geography or depth grammar of the expressions that we use in such games (see Price, 2011; see also Price et al., 2013).

Thus far, we've sketched out the main characteristics of the Wittgensteinian, pragmatist kind of non-descriptivism that we're favoring here. Roughly, it characterizes the meaning of an expression as dependent on a network of inferential connections between such expression and others, as well as with the possible courses of action that are logically connected to its use. In addition, it recognizes the multiplicity of things that may merit the name "language", and grounds these diverse language-games in the common social practices that characterize a given form of life. Contrary to the early Wittgenstein, the later Wittgenstein doesn't idolize any particular norm or model of language as the benchmark against which to compare its many possible uses; hence descriptivism, the dogmatic prescription that the only truth-evaluable expressions are those that represent some potential state of affairs, is dismissed here as an unfounded philosophical preconception, a mere fruit of philosophical overthinking -and one that carries with it the most terrible ghosts and puzzles. Instead, it's replaced by an explicit recognition of the vast plurality of communicative practices that we engage in, as well as the plurality of criteria that we might use in different social interchanges to assess the meaning (and truth value, when relevant) of different uses of language. Rather than prescribing what can and cannot count as "meaningful" or "truth-evaluable", we're invited to take a look at our actual communicative practices. In the following section, we'll see how Wittgenstein -as well as Ryle- applied this methodological advice to the analysis of the particular kind of interpersonal practice that we're interested in here; i.e., folk-psychological interpretation.

4.2. Pragmatist non-descriptivism and folk-psychological interpretation

So, what do we see when we take a close look to our folk-psychological interpretative practices? Recall the various examples that we saw in Chapter 3 (sections 3.2.1. and 3.2.2.). What's the characteristic function of the mental-state ascriptions that the different characters use in the diverse *mind games* in which they engage? Is it a descriptive, causal-explanatory function? If not, what is it? In the following sections, we'll develop the consequences of the Wittgensteinian meaning-as-use conception of language for the analysis of the meaning and function of mental-state ascriptions. First, we'll see some further arguments against the idea that mental-state ascriptions are descriptive or representational linguistic devices. After that, we'll sketch out an alternative conception of folk psychology, according to which the main function of mental-state ascriptions is not to describe an agent's doings nor their overt or covert causes, but to evaluate such doings in normative terms, as norm-conforming or norm-divergent. Finally, we'll come back to the issue of truth-aptness, and we'll see how the pragmatist kind of non-descriptivism advocated for here leads us to a post-ontological account of the place of mind in nature, which affords the kind of non-reductive, yet compatibilist kind of naturalism about the mind that we need to escape the puzzle of translatability.

4.2.1. Mental-state ascriptions as non-descriptive devices

There are several reasons why we can reject that the primary function of folk-psychological interpretation is to describe some given state of affairs. We saw one such argument already in Chapter 3 (see section 3.2.2.), related to the recognition of the normative or prescriptive force of mental-state ascriptions. As we saw, this normative force is lost when we replace mental-state ascriptions by some description of some particular state of affairs, and attempting to do so constitutes an example of the is-ought fallacy. Since such normative force seems to be a core characteristic of mental-state ascriptions, these don't seem to be descriptive.

We'll come back to this characteristic feature below, since it points to a better conception of the primary function of our folk-psychological interpretative practices. Before that, let's first see some more indicators that suggest that our folk psychology is not some kind of proto-scientific, descriptive practice. According to Villanueva (2019) Wittgenstein's work provides at least two other observations that point to the non-descriptive character of mental-state ascriptions -and, particularly, of ascriptions of propositional attitudes like beliefs, desires, intentions, and so on. We'll refer to these two observations as *non-durability* and *truth-conditional dependence*.

To begin with, we might question: is folk-psychological interpretation a unitary linguistic practice? Or are there different possible uses of what commonly falls under the broad category of "mental states and processes"? In this regard, both Wittgenstein (1953/1958; see also Wittgenstein, 1980a, 1980b, 1982, 1992) and Ryle (1949/2009) note that this category applies to many different things; from understanding the meaning of the words we use, or believing that certain state of affairs is the case, to having intrusive thoughts or catchy tunes getting "replayed" over and over "in our heads" (see Ryle, 1949/2009, p. 24). Drawing from this observation, both establish a somewhat loose distinction between *dispositions* and *occurrences* or *states of consciousness* (Ryle, 1949/2009; Wittgenstein, 1953/1958, §149, p. 59, 1980b, §45, pp. 9-10; see also Villanueva, 2019, p. 154)[40]. Sticking to Villanueva's (2019, p. 155) construal of Wittgenstein's approach, he distinguishes dispositions and states of consciousness by pointing out that the former, unlike the latter, aren't "switched off" nor "interrupted when there is some kind of 'consciousness breakdown', as when we fall asleep (…) nor their duration can be exactly measured by using a stopwatch […]. That's why Wittgenstein claims that dispositions lack *genuine duration*" (Villanueva, 2019, p. 155, author's translation, emphasis added; see also Wittgenstein, 1980b, §§ 45, 51, 178, pp. 9, 11, 34, 1992, §MS169, p. 9). In this sense, we don't stop believing that the wallpaper of our former flat resembled that of an old hairdressing salon when we fall asleep, nor we can establish the exact location or duration of our secret, filthy desire that the sanitary lockdown had lasted a couple of months more (at least not in the same sense in which we can point to a physician where exactly our twisted ankle hurts and for how many seconds it unbearably aches if our cat steps on it). This lack of "genuine duration" was repeatedly pointed out by Wittgenstein to show that mental-state ascriptions (specifically, ascriptions of mental dispositions, such as propositional attitudes) do not describe *any* state of affairs; after all, even in the Cartesian framework, minds were taken to be temporally distributed -the *res* cogitans was exempted from *extension* (i.e., spatial distribution), but not from *temporality*.

Wittgenstein offers another insightful reason to reject the idea that ascriptions of propositional attitudes describe some given state of affairs. He articulated it in response to Russell's (1913) relational theory, which constitutes a paradigmatical example of an

[40] A critical difference between Wittgenstein's and Ryle's conception of the distinction between dispositional and occurrent mental states and that of other contemporary approaches is that, for Ryle and Wittgenstein, the distinction doesn't capture an actual, empirically testable difference between two different mental kinds (as, for example, Carruthers, 2013, implies); on the contrary, Wittgenstein and Ryle are concerned with establishing a *grammatical* or conceptual distinction, one that aims at capturing the differential *depth grammar* of superficially similar mental terms (see section 4.1.2.).

internalist descriptivist approach to mental-state ascriptions (see Chapter 3, section 3.1.2.). Already before the *Tractatus* (Wittgenstein, 1961, p. 121; see Villanueva, 2019, p. 151; see also Thornton, 2007), Wittgenstein saw a fundamental problem with this account, specifically in the case of belief ascriptions: roughly, that it allows for *impossible* things to be believed; within the descriptivist view, it would *make sense* to say that one believes something that is, by definition, *unbelievable*, i.e., that cannot even be true nor false, simply because it's senseless (e.g., the existence of a "square circle").

The problem can be stated as follows. If we think that sentences of the form "*S* believes that *p*" describe a relation between an agent and a free-standing mental object "in their head" (e.g., the believed proposition *p*), then it seems to follow that the truth-conditions of the belief ascription are not necessarily dependent from the truth-conditions of the proposition that is believed. After all, two descriptions of two different facts (say, "The pizza is in the oven" and "The cat is in the kitchen") may be independently true or false; the former might be true while the latter is false, and vice versa. This seems to make sense when we think of truth-evaluable propositions like "The pizza is inside the oven" and their related belief ascriptions (e.g., "You believe that the pizza is inside the oven"). In these cases, it's clear that the truth-value of the former and of the latter might differ: the pizza could be in the oven and someone might -regretfully- not believe so, or vice versa. Internal descriptivism takes this as evidence that the truth conditions of each sentence are independent; each represents two different states of affairs, one constituted by a relation between two material objects (e.g., the pizza and the oven), and another constituted by a relation between a material object (e.g., the agent) and a mental object (e.g., the proposition expressed by "The pizza is inside the oven"). However, the problem with this account comes when we think of sentences like "The square circle is in the table" or "I believe that all mimsy were the borogoves", which are obviously senseless, and thus unthinkable. Here, the internal descriptivist would be committed to the idea that the truth conditions of these sentences would be independent from those of their related belief-ascriptions; thus, they would make sentences like "They believe that the square circle is in the table" potentially *truth-evaluable*, and consequently "The square circle is in the table" something *thinkable* or *believable* when, by definition, it's not.

This argument is closely linked to what is commonly known as "Moore's paradox" (Wittgenstein, 1953/1958, § X, pp. 190-191), which also puts unbearable pressure on the idea that mental-state ascriptions describe some free-standing state of affairs. This time, the observation considers the case of belief self-ascriptions (i.e., sentences of the form "I believe that *p*"). Moore's paradox consists in the observation that the truth value of "It's the case that

p" and "I believe that *p*", considered separately, may not be the same: again, the truth of "The pizza is in the oven" is perfectly compatible with the falsity of "I believe that the pizza is in the oven"; the former might be the case when the latter is not, and vice versa. However, their conjunction in the first-person and present tense produces an absurd sentence: one cannot sensibly claim "The pizza is in the oven, but I don't believe that it's in the oven". The point against descriptivism is the following: if both sides of the conjunction described different, independent states of affairs, this kind of sentences should make sense -as it does, for example, when we claim "The cat is in the kitchen, but the cat's food is in the living room". However, it clearly doesn't. What we can conclude from this is, at least, that the truth-conditions of a present-tense belief self-ascription and those of the believed proposition are not logically independent, as those of two descriptions of two different states of affairs may be.

These three arguments (i.e., the argument from the normative force, the argument from non-durability and the argument from truth-conditional dependence) help us see that competent speakers don't use mental-state ascriptions (or, at least, ascriptions of propositional attitudes) to describe some given state of affairs. They thus reveal what Ryle (1949/2009) called the *category mistake* that seems to be underlying the whole Dogma of the Ghost in the machine (see Chapter 2, section 2.1.2.); the one that derives from representing "the differences between the physical and the mental (...) *inside the common framework* of the categories of 'thing', 'stuff', 'attribute', 'state', 'process', 'change', 'cause' and 'effect'" (Ryle, 1949/2009, p. 9, emphasis added). It's important to note here that the root of the category mistake that Ryle has in mind is not in the distinction between two general kinds of substances, but rather on the consideration of minds in factual terms itself. In other words: the root of the category problem is not substance dualism, but factualism -and, relatedly, descriptivism, which forces us to think of the mind in factualist terms.

Similarly, Wittgenstein (1953/1958) wonders:

> 308. How does the philosophical problem about mental processes and states and about behaviourism arise? -The first step is the one that altogether escapes notice. We talk of processes and states and leave their nature undecided. Sometime perhaps we shall know more about them- we think. But that is just what commits us to a particular way of looking at the matter. For we have a definite concept of what it means to learn to know a process better. (The decisive movement in the conjuring trick has been made, and it was the very one that we thought quite innocent.)- And now the analogy which was to make us understand our thoughts falls to pieces. So we have to deny the yet uncomprehended process in the yet unexplored medium. And now it looks as if we had denied mental processes. And naturally we don't want to deny them. (Wittgenstein, 1953/1958, §308, p. 103)

As we view it, "the decisive movement in the conjuring trick" is the commitment to descriptivism about mental-state ascriptions, in both its shallow and deep varieties, i.e., both the assumption that mental-state ascriptions in fact represent some given state of affairs and the assumption that, if they didn't, then they wouldn't be truth-apt). As of now, we've focused on discussing the former claim; the arguments from normative force (Chapter 3, section 3.2.2.), non-durability, and truth-conditional dependence just provide some evidence that mental-state ascriptions don't in fact describe any state of affairs. Now, two questions remain: first, if the primary function of mental-state ascriptions is not to describe some given state of affairs, what is it then? And second, if they don't describe any given state of affairs, are they really truth-apt? In the next section, we'll delve into the former, and in section 4.2.3. we'll turn back to the latter.

4.2.2. The evaluative and regulative function of mental-state ascriptions

A plausible answer to our first question arises from a proper analysis of the argument from the normative force of mental-state ascriptions against their descriptive character. Recall the case of Mustard, Emerald, and Aquamarine (Chapter 3, section 3.2.2.). For the sake of clarity, consider now the following simplified versions of Mustard's and Emerald's claims about Aquamarine:

(1) Aquamarine displays the brain activity pattern BRN.
(2) Aquamarine believes that the shop is open and desires to buy the camera already.

In our example, (1) and (2) were used to establish a prediction about Aquamarine's behavior. Now, sentences like (1) and (2) can also be part of an *explanation* of someone's behavior. Imagine that Aquamarine had never told Emerald about her beliefs and desires, and that Mustard had forgotten to check her teleanalyzing device. Suddenly, they see Aquamarine getting dangerously close to the photography shop. Mustard runs to check her teleanalyzing device, while Emerald tries to guess Aquamarine's mental states. They finally form the following explanations for Aquamarine's behavior, respectively:

(3) Aquamarine goes to the shop to buy the camera because she displays the brain activity pattern BRN.
(4) Aquamarine goes to the shop to buy the camera because she believes that the shop is open and she desires to buy the camera already.

As we've seen, many reductive compatibilists would presuppose that (3) and (4) serve the same purpose: to *causally* explain Aquamarine's behavior by establishing a relation between it and some given facts. Ryle's (1949/2009) repeated attacks to this intellectualist and causalist understanding of mental concepts and his insistence on the distinction between *know-how* and *know-that* (see section 4.1.2.) yield a radically different view.

> [The supporters of the dogma of the ghost in the machine] postulate an internal shadow-performance to be the real carrier of the intelligence ordinarily ascribed to the overt act, and think that in this way they explain what makes the overt act a manifestation of intelligence. They have described the overt act as an effect of a mental happening, though they stop short, of course, before raising the next question—what makes the postulated mental happening manifestations of intelligence and not mental deficiency (...) But when a person talks sense aloud, ties knots, feints or sculpts, the actions which we witness are themselves the things which he is intelligently doing, though the concepts in terms of which the physicist or physiologist would describe his actions do not exhaust those which would be used by his pupils or his teachers in appraising their logic, style or technique. He is bodily active and he is mentally active, but he is not being synchronously active in two different 'places', or with two different 'engines'. There is the one activity, but *it is one susceptible of and requiring more than one kind of explanatory description*. (Ryle, 1949/2009, pp. 37-38; emphasis added)

> In short, then, the doctrine of volitions is a causal hypothesis, adopted because it was wrongly supposed that the question, 'What makes a bodily movement voluntary?' was a causal question. This supposition is, in fact, only a special twist of the general supposition that the question, 'How are mental-conduct concepts applicable to human behaviour?' is a question about the causation of that behavior. (Ryle, 1949/2009, pp. 37-38)

What these passages point to is that when we use mental language to ascertain the intelligent, rational, volitive -in sum, normative- character of our actions we are not providing a causal explanation, but doing something else. In Wittgenstein's (1953/1958, §664, p. 168) terms, we'd say that (3) and (4) are only similar in their *surface grammar*, i.e., in that both share a similar grammatical structure: both (3) and (4) use a causal connector ("because") to establish a relation between an *explanans* ("Aquamarine goes to the shop to buy the camera") and an *explanandum* ("she displays the brain activity pattern BRN", "she believes that the shop is open and she desires to buy the camera already").

However, the different reactions that we should expect from Mustard and Emerald if their expectations were not fulfilled -as it happened in the original example- reveal that the *inferential connections* that can be established between sentences like (1) and (2) and

Aquamarine's behavior are overly different. As we saw in section 4.1.2., these "inferential connections", i.e., the kind of inferences that we're allowed to make when we use an expression in a certain context, are what Wittgenstein (1953/1958) refers to as the *depth grammar* of an expression, and what Ryle (1949/2009) calls its *logical geography*, which amount to its meaning. In the case of explanatory sentences like (3) and (4), the similarities in their surface grammar hide the overt differences in their depth grammar or logical geography. The difference, as we saw, is the following: while a description of someone's neural activity (or patterns of interaction with the environment) may allow us to *scientifically explain* their behavior (i.e., to establish its *causes*), mental-state ascriptions allow us to *rationalize* or *justify* it (i.e., to explain it in terms of their *reasons* to act in a certain way). As Ryle puts it:

> When we ask 'Why did someone act in a certain way?' this question might, so far as its language goes, either be an inquiry into the cause of his acting in that way, or be an inquiry into the character of the agent which accounts for his having acted in that way on that occasion. I suggest, what I shall now try to prove, that explanations by motives are explanations of the second type and not of the first type. It is perhaps more than a merely linguistic fact that a man who reports the motive from which something is done is, in common parlance, said to be giving the 'reason' for the action. (Ryle, 1949/2009, p. 75)

That's why, if Aquamarine failed to act in accordance with Mustard's and Emerald's expectations, the former being based on some empirical inquiry and the latter just based on Aquamarine's self-ascriptions, their reactions would be different. To use Sellars's (1956) and McDowell's (1996, p. xiv) terms, while Mustard's use of (1) and (3) constitutes a move in the *realm of law* or the *logical space of nature*, Emerald's use of (2) and (4) constitutes a move in the *logical space of reasons* (see also Pinedo-García, 2014); at best, moves of the first kind -e.g., scientific descriptions, explanations, and predictions- allow us to make inferences about what behaviors will *in fact* occur given certain circumstances; by contrast, moves of the second kind (e.g., justifications, rationalizations, etc. -what we might call "normative explanations") allow us to draw inferences regarding what behaviors *should* occur (Sellars, 1956; see McDowell, 1996; Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2014, 2020; see also Almagro-Holgado, 2021; Fernández-Castro & Heras-Escribano, 2020; Pérez-Navarro et al., 2019). That's why, as we saw, sentences like (2) and (4) cannot be reduced (i.e., translated) to sentences like (1) and (3).

This distinction between justifications and causal explanations -or between the "logical space of reasons" and the "logical space of nature"- points to a core characteristic of the kind of Wittgensteinian, pragmatist non-descriptivism that we've defended here. As we've

seen, the meaning-as-use conception of language that underlies this approach draws from the assumption that the meaning of a given expression is given by its depth grammar or logical geography, i.e., by the set of "inference tickets", to use another Rylean term, that we acquire when we use it in a particular context, and that such logical geography depends on the norms of the language-games in which it's employed. In addition, what gives this approach its pragmatist flavor is the assumption that the distinction between language-games is not something that can be established beforehand, but that requires the consideration of the different social practices in which the different language-games are circumscribed.

What we want to claim here is that expressions like (1) and (3), on the one hand, and (2) and (4), on the other, when used in situations similar to our example -which we take to be their prototypical uses-, constitute "moves" in different language-games because they subserve radically different *practical* purposes. We can follow here a rough distinction between two general kinds of practices, which amount to adopting two different possible stances when attempting to give an account of someone's actions and reactions. On the one hand, we can talk of a *nomological stance*, defined by the principles of the scientific image (see Chapter 2 section 2.1.). We adopt such a stance when we engage in *causal-explanatory* practices, whose primary purpose is to causally explain, predict, and control a person's actions and reactions; here our main task is to correctly describe and operationalize them, as well as the events that may causally explain them, in order to improve our predictions and our intervention abilities. From the nomological stance, the person's doings are viewed in *subpersonal* or *objectifying* terms, i.e., in terms of natural events, as potentially explainable, predictable, and modifiable as any other natural phenomenon.

On the other hand, we might also talk of an *agential stance*[41], typically defined by our manifest image of the world and ourselves. We adopt such a stance when we engage in *rationalization* practices, where our main goal is not to causally explain, predict, and modify each other's doings, but to *understand* or *comprehend* them, to render ourselves intelligible to one another, and thus morally or epistemically evaluable; in other words, to assess the norm-conforming or norm-deviant character of other's doings regarding different normative standards (e.g., of intelligibility, rationality, morality, psychological wellbeing, and many others) (see also Pinedo-García, 2020; Ramberg, 2000). From such a stance, a person's actions and reactions are viewed in *personal* or *humanizing* terms, i.e., in terms of freedom,

[41] Other authors have advanced similar "stances" to convey this kind of attitude towards each other's doings. Dennett's (1979/1987) famous "intentional stance" is somewhat similar, but insofar as he presents it as a particular kind predictive strategy, it's more appropriate to think of it as a special kind of nomological stance. By contrast, De Haan's (2020, 2021) "existential stance", whereby we view each other's actions in terms of their meaning and intelligibility, is more closely linked to what we want to convey here.

intentionality, autonomy, and responsibility one's their actions (for similar distinctions, see de Haan, 2020a, 2021; Pinedo-García, 2020; Ramberg, 2000; Thornton, 2007).

Critically, there's nothing necessarily "good" or "bad" about each possible stance. In fact, both can be used for disastrous purposes: on the one hand, take for instance the worldwide use of physical, mechanical, and chemical restraints in psychiatric institutions, or the growingly alienating character of labor conditions in ever more productive areas; on the other hand, take the current worldwide pleas for "personal liberty" or "freedom of expression" to remain happily unchallenged for one's decision to make harassing gypsyphobic jokes or demonize measures intended to reinforce the welfare state. What we're claiming here is just that they are *different* stances, different explanatory and intervention strategies, which are not necessarily incompatible nor mutually exclusive, but neither assimilable to one another.

It is these different stances and the different social practices they reflect which ground the different language-games in which we might engage when trying to give an account of someone's doings -and, consequently, the different logical geography of different kinds of claims. Based on our rough distinction between causal-explanatory and rationalization practices, we can draw another rough distinction between *descriptive* and *evaluative* language-games. What our example above shows is that when we self-ascribe and ascribe mental states to one another, when we engage in our daily folk-psychological interpretative mindgames, we're not trying to describe some private and nebulous microcosmos (as internal descriptivists would take it), nor someone else's patterns of neural or behavioral activity (as external descriptivists would take it) that may be causally related to someone's actions. Instead, what we're doing is expressing and acquiring certain commitments to certain norms (e.g., of rationality, intelligibility, moral adequacy, psychological wellbeing, etc.), in light of which our actions and others' can be assessed in personal terms. In other words: the primary function of folk psychology is not descriptive, but *evaluative* and *regulative*, and it primarily finds its place in rationalizing accounts of one another (see Almagro-Holgado, 2021; Almagro-Holgado & Fernández-Castro, 2019; Fernández-Castro, 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019; Kalis, 2019; McGeer, 2007, 2015, 2021; Pérez-Navarro et al., 2019; Pinedo-García, 2020; Villanueva, 2018, 2019; Zawidzki, 2008). Some recent authors have expressed this view by claiming that folk psychology is not -or at least not primarily- about *mindreading* (i.e., describing and causally explaining one another), but about *mind-making* or *mindshaping* (i.e., reciprocally regulating our actions in order to make them norm-conforming) (Fernández-Castro, 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019; McGeer, 2007, 2015, 2021; Zawidzki, 2008, 2013).

Another way to put this is as follows: *genuine* -or, at least, prototypical- uses of mental-state ascriptions make ourselves and others *responsible* for acting in some ways and not acting in others. This "responsibilizing" function thus is a double-edge tool: it's used both to *evaluate* our past and ongoing actions and, at the same time, to *prescribe* and *regulate* how we should act in the future if we are to remain being viewed as correctly following such norms.

This applies both to the first- and the third-person uses of mental-state ascriptions. Let's first analyze first-person uses (i.e., mental-state *self*-ascriptions), for it will then be easier to understand our position in the case of third-person uses. From our non-descriptivist point of view, when we self-ascribe a mental state, we put our past, present, and future behavior under the light of certain shared norms, and so others -or even ourselves adopting a third-personal stance towards our own behavior- can assess whether we act in accordance with such norms or not, i.e., whether we're acting in accordance to what follows, within a certain form of life, from such self-ascription. Another way to put this is that we acquire a series of *commitments* to undertake certain courses of action (Almagro-Holgado, 2021). Take belief self-ascriptions, for example. What specific commitments do we acquire when we self-ascribe the belief that "the pizza is in the oven", for example? According to the Wittgensteinian (and Rylean) view of language that we presented above, what we acquire is a commitment to undertake any conceivable course of action that, within a certain language, community and form of life, follows from what we claim to believe in, desire, intend to produce, etc. (i.e., what follows from the "p" in "I believe that p") (section 4.2.1.). That's why Wittgenstein (1921/2001, §5.542, p. 64) pointed out that "It's clear (...) that 'A believes that p', 'A has the thought p', and 'A says *p*' are of the form '"*p*" says *p*'". What this means is that, as we saw earlier, the truth-conditions of sentences of the form "I believe that *p*" and "*p*" (i.e., the believed proposition) are related; hence one cannot sensibly say something like "The pizza is in the oven, but I don't believe that the pizza is in the oven", nor "I believe that the square circle is on the table" (because nothing follows from such sentence). Thus, when one says "I believe that the pizza is in the oven", what one is doing is expressing a commitment to what follows from asserting the sentence "the pizza is in the oven" in our language, i.e., its logical geography.

Now recall that, in the Wittgensteinian (and Rylean) conception of language, understanding what follows from a certain assertion (i.e., grasping the "norms" that determine its meaning and its truth-conditions) is not a matter of contemplating some regulative proposition in our heads, but a matter of displaying some kind of practical ability; in other words: it's not a matter of know-that, but of know-how. Thus, in uttering "I believe that the pizza is in the oven", we acquire a series of *practical* commitments: we commit ourselves to

undertake whatever possible courses of actions may follow from self-ascribing such mental state; one's past, current, and future actions can thus be evaluated in terms of their conformity to the norms that, in our language, determine what is logically linked to asserting the content of our self-ascribed belief: provided certain conditions (e.g., that one wants to eat pizza, that one doesn't have any movement limitations, etc.), if we make such self-ascription, we acquire a commitment to look for the pizza in the oven -and not, say, in the fridge-, to answer "in the oven" if asked "where's the pizza?" -and not answering "in the fridge"-, and a potentially infinite set of other courses of action that, in our community, we would sanction as "correct" or "incorrect" in light of such belief self-ascription.

And the same goes for those cases where we interpret other's doings in terms of mental states. Again, what we're doing here is not pointing to some hidden cause nor merely describing the other's behavior, but just evaluating their actions in terms of rule-following, and prescribing how they should act in the future if they're to remain being viewed as proper rule-followers. In doing so, we're ourselves also expressing a commitment to what follows from our mental-state ascriptions: when we take others' doings to be cases of believing, desiring, intending, etc., we're showing what we view as normatively linked to such ascriptions; if we acted in the same way, in the same circumstances, we should evaluate ourselves in terms of the same mental states (Almagro-Holgado, 2021; Fernández-Castro, 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2020; Frápolli & Villanueva, 2012, 2013; Pérez-Navarro et al., 2019; Pinedo-García, 2020).

What we've just said is well-captured by both Ryle and Wittgenstein's understanding of mental states (and, particularly, propositional attitudes) as *normative dispositions*. To ascribe a mental state to someone (or to oneself) is to view them as disposed to take the *right* courses of action in the *right* circumstances. However, it's important to note here that neither Ryle nor Wittgenstein view mental states as *causal* devices (as some functionalists take it), nor as mere behavioral dispositions (e.g., mere "ascriptive shorthands" to refer to a finite set of behavioral counterfactuals), as it's typically been claimed of both (see Tanney, 2009). Instead, as Almagro-Holgado (2020) has pointed out, Ryle (1949/2009; pp. 31-32) is careful to draw a distinction between "simple" or "single-track" and complex or "higher-grade" dispositions; while the former can be understood in purely factual and causal terms (e.g., the solubility of sugar in water would be one such example), the latter necessarily involves a reference to norms. Complex dispositions are thus *normative* dispositions, and mental vocabulary is dispositional in this last sense: again, ascribing a mental state to someone or to ourselves is not just describing what someone would do if certain circumstances hold, but what someone should do in certain circumstances if they're proper norm-followers.

Thus, this regulativist and evaluativist approach to folk psychology dispels the intellectualist legend for good. To characterize someone as being in a certain mental state is to characterize them as following certain rules, but this doesn't mean to describe them as "contemplating some regulative propositions in their heads", nor as having some "neural representations" encoded in their brains, nor as verbalizing (overtly or covertly) those propositions out loud from time to time; it's not to describe *two* occurrences (the person's behavior and some internal or external fact that may cause it), but to *evaluate* one occurrence (the person's behavior) as an actual case of following a certain rule, as a correct movement within a social practice. In this sense, a crucial aspect of this regulativist and evaluativist approach to folk psychology is that it allows us to establish a distinction between *expressing* one's mental states and *saying* that one is in a certain mental state (Almagro-Holgado, 2021; Villanueva, 2014, 2018). As we saw in section 4.1.2., "to *think* one is obeying a rule is not to obey a rule" (Wittgenstein, 1953/1958, §201, p. 81), and hence to think (or say) that one is in a certain mental state is not the same as actually being in that mental state (i.e., really expressing it in one's actions). To see the intuitive grip of this distinction, consider the following example:

> Chalk, a White, cis-heterosexual young man in his 20's, doesn't consider himself a racist, sexist nor LGBT-phobe person. In fact, when asked about or reflecting himself on these topics, he covertly and overtly affirms things like "I believe that racialized people deserve equal treatment", "I hope that, one day, men and women will finally be treated as intellectual peers", or "I love that people nowadays are so free to explore their sexuality and question traditional gender roles". However, Chalk is unaware that, in numerous occasions, he is biased to pay more attention to his work colleagues' opinions when they come from men rather than from women. He also sometimes changes sidewalk when facing a big group of racialized people in the street, and he would have serious difficulties accepting that his future children decided to undertake a testosterone treatment because they don't self-identify with any of the two options imposed by gender binarism.

In this example, Chalk's overall patterns of actions and reactions (taking this to include both Chalk's overt and covert behavior) seem to at least preclude a straightforward evaluation of his sincere mental-state self-ascriptions (e.g., his professed beliefs and desires about gender, racial, and LGBT+ equality) as true. In other words: depending on our evaluative framework, we'll be more or less inclined to say that there's a mismatch between what Chalk says to others and to himself about his mental states and what should follow from being in such mental states. Other than our example, it's evident that we sometimes fail -and sometimes spectacularly- to identify or correctly assess which are our own mental

states (e.g., see Coliva, 2016; Schwitzgebel, 2008; Srinivasan, 2015; see also Almagro-Holgado, 2021). In this sense, our pragmatist kind of non-descriptivism helps us to do away with one foundational epistemological tenet of Cartesianism: the idea of the privileged access to one's own mental states, or the idea that one is always in a privileged epistemic position to determine whether one *is* in a certain mental state or not. In fact, from our perspective, we need multiple, constant engagement with others to gain a more precise view of what our own mental states are (i.e., what norms we in fact follow) and how to effectively change them (i.e., how to start acting so that our behavior can be evaluated as meriting the "desired" mental-state ascriptions).

Obviously, an underlying assumption here is that there's actually *something* to say about the "truth" or "falsity" of mental-state ascriptions -that there's someone, if not always oneself, who might be at some point in an authoritative position regarding whether someone *really* has certain mental states. How come, if mental-state ascriptions don't describe any specific state of affairs? This was the second question that we asked at the end of the previous section, and the one that we must answer if we're to escape the puzzle of translatability.

4.2.3. Truth-evaluability and the post-ontological approach to the mental

So far, we've challenged the shallow or affirmative version of the descriptivist stance regarding mental-state ascriptions (i.e., that mental-state ascriptions do in fact describe some given state of affairs). We've seen that some of the arguments against descriptivism in its shallow version point to an alternative evaluativist and regulativist conception of folk psychology, according to which the primary function of mental-state ascriptions is not to describe someone's doings nor attempting to causally explain them, but to evaluate their correctness or incorrectness under the light of myriad different possible norms. Mental-state ascriptions thus find their place within rationalizing or agentializing, rather than nomological practices.

This provides a working answer to the first of the two questions that we posited at the end of section 4.2.1. (i.e., if the primary function of mental-state ascriptions is not to describe some given state of affairs, what is it then?). Now we must turn to our second question: if they don't describe any given state of affairs, are they really truth-apt? This last question leads us to the crux of the matter: can our favored pragmatist kind of non-descriptivism provide a successful way out of the puzzle of traducibility? In other words: can it account for how mental-state ascriptions might be truth-apt without buying into either reductivism or non-naturalism about the mental?

Now it's time to challenge the deep version of descriptivism, or the idea that only *if* mental-state ascriptions describe or represent some given state of affairs, then they can be

assessed in terms of their truth or falsity. This was, as we saw at the end of Chapter 3, the underlying premise that forced naturalism into the self-defeating path of reductivism or incompatibilism (and non-naturalism into the self-defeating path of private rule-following); or, to put it in Wittgenstein's (1953/1958) terms, "the decisive movement in the conjuring trick" (§308, p. 103). Cartesianism draws from it to establish a distinction between *two ways of existing*. Instead, naturalists assume that there's *only one possible way of existing*. On one side, reductive compatibilists attempt to make room for the mind within their unitary ontology, thus assuming that mental-state ascriptions ultimately describe some given natural fact. However, the aforementioned arguments from normative force, truth-conditional dependence, and non-durability put unsurmountable pressure on this assumption. Non-reductive incompatibilists, in turn, draw from similar arguments to assume that, if we want to remain faithful to naturalism, we're now left with only one option: to claim, tragically, that minds -as well as any other thing that can't be defined in pure descriptive terms, e.g., logics, truth, etc.- *don't really exist*, and so on. In doing so, they fail to reject descriptivism, in the conditional sense. In fact, they express a renewed commitment to it; ironically, in their furious attempt to eradicate all trace of Cartesianism from their philosophical and scientific accounts, their incompatibilism ends up falling short of their own radical aspirations, since it just expresses an inability to break through the "logical mould into which Descartes pressed his theory of the mind" (Ryle, 1949/2009, p. 9).

To stress the unfounded absurdity of the descriptivist framework in which non-naturalists and naturalists collide, let us present one last example. Keeping in line with Wittgenstein's game language, let's see what descriptivism would tell us about some chess piece; pawns, for instance. Let "pawn-sentences" be those sentences that include the term "pawn". What would a descriptivist analysis tell us about such term and about the sentences in which it appears?

One the one hand, it's clear that "pawn" cannot be translated to a mere description of material facts: "pawn" cannot be defined, for instance, as "a wooden block that people usually move in such and such directions over an 8x8 black and white square board"; after all, it may be made of any possible material and, most importantly, such definition wouldn't allow us to distinguish between correct and incorrect moves with it -which, we might say, is *of the essence* of the concept of "pawn". It seems clear that a correct definition of "pawn" must necessarily contain an indication of the possible movements that we're allowed to do with it (e.g., the chess piece "of least value having the power to move only forward ordinarily one square at a time, to capture only diagonally forward, and to be promoted to any piece except a king upon reaching the eighth rank", see Merriam-Webster, n. d.). By contrast, all that a

purely materialist description of pawns could tell us is which are the most or the least statistically frequent movements. If so, we seem to encounter a problem here: what we could call *the problem of "chessity"*.

Now, it would be absolute nonsense if, in order to avoid the problem of chessity, we decided that "pawn" in fact describes some "internal essence" of certain blocks, a non-extensional, ghostly entity inside them that somehow causally explained the movements that we do with them on the board. This would lead us straight to what we could call *the pawn-block problem*. At this point, we seem again forced to choose between non-naturalism or incompatibilism about pawns: either we assume some kind of ontological duality or we just proclaim, with grievous countenance, that "pawns don't really exist", and that pawn-sentences are either literally false or not even truth-evaluable.

As we can see, descriptivism does with pawns and pawn-sentences the same that it does with minds and mental-state ascriptions. By contrast, our favored Wittgensteinian and Rylean kind of non-descriptivism about the mind points in a radically different direction. It comes along with ontological radicals in that the normative force of concepts like 'pawn' or 'belief' is "of the essence" of such concepts: to say of some wooden piece that it's a pawn amounts to saying that such piece *must* be moved in a certain way, and not that people in fact tend to move it that way; and to say of someone that they believe something amounts to saying that they *should* act in certain ways, and not that they in fact tend to act in certain ways or that there's something inside or outside their skull that makes them act like that. At this point, however, Wittgenstein's radical methodological maxim is recalled: "don't think, but look!". That is: instead of drawing from a preconceived prescription to judge which sentences have meaning and which not, or which might be truth-evaluable and which not, the Wittgensteinian conception of language and meaning assumes that we must draw from what is obvious at looking straight into the communicative practices in which they're used. We must thus take the aforementioned observations at face value: if such normative or prescriptive force is "of the essence" of "pawn-sentences", or of the kind of mental-state ascriptions that we use in our daily mindgames, then well, *that's* how pawn-sentences and belief ascriptions work, i.e., that's how they are used by competent speakers, and that's what reveals their depth grammar or logical geography (see Price, 2011, Price et al., 2013; see also Fulford & van Staden, 2013).

This movement helps us see that from the fact that we cannot translate pawn-sentences nor belief ascriptions to descriptions of fact it does not necessarily follow that pawns or beliefs either "exist on a different ontological realm" or "don't really exist" (i.e., that pawn-sentences or belief ascriptions either describe some inner phantom or are literally false or

not even truth-apt); these absurdities only arise from maintaining a preconceived and unfounded view of meaning and language: that the only function of language is to describe the world and that only descriptions of some possible state of affairs are meaningful or truth-apt. Instead, our favored "deep" or radical variety of non-descriptivism takes it that what the arguments from non-durability, normative force, and truth-conditional dependence reveal is just that we might use "exist" in many different ways; in other words, that sentences containing terms like "belief" or "pawn" may be as true or false as any description of some material state of affairs, the difference being in the different norms that govern different kinds of linguistic interchanges within a given community. Wittgenstein's (1953/1958, §304, p. 102) seemingly paradoxical remark about the mind, i.e., that "It is not a *something*, but not a *nothing* either!" can thus be understood in this sense (see also Pinedo-García, 2014). A similar observation can be made about Ryle's (1949/2009, pp. 11-12) analysis of the root of the category-mistake:

> When two terms belong to the same category, it is proper to construct conjunctive propositions embodying them. Thus a purchaser may say that he bought a left-hand glove and a right-hand glove, but not that he bought a left-hand glove, a right-hand glove and a pair of gloves. 'She came home in a flood of tears and a sedan-chair' is a well-known joke based on the absurdity of conjoining terms of different types. It would have been equally ridiculous to construct the disjunction 'She came home either in a flood of tears or else in a sedan-chair.' Now the dogma of the Ghost in the Machine does just this. It maintains that there exist both bodies and minds; that there occur physical processes and mental processes; that there are mechanical causes of corporeal movements and mental causes of corporeal movements. I shall argue that these and other analogous conjunction are absurd; *but, it must be noticed, the argument will not show that either of the illegitimately conjoined propositions is absurd in itself. I am not, for example, denying that there occur mental processes*. Doing long division is a mental process and so is making a joke. But *I am saying that the phrase 'there occur mental processes' does not mean the same sort of thing as 'there occur physical processes'*, and, therefore, that it makes no sense to conjoin or disjoin the two. (Ryle, 1949/2009, pp. 11-12, emphasis added)

Ryle thus sees the absurd character of the official doctrine as residing in descriptivism and factualism, not in the truth-evaluability of mental-state ascriptions. What this Wittgensteinian and Rylean conception of mental language points to is that, if we want to dispel the "conjuring" and find "a way out of the fly-bottle" (Wittgenstein, 1953/1958, §§ 308-309, p. 103) -i.e., a way to escape both the mind-body problem and the problem of normativity at once (or the pawn-block problem and the problem of chessity, for that matter)- we

must recognize that "it is perfectly proper to say, in one logical tone of voice, that there exist minds and to say, in another logical tone of voice, that there exist bodies. But these expressions do not indicate two different species of existence" (Ryle, 1949/2009, p. 12)[42]. What Ryle is advancing here is a *post-ontological* account of mind (see Ramberg, 2000; see also Heras-Escribano & Pinedo-García, 2018; Pinedo-García, 2020), i.e., an approach that does not think of the difference between the mental and the non-mental in terms of an "ontological plurality" but in terms of a "linguistic plurality"; not in terms of a plurality of "ways of existing", but in terms of the plurality of language-games that competent speakers within a given linguistic community engage in and the plurality of *criteria* that they use to determine the truth or falsity of different sentences.

Importantly, this approach not only excludes strongly "realist" (i.e., factualist) approaches to the mind, but also certain forms of "subjectivism" or "anything goes" relativism about our folk-psychological interpretative practices, whereby one would get to decide upon the truth value of a given belief ascription at will (see Pérez-Navarro et al., 2019; see also Pérez-Navarro, 2021). Quite to the contrary, on this approach, the truth value of mental-state ascriptions depends on the norms and standards that interpreters *share* as members of the same linguistic community and participants of shared forms of life -norms which we're trained to follow by our respective communities, whose following we express in practice, and over which we cannot "jump" at will (see Pérez-Navarro, 2021). It's these shared norms that sanction the exercise of our interpretative abilities, and although we might sometimes come to apparently unsolvable disagreements about our interpretations of someone's mental states or about the criteria that we should follow in some particular case (see Curry, 2020; Pérez-Navarro et al. 2019), this doesn't mean that mental-state ascriptions "lack real truth value" or are "not truth-apt".

In sum, the pragmatist and pluralist kind of non-descriptivism afforded by Wittgenstein's and Ryle's work provides us with a way out of the puzzle of translatability; in other words, it provides us with a naturalist, non-reductivist, yet compatibilist account of the place of mind on nature. In this approach, mental-state ascriptions do not amount to a description

[42] Similar arguments to the ones we've seen here and in Chapter 3 regarding the self-defeating character of traditional naturalisms have motivated more *liberal* or *relaxed* forms of naturalism (see also Caro & Macarthur, 2004, 2022; Hutto, 2022; McDowell, 2004; Price, 2004, 2011; see Thornton, 2007, for an application to the philosophy of mental health). Although we won't endorse any of these proposals in particular, we think that the post-ontological view of mind offers a way to develop such kind of position: if we allow for different meanings of "existing" or "being true", then "the world", identified with "what exists" or "what is the case", can include things other than natural objects (e.g., meaning, norms, inferential connections, minds, etc.), many of which are preconditions themselves of scientific thought. In this sense, our approach is akin to some of these alternative forms of understanding our naturalist commitments -e.g., Price's (2004, 2011) *subject naturalism*.

or representation of some particular state of affairs that may or may not be causally related to the agent's actions. Instead, "*having* a mental state*"* or "*being in* a mental state" just amount to being "truthfully ascribed such mental state" within certain evaluative language-games; what this approach challenges is that describing how different states of affairs are spatial-temporally and causally related is the only possibly truth-evaluable practice, and that the only way for mental-state ascriptions to be truth-apt is by representing some given state of affairs. Instead, the truth or falsity of different mental-state ascriptions will depend on the norms of interpretation that competent language users follow when evaluating others' or even one's own doings in normative terms.

At this point, we're left wondering: "but what norms, *exactly*"? In the next section, we'll introduce some final considerations in this regard. We won't delve into a full-fledged development of the different possibilities at stake here, but a couple of remarks on this issue might be in place to understand the hallmark of the pragmatist kind of non-descriptivism that we've endorsed here. In addition, some of the issues we'll discuss here will be relevant in Chapters 5 and 6 when we address the contemporary debates around the conceptualization of delusions.

4.2.4. Pluralism and the norms of folk-psychological interpretation

As we've seen in the previous section, non-descriptivism encourages a shift from thinking about the mind in factual terms -i.e., in terms of more or less bizarre objects, properties, or relations among them- to thinking of it in post-ontological, primarily normative terms -i.e., in terms of the norms that competent speakers follow in their interpretative practices.

A prime example of this way of thinking about the mind is reflected in some contemporary treatments of the issue of self-knowledge, understood as a person's capacity to know what their own mental states are, and its relation to first-person authority, i.e., understood as the idea that we shouldn't doubt a person's mental state self-ascriptions. As we saw in Chapter 2 (section 2.1.3.), the Cartesian characterization of self-knowledge is utterly traversed by troubling ontological commitments: we are supposed to have first-person authority because we have a privileged and immediate method (i.e., introspection) to access our mental states, understood as inner and private ontological weirdos. By contrast, many contemporary authors have attempted to characterize self-knowledge and first-person authority in non-ontological and "non-detectivist" ways, emphasizing their normative and inherently social character (see Bar-On, 2015; Borgoni, 2019, forthcoming; Coliva, 2016, Davidson, 1984; Srinivasan, 2015; Wright, 1998; Villanueva, 2014; see also Almagro-Holgado, 2021; Fernández-Castro & Heras-Escribano, 2020). On these approaches, the problem of the knowledge of other minds and the problem of the knowledge of one's own mind are given a similar

treatment; rather than positing different epistemic methods to access each other's private "theatres of consciousness", the focus is put on mental-state ascription and self-ascription practices: which norms govern these interpretative practices? Which criteria do we follow when we self-ascribe mental states to ourselves? And which do we follow when we assess the truth of such self-ascriptions?

Against this background, the issue of one's authority over one's own mental states is treated as a particular norm of ascription (Almagro-Holgado, 2021, Almagro-Holgado et al., 2021; Borgoni, 2019, forthcoming, Coliva, 2016; Villanueva, 2014). Some take it that, once we reject the Cartesian notion of privileged access, we must also reject first-person authority, or at least give it less importance than it's been usually given (e.g., Schwitzgebel, 2008, 2013). Others, by contrast, think that we can still retain the main intuitions behind the idea of first-person authority without endorsing Descartes's theory of mind (e.g., Borgoni, 2019, forthcoming; Wright, 1998).

Among the former, several authors have insisted that the notion of "privileged access" faces unsurmountable objections. There are some core conceptual challenges to this idea, chiefly among them Wittgenstein's (1953/1958; see also Kripke, 1982) argument against the possibility of following a rule privately (see Chapter 3, section 3.2.3.). In addition, there're several empirical challenges: as various authors have pointed out, cases of self-deception or error about one's mental states (e.g., Chalk's example above, Red's case in Chapter 5) are relatively common, and they suggest that we're surprisingly bad at identifying our desires, beliefs, or even our ongoing phenomenal states (e.g., Schwitzgebel, 2008; see also Almagro-Holgado, 2021; Coliva, 2016). Drawing from these and other similar cases, some authors have challenged to a lesser or greater extent the presumption of first-person authority, i.e., the idea that the sincerity of one's mental self-ascriptions is sufficient for determining their truth or falsity. Instead, these authors emphasize the role of one's overall degree of consistency in one's behavioral, cognitive, and phenomenological activity (e.g., Schwitzgebel, 2002, 2013, 2021). Some of the most relevant approaches in the debate on the doxastic status of delusions that we'll see in Chapters 5 and 6 endorse similar views: classical *interpretivists* (see Davidson, 1986; Dennett, 1979/1987; see also Byrne, 1998), for example, propose other interpretation rules, such as rationality, intelligibility, or predictability of one's overall behavior. In a similar vein, some varieties of *functionalism* (e.g., Schwitzgebel, 2002, 2013) have pointed out that being in a certain mental state amounts to displaying "a dispositional profile that matches, to an appropriate degree and in appropriate respects, a stereotype for that attitude, typically grounded in folk psychology" (Schwitzgebel, 2013, p. 75).

We're sympathetic to this kind of approaches, and we agree that, on many occasions, one is not one's "best acquaintance"; it's others (most often, significant others) who are in a better position to tells us what we really think or how we feel about some matter. However, these approaches don't seem to accommodate many cases where competent speakers seem to rely almost exclusively on a person's sincere self-ascription to determine their mental state, regardless of how consistent the person's behavior is (see Blue's and Green's cases in Chapter 5). As we'll see in Chapters 5 and 6, this seems to be the case with people with delusions whom, although sometimes fail to act upon their delusions, are nonetheless straightforwardly interpreted as believers of the delusional content by other people (see Rose et al., 2014).

Other authors have still insisted that, at least for certain kinds of mental states and in most cases, one is (or should be) considered as authoritative regarding one's own state of mind by default -even if we're prone to fail at identifying them sometimes (Borgoni, 2019, forthcoming; Coliva, 2016; Davidson, 1984; Wright, 1998). Borgoni (2019, forthcoming) provides a case in point. She has explicitly argued against one implicit commitment in many approaches to self-knowledge and first-person authority: namely, that it's the former, given its alleged special features, which should ground the latter. According to her, first-person authority isn't grounded on some privileged way to access our mental states (e.g., immediate knowledge by introspection); rather, it's just a characteristic feature of how many of our daily interactions with each other work. As she views it, "saying that "someone has first-person authority" [...] means that she has the right to be deferred to when it comes to communicating her mind" (Borgoni, forthcoming, p. 15). She thus views first-person authority as a norm; in particular, as an *interpersonal* norm:

> Default acceptance, absence of doubt and explicit query are ways of treating of our interlocutors' expressions of their mental states that imbue them with first-person authority. People are granted authority with regard to their minds to the extent that they are believed by default when they communicate what they feel, fear, wish, or believe. The phenomenon is thus intrinsically social: it governs how we deal with people's expressions of their minds in interpersonal communication. [...] If first-person authority governs how we treat our peers' expressions of their minds in communicative interpersonal relations, it seems correct to characterize it as a *norm*: an interpersonal norm. When we defer to others with regard to their states of mind as pictured in the three circumstances listed above, we are following the norm. (Borgoni, forthcoming, p. 15)

Thus, according to this kind of approach, first-person authority is the primary norm or criteria that competent language speakers use (or should use) when assessing the truth or falsity of a mental-state ascription: one is or should always be treated as authoritative regarding one's own mental state self-ascriptions. Here, first-person authority involves an ethical component: if someone claims "I believe that *p*", and they're sincere, then a proper thing to do in most cases is to take their self-ascription at face value; others should grant the truth of one's belief self-ascription by default.

We're also sympathetic to this kind of view of the relation between self-knowledge and first-person authority. To begin with, it properly reflects how many of our interpretative practices work. After all, when we want to know what other people think or feel about some matter, a common method is to ask them, and we usually accept such self-ascriptions without further questions; this approach is able to accommodate these practices without resorting to the epistemological and ontological resources of Cartesianism. In addition, as we'll see in Chapter 6, it opens up a way to understand why we tend to interpret cases of delusions in terms of beliefs and, most importantly, why we *should* continue to interpret them as such.

However, as Almagro-Holgado (2021) and Villanueva (2014) have pointed out, if we put too much weight on first-person authority, we're left unable to account for many cases where we don't seem to be interested in the person's sincerity to determine the truth of their mental-state self-ascription (for several examples, see Villanueva, 2014). From our point of view, both strands of the debate -those against first-person authority vs. those in favor of adopting it by default- reflect the functioning of our interpretative practices, but only partially. A major problem of some of these approaches is that they seem to advance one or another *golden rule* for interpretation (e.g., first-person authority, overall rationality, predictability, conformity to some established folk-psychological stereotype, etc.). In doing so, some of them end up endorsing a view according to which ascribing beliefs or other mental-state ascriptions is, after all, a description: not of some brute fact, but of whatever falls under such golden rules (e.g., some given pattern of behavior, the person's mental-state self-ascription, etc.) (see Almagro-Holgado, 2021; Villanueva, 2014, 2018).

By contrast, on the Wittgensteinian and Rylean view of mind and language that we've endorsed here, there are no such golden rules. As we've seen, "following a rule" in this framework is not a matter of *knowledge-that* (i.e., a matter of contemplating some pre-specified -or even specifiable- regulative proposition and then acting in consequence), neither in the case of the person who claims to be in a certain mental state, nor in the case of the person who assess the truth-value of such mental-state ascription. It is a matter of *knowledge-how*, of having been sufficiently trained in certain practices by a certain

community, and then showing that practical ability in *actual*, concrete cases. In this sense, the pragmatist non-descriptivist view endorsed here allows for a pluralistic and contextualist view of the norms and criteria that competent language users employ to decide on the truth or falsity of mental-state ascriptions; no fixed, golden rule can be given once and forever to account for how competent speakers assess the truth or falsity of different mental-state ascriptions, because this will vary depending on the context of ascription (i.e., the specific features of the situation in which an ascription or self-ascription is made) and on the evaluative standards of the community or form of life of reference (Almagro-Holgado, 2021; Curry, 2020; Fernández-Castro & Heras-Escribano, 2019; Pérez-Navarro et al., 2019).

This kind of approach to our belief ascription and self-ascription practices has been deployed in Villanueva's (2014) "expressivist strategy" to account for self-knowledge and, in particular, for the relevance of first-person authority when assessing the truth of a particular mental-state self-ascription. According to such strategy, a particular mental-state self-ascription should be considered to be true "when uttered in a suitable context c, if and only if a contextually salient set of features only makes sense if the avowal is not taken to be false" (p. 7). Almagro-Holgado's (2021) notion of "contextual first-person authority" or "contextual authority", according to which "there are contexts in which there is a presumption of authority regarding a mental self-ascription, contexts in which the speaker exhibits a strong authority, and contexts in which there is neither strong nor presumptive authority" (p. 179) echoes this strategy.

In line with Borgoni's (2019, forthcoming) view of first-person authority, this approach takes it to be an important interpersonal norm of interpretation; one which we probably follow in many cases and which we *should* actually follow in many others. In this respect, Borgoni's (2019) examples of slaves and women being systematically denied the authority over their own mental states are compelling enough, and they clearly reveal the ethical dimension of first-person authority, which we'll further delve into in Chapter 6 when considering the case of people with delusions. However, Villanueva's (2014) expressivist strategy or Almagro-Holgado's (2021) contextual authority allow us to retain this while, at the same time, accommodating the irreducible context-relativity of our interpretative practices. In line with Wittgenstein's characteristic philosophical method, this kind of approach emphasizes the need to pay attention to the specific details of each singular case. Sometimes, the variables of the context will prompt the consideration of the person's sincerity as the major piece of evidence in favor of the truth of a given mental-state ascription; in other cases, like in Chalk's case above, the overall consistency in the person's doings will tend to weight more, thus

rendering the sincerity of their self-ascriptions less relevant (Almagro-Holgado 2021, Villanueva, 2014).

In sum, the main maxim of this pluralist account of folk-psychological interpretation is the following: if we want to determine which criteria govern our mental-state ascription practices, we must proceed on a case-by-case manner; no idealized model will be able to capture the wide variety of norms that we might follow in different contexts to determine the truth or falsity of different mental-state ascriptions. In Chapter 6, we'll highlight the main implications of this approach to current debates on the proper conceptualization of delusions.

4.3. Conclusion

In the previous chapter, we saw how descriptivism, the implicit semantic commitment of Cartesianism and many contemporary approaches to the philosophy of mind, leaded us to a two-way dead-end, whereby we seemed forced to choose between a self-defeating kind of normativism (i.e., non-naturalisms like Descartes's) or a self-defeating kind of naturalism (i.e., reductive compatibilism or non-reductive incompatibilism). In this chapter, we've seen how a pragmatist kind of non-descriptivism, inspired by Wittgenstein's and Ryle's conception of mind and language, frees naturalism and compatibilism from the reductivist straightjacket into which descriptivism forces them: in its rejection of descriptivism about mental-state ascriptions, it provides us with a *non-reductive, yet compatibilist* approach to the place of mind on nature, thus affording a way out of the puzzle of translatability. It thus enables us to accommodate all three attractive features of non-reductivism and compatibilism: *Nomological Power* (i.e., the idea that folk psychology shouldn't constrain scientific psychology), *Truth-Aptness* (i.e., the idea that mental-state ascriptions are truth-apt), and *Normative Force* (i.e., the idea that mental-state ascriptions rationalize behavior) (see Chapter 3, section 3.2.1.).

Specifically, the Wittgensteinian framework endorsed here rejects both descriptivism about what we do with words and about what is said with them; instead, it assumes functional pluralism about language (i.e., the idea that language may be used for myriad purposes, including describing the world, but also issuing orders, making promises or declarations, etc.), as well as the idea that the meaning (and truth-conditions, when relevant) of a given expression doesn't depend on its representational capacity. In this sense, we've seen that Wittgenstein's conception of language goes further than other non-descriptivist approaches in that it not only rejects descriptivism in its shallow, affirmative version (i.e., the claim that all declarative sentences in fact represent some state of affairs), but also in its deep, conditional version (i.e., the claim that only successfully representational or descriptive sentences

are truth-apt). Instead, it assumes that it is what we do with different linguistic expressions, the kind of moves that we make with them in different language-games, what determines their meaning, their truth-evaluability, and their truth-conditions. Specifically, the notion of "language-game" captures three core features that we see as central to the Wittgensteinian view of language: a) the idea that language is a normative system of logical relations among concepts, where the meaning of an expression is given by its "depth grammar" or "logical geography" (i.e., the conceptual or inferential connections that it has with other expressions, namely those that could potentially justify it and those that could potentially be justified by it); b) the idea that such normative system is grounded in the different social and communicative practices that characterize the forms of life into which we are born, and in which we're systematically trained by our linguistic community; and c) that there need not be any necessary and sufficient condition for something to qualify as a "language", and that the possible similarities among language-games are best understood in terms of "family resemblances".

Applied to the analysis of the meaning of mental-state ascriptions, this approach yields three arguments in support of the idea that mental-state ascriptions don't describe any state of affairs, neither internal or "private" nor external or "public": the argument from non-durability (i.e., the idea that mental dispositions don't have "genuine duration"), the argument from truth-conditional dependence (i.e., the idea that the truth-conditions of "p" and "I believe that p" are not independent from each other) and the argument from normative force (i.e., the idea that the normative or prescriptive force of mental-state ascriptions is "of the essence" of the meaning of such linguistic expressions). Taken together, these three arguments reveal what Ryle viewed as the characteristic category-mistake of the official doctrine: the framing of "the differences between the physical and the mental (…) inside the common framework of the categories of 'thing', 'stuff', 'attribute', 'state', 'process', 'change', 'cause' and 'effect'" (Ryle, 1949/2009, p. 9). Instead, these three arguments provide support for the idea that the main or primary function of our folk-psychological mindgames is not descriptive, but evaluative and regulative, i.e., it's not to describe each other's doings or some inner or outer relevant fact that may cause them, but to assess their correctness or incorrectness, their norm-conforming or norm-deviant character. In this sense, we've claimed that folk-psychological interpretation is not primarily exercised from a nomological stance, defined by the adoption of a subpersonal or objectifying view of someone's actions and reactions, and whose main purpose is to predict and control them; instead, it's primarily exercised from an agential stance, characterized by the adoption of a personal or humanizing

view of our doings in order to rationalize them, making them intelligible or understandable, and thus evaluate their epistemic or moral merits or demerits.

Finally, we've seen that the pragmatist non-descriptivist framework advocated for here allows us to see that, from these observations, it doesn't necessarily follow that mental-state ascriptions are necessarily false or lack truth-conditions; this is only the case when we make "the decisive movement in the conjuring trick" and assume that either mental-state ascriptions describe some given state of affairs that may be subject to some sort of empirical scrutiny, or they are necessarily false or plainly senseless, i.e., when we remain committed to descriptivism in its deep version. By contrast, the pragmatist kind of non-descriptivism that can be read off Ryle's and Wittgenstein's work shows the absurd and, above all, unnecessary character of this assumption: to remain faithful to naturalism, all we need to endorse is the idea that the truth-evaluability of mental-state ascriptions is not given by their representing some specific state of affairs, but by the different criteria that competent language speakers use when ascribing mental states or assessing their truth or falsity in actual cases. In this sense, we've seen that a characteristic feature of our pragmatist non-descriptivism is that it rejects a monolithic and static view of the norms at play in folk-psychological interpretation; against the idea that there's some fixed, golden rule that determines in all and every possible case what counts as "believing", "desiring", "intending" and so on, competent speakers follow myriad different rules to assess the truth or falsity of different belief ascriptions in different cases. Moreover, what exact rules are followed will vary across contexts of ascription; in some settings, given certain salient features of the context, we might privilege the person's sincerity in their mental state self-ascriptions; in others, we might give more weight to their overall behavioral consistency or their predictability; and yet in others we might privilege other contextual features (Almagro-Holgado, 2021; Villanueva, 2014).

It's been a while now since we left aside the harsh debates among competing therapeutic models (Chapter 1) and we began our exploration of the different philosophies of mind underlying them and their more or less implicit Cartesian commitments (Chapter 2). After discussing the conceptual perils of the reductivist and eliminativist or incompatibilist tendencies that predate these discussions, as well as their common root in the dogma of descriptivism (Chapter 3), we've now seen how Wittgenstein's and Ryle's work, offer a sounder, non-descriptivist, framework; one which affords a non-reductivist, yet compatibilist naturalist approach to the relation between mind, normativity, and nature. Now it's time to see what the main payoffs of this approach are in the field of mental health. We won't provide a full-fledged account of this "philosophy of mental health without mirrors", as we might call it, nor all its possible implications for the four overarching conceptual problems

that we saw in Chapter 1, i.e., the analogy, boundary, priority, and integration problems - although we'll outline some of these at the end of this dissertation (see Chapter 9). Instead, now we'll mainly focus on the particular contributions that our non-descriptivist approach can make to a debate that has occupied mental health philosophy and research during the last 20 years: the debate on the doxastic status of delusions and its implications for their assessment and treatment.

PART II
Non-descriptivism and the intervention with people with delusions

Chapter 5
Believe it or not: Delusions and the typology problem

Consider the following two cases, where the experiences of two people from a non-clinical population are described[43,44]:

> RED, THE HALFWAY "LGBT+ ALLY": Red, a young cis heterosexual man in his twenties does not consider himself to be LGBT+phobic. In fact, when asked directly about this and other related issues, he overtly (and covertly) asserts things like "I believe that the LGBT+ movement deserves our full support", "We should definitely celebrate sexual-affective diversity", or "Trans women are women and trans men are men". Far from just parroting these and other pro-LGBT+ claims and slogans, Red's behavior is often in line with his self-professed support for the LGBT+ movement: for example, he has attended several LGBT+ Pride parades with his friends in the last years; he has immediately defended some LGBT+ friends when verbally attacked in the street; and even his irreflective attentional behavior, interest and valuing of a speaker's discourse is completely unbiased by the knowledge of their gender identity and sexual orientation. However, Red's thoughts and behavior also show some degree of inconsistency: for example, Red has liked and shared several posts by well-known trans-excluding collectives and personalities criticizing pro-trans laws (in fact, he feels somewhat inclined to agree with these trans-excluding organizations when they claim things like "these laws pose a threat to *women*'s rights"); he has consciously and repeatedly refused to use "they" and other inclusive language pronouns with non-binary people; he has admittedly avoided going to LGBT+ Pride parades with his gay friends if he was going "alone" (that is, without any other straight friend); and he would have some serious difficulties to understand that his future

[43]We would like to thank Manuel Almagro Holgado for these two examples, which constitute an adaptation of his Cases 1 and 2 discussed in the Chapter 4 of his PhD thesis *Seeing hate from afar. The concept of affective polarization reassessed* (2021).

[44]All the examples that appear here are adapted from real cases.

non-binary offspring would decide to undertake hormonal treatment with testosterone to achieve a more androgynous look.

BLUE, THE RACIALIZED CLINICAL PSYCHOLOGIST: Blue is a young clinical psychologist of Brazilian origins in her early thirties. After more than a decade living in Spain, her ability to speak Spanish verges on that of native speakers. However, Blue has recently found some trouble with one of her clients in the clinic where she works at, which seem to be related to her ability to speak Spanish. At the beginning of therapy, Blue described this client to a friend as "a lovely old Catholic devote", and commented on how well the therapy had started. However, after several sessions, Blue has become unable to shake the conviction that her client is displaying a racist attitude towards her. Specifically, Blue feels that her client is not taking her advice seriously due to her origins. In conversation with another friend, she confesses that she is unsure as to why exactly she thinks so; all she can tell is that, during their first sessions, her client commented a couple of times that he sometimes found some problems understanding and thus following Blue's therapeutic instructions and recommendations. Blue also feels that whenever she doesn't understand a certain word or expression, the client raises the tone a little bit and repeats the word in a somewhat condescending manner. She's aware that there are a number of competing reasons why she might be failing to be perceived as an authority by her client; after all, she does have some problems understanding her client's jargon and making herself understood, she's still a rookie in terms of clinical practice, and she is much younger than her client -indeed, she is a young *woman*. Moreover, in situations where she has come to believe that someone was being racist towards her, Blue has typically despised that person, refused to maintain any kind of contact and even publicly shamed that person, no matter who they were; on the contrary, she now still feels inclined to describe her client as a kind, charming, and gentle old man; the thought of referring her client to another psychologist does not even cross her mind, and she even feels a bit embarrassed for having such thoughts towards her client.

These two cases involve people who assert or claim to believe a certain content. For the sake of clarity, take (1) and (2) to be Red's and Blue's belief claims, respectively:

(1) I believe that the LGBT+ movement deserves our full support
(2) I believe that my client is displaying a racist attitude towards me

However, despite they sincerely assert (1) and (2), they behave, cognize and feel in ways that seem somewhat inconsistent with what one would expect of them. Thus, someone might question whether Red and Blue *really* believe the contents they claim to believe; or, in other words, whether they can be *truthfully* ascribed the belief that they self-ascribe. We can easily

think of situations where it might be of a great practical importance to ask and answer this kind of questions: in Red's case, the question as to whether he really believes that "the LGBT+ movement deserves our full support" would be vital for LGBT+ people to evaluate whether Red deserves to be considered a political ally; in Blue's case, whether she can be truthfully ascribed the belief that her client is being a racist towards her or not might be critical to determine if she should seriously consider the possibility of referring her client to another psychologist. Considering the two examples, we think that a general inclination would be to provide a negative answer in Red's case and a positive one in Blue's case; or, at least, that we would feel more inclined to question Red's status as a believer of the content he claims to believe than we would be in Blue's case.

Now consider the following case, where the delusional experience of a person with a clinical condition is depicted:

> GREEN, THE PART-TIME KARMIC TRASH COLLECTOR: Green is an 18-year-old boy who got out of a romantic relationship some months ago. Soon after that, he started a new relationship with a different person. Although Green was feeling much happier at the beginning, he eventually developed an overwhelming feeling of guilt and shame for the way in which things ended up with his former couple. Until now, Green had proudly identified himself as an agnostic regarding religious matters, a comfortable middle-ground position for him given his mother's Christian beliefs and father's mixed Christian-Buddhist creed, on the one hand, and his steadfast atheist friendships, on the other. In fact, he has been normally inclined to endorse quite die-hard physicalist and nihilistic positions when discussing metaphysical, ethical, and epistemological issues with his family and friends. However, due in part to his strong feelings of guilt, Green has now come to be convinced that he is just about to be severely punished by some supernatural and omniscient "Karmic force", as he describes it. Specifically, he cannot get rid of the idea that his new partner is in danger, due to his past actions, and claims that he now "owes something to the Universe" in return for keeping his new partner safe. The "tasks" he says that he has been commanded to do by this punitive and vengeful Karmic force are the following: first, he must order the objects in the shelf of his room in a "rectangular fashion" (i.e., he must strive to keep 90º angles between the different objects); second, he must pick every single cigarette bump and piece of trash that he sees in the street and throw them to a garbage container. Strangely (and luckily) enough though, he does not seem to be fully committed to such tasks. For example, he just "remembers" to collect cigarette bumps and trash from the street on Friday and Saturday nights, as well as on some random Thursdays and Sundays; he barely takes some time to tidy up his room, much less to properly order the objects on his shelf in an exact rectangular manner; and even when he does comply with his Karmic duties, he sometimes "cheats" (for example, he pretends not to

> have seen a particular cigarette bump or piece of garbage, or he only orders half of the shelf and promises himself to finish later, yet knowing that he will later be unable to do so or that he will just forget about it). Moreover, he has recently started to throw his own cigarette bumps and garbage to the street floor, partly due to a somewhat rebellious attitude against the mean Karmic force controlling him, but also to a still blurry and vague recognition that such force might in fact not exist.

Now take (3) to be Green's belief claim:

(3) I believe that if I don't comply with my Karmic duties, my partner will be attacked.

Is Green's example more similar to Red's case or to Blue's case? Does Green really believe that he must collect every cigarette bump and piece of trash on the street or otherwise a Karmic force will eventually let his partner be attacked? In other words, can Green's delusion be appropriately conceptualized in terms of belief? And what implications might this have for the understanding, evaluation and clinical treatment of Green's experience? The present chapter will mainly delve with this kind of questions. Regarding the first one, we think that the most obvious and immediate answer is that Green's case is more similar to Blue's case; most people would be straightforwardly inclined to say that Green surely believes what he claims to believe (see Rose et al., 2014, for evidence in this direction). Indeed, we think that, contrary to Red and Blue's cases, many people would find it utterly strange to even wonder whether Green really believes in such bizarre causal link between his rectangular shelf organizing and trash collection duties and the probability that her partner is attacked. Many would take it to be unnatural to ask such question; why on earth would he say that kind of things if he did not really believe them?

In fact, the most common and straightforward way to classify and describe delusions in clinical practice has been to understand them in terms of *irrational* beliefs. However, as we'll see, this intuitive or straightforward understanding of delusions has not remained unchallenged, and many authors have pointed out that this kind of questions, however strange they might seem, are important and informative regarding the establishing of a proper scientific, clinical and ethical approach to the intervention with people with delusions.

This and the following chapters will delve into what López-Silva (2018) has called *the typology problem*, i.e., "the problem about the *specific type of mental state* that grounds a delusional report" (2018, p. 204), exploring whether conceptualizing delusions as beliefs has any practical implications for the intervention with people with delusions. In this chapter, our main goal will be to introduce the different positions in this debate. In section 5.1., we'll

begin by presenting the standard conceptualization of delusions as irrational or strange beliefs -which is why this position has been dubbed *standard doxasticism*. In addition, we'll present the main criticisms against this standard approach, which thus go under the name of *antidoxasticism*. We'll focus on two kinds of objections to doxasticism: the objection from interpretivism, based on what we'll call the Rationality Constraint argument (hence RC) and the objection from functionalism, based on what we'll call the Stereotypical Causal Role argument (hence SCR). According to the former, delusions are not beliefs because people with delusions often fail to conform to certain rationality criteria. According to the latter, delusions are not beliefs because delusional states often fail to display belief-like causal roles; thus, people with delusions cannot be considered to actually believe in the content of their claims.

In section 5.2., we'll consider two different kinds of responses to these objections. On the one hand, *revisionist* defenses of doxasticism aim to preserve our default conception of delusions in terms of beliefs by means of revisiting the fundamental assumptions of interpretivism (see Chapter 4, section 4.2.4.), and functionalism (see Chapter 2, section 2.2.2.1.). Specifically, we'll focus on the description of two varieties of revisionist doxasticism: Bortolotti's (2010) *modest doxasticism*, which entails an interpretivist defense of doxasticism, and Bayne & Pacherie's (2005) *dispositionalist* defense, which entails a functionalist defense of doxasticism. As we'll see, while the former puts a greater emphasis on the articulation of a response to RC, the latter puts a greater emphasis on the rejection of SCP. On the other hand, we'll also consider a *non-revisionist* response to the antidoxasticist arguments, recently exemplified by Clutton's (2018) cognitive phenomenological defense of scientific doxasticism, which rejects both RC and SCP via the rejection of the functionalist and interpretivist frameworks altogether. At the end of this section, we'll introduce the two desiderata that pro-doxastic approaches aim to retain: a) a scientific desideratum, related to the claim that doxasticism leaves us in a better position to account for delusions in a scientific and clinically-informative way; and b) an ethico-political desideratum, related to the claim that doxasticism is better equipped to inform judgements about the agential status of people with delusions and thus provides a further barrier against abusive or unjust treatment.

Finally, in section 5.3., we'll resume the contents of this chapter and establish the guiding questions for the following one.

5.1. The typology problem

The conceptualization of delusions in *doxastic* terms (i.e., relating to an agent's beliefs) is commonplace in scientific and clinical literature. In this sense, it is almost preceptive to

begin our discussion by making reference to the definition of delusions according to traditional diagnostic manuals. The DSM-V (APA, 2013, p. 87) defines delusions as:

> [...] fixed beliefs that are not amenable to change in light of conflicting evidence. Delusions are deemed bizarre if they are clearly implausible and not understandable to same-culture peers and do not derive from ordinary life experiences. [...] The distinction between a delusion and a strongly held idea is sometimes difficult to make and depends in part on the degree of conviction with which the belief is held despite clear or reasonable contradictory evidence regarding its veracity. (APA, 2013, p. 87)

However, this *standard doxasticist approach* (see Frankish, 2009, p. 269) has long been challenged within philosophical discussion on the grounds that delusional states do not properly meet the necessary criteria that certain theories of belief establish for true or appropriate belief ascription. Roughly, the idea is that many people with delusions often fail to reason, act, or react as we would expect of some who believed what they claim to believe (see Berrios, 1991; Currie, 2000; Currie & Jureidini, 2001; Egan, 2008; Frankish, 2009, 2012; Graham, 2010a; Hamilton, 2007; Hohwy & Rajan, 2012; Murphy, 2012; Radden, 2010, 2013; Sass, 1994; Schwitzgebel, 2012; Stephens & Graham 2007; Tumulty, 2011, 2012; Young, 1999; see also Bayne & Pacherie, 2005; Bortolotti, 2010). One of the most discussed examples is that of people with Capgras delusion (where the person asserts that a close one -a partner, a relative, etc.- has been replaced by an identically looking impostor). However, many people with Capgras delusion fail to act and reason on the grounds of what they claim to believe: for instance, they continue to live and engage with the impostor as they did with the replaced loved one, often not even trying to search for the latter, or they fail to provide sufficient reasons for their claim, or to solve certain contradictions (e.g., they don't offer any "excuse" as to why the impostor knows every single detail of their relationship with the replaced loved one) (see Coltheart et al., 2011).

Specifically, the debate has been mainly framed by two inter-related theoretical frameworks: interpretivism and functionalism, which we've mentioned in Chapters 2 (section 2.2.2.1.) and 4 (section 4.2.4.). Thus, before analyzing the argumentative structure of the antidoxasticist criticisms that gave rise to the typology problem, let's first review the main features of these two theories of belief.

5.1.1. Interpretivism and functionalism: Two theories of belief?

On the one hand, according to interpretivist approaches, the concept of belief is primarily individuated in terms of rationality rules or criteria (Davidson, 1986; Dennett, 1979/1987; see

also Byrne, 1998). According to Bortolotti's (2010) construal of these *rationality constraints*, a belief can be ascribed to an agent if and only if the agent's intentional state is: a) contentful (i.e., evaluable in terms of its truth or falsity); b) epistemically rational (i.e., it must be grounded on sufficient evidence and responsive to counter-evidence); c) procedurally rational (i.e., it must be well integrated in the agent's belief system and hold some appropriate inferential relations with the agent's other mental states); and d) agentially rational (i.e., it must be action-guiding and its endorsement must be grounded on intersubjective good reasons) (see also Bayne & Pacherie, 2005; Bortolotti & Miyazono, 2014; Clutton, 2018; Lopez-Silva, 2018; Miyazono, 2019).

On the other hand, according to functionalist approaches (see Chapter 2, section 2.2.2.1.), the concept of belief is primarily individuated in terms of its stereotypical *causal* roles (Block & Fodor, 1972; Lewis, 1966, 1980; Putnam, 1967/1975); beliefs, understood by functionalists as causal devices, are no different from computational states, which are typically defined by their inputs and their outputs. In this sense, beliefs and other mental phenomena aim to fill the alleged "explanatory gap" between perception and action, perception and other mental states, or between other mental states and action. For functionalists, a certain mental state is a belief if and only if it displays certain belief-like causal roles (i.e., if it displays a particular functional profile). Following the classical distinction, this minimal functionalist stance can be fleshed out either in occurrentist (e.g., Carruthers, 2013) or dispositionalist terms (e.g., Schwitzgebel, 2013) (see Chapters 2 and 4, sections 2.1. and 4.2.1.; see also Nottelman, 2013). According to *occurrentism* -or *standard representationalism*, to use Miyazono and Bortolotti's (2014) term, "beliefs are occurrences (e.g., phenomenological states or distinctive activations of the cognitive system)" (Nottelman, 2013, p. 23); thus, "to believe is to have a representation that plays belief roles" (Miyazono and Bortolotti, 2014, p. 32). On the other hand, according to *dispositionalist* accounts of belief (Nottelmann, 2013; Schwitzgebel, 2002, 2012, 2013), beliefs are dispositions to behave, cognize, or experience in certain ways. Belief ascriptions here play a similar role as terms like "soluble" or "flammable" for describing sugar or gasoline; though they don't describe any further facts, they are useful for us when causally explaining and predicting how will sugar or gasoline react when in contact with water or fire, respectively. Finally, contra dispositional realists (see Molnar & Mumford, 2003), the kind of dispositionalism that we'll mainly focus on here takes it that all there is for an agent to be truthfully ascribed a certain belief is that they systematically display the appropriate behavioral, cognitive, and phenomenological patterns whenever certain triggering conditions are met (e.g., Schwitzgebel, 2002, 2013).

In a sense, functionalism and interpretivism can be understood as non-identical twin theories of belief. To begin with, they both share an ontologically noncommittal approach to belief ascription, since neither is -at least necessarily- committed to any particular ontological stance with regard to the nature of beliefs; they avoid such ontological commitment by equating the possibility for an interpreter to truthfully ascribe the belief that *p* to an agent (on the grounds of the agent's behavior, cognition, and phenomenology meeting certain criteria) with the agent's *having* the belief that *p* (or *being in* the mental state of believing that *p*) (see Byrne, 1998).

In addition, both are grounded on a similar *mindreading* view of folk psychology (see Chapters 2 and 3, sections 2.2.1. and 3.1.1.), which takes it that mental-state ascriptions subserve some kind of pre-scientific, causal-explanatory function (Almagro & Fernández-Castro, 2019; Fernández-Castro, 2017a, 2017b; Fernández-Castro & Heras-Escribano, 2019; McGeer, 2007, 2015, 2021; Zawidzki, 2008). Indeed, functionalism could also be read as adding a further conceptual commitment to the interpretivist proposal: the nomological construal of the interpretivists' rationality constraints. Thus, an agent's intentional state can be appropriately or truthfully conceived of as a belief if and only if it is in an *appropriate* (that is, rationally understandable) *causal* relation with the available evidence, with other beliefs and mental states, and with the agent's subsequent actions and reason-giving (see Bayne & Hattiangadi, 2013). In turn, some interpretivist approaches conceive of mentalistic or intentional explanations as a certain kind of nomological or causal-predictive tool, one reserved to those creatures that we deem to be rational. According to Dennett (1979/1987), adding to the *physical stance* and the *design stance* -which we adopt when we explain an agent's behavior in purely physical or biological terms, respectively- we can adopt what the author calls the *intentional stance* towards the explanation of the behavior of rational creatures. When we causally explain and predict an agent's behavior in mentalistic terms (i.e., by attributing beliefs, desires and intentions to it), we are taking such an intentional stance. Antidoxasticism, as we'll now see, has built up on these two inter-related theories of belief.

5.1.2. Antidoxasticisms

Drawing from interpretivism and functionalism, many have contested the standard conception of delusions as beliefs (Berrios, 1991; Currie, 2000; Currie & Jureidini, 2001; Egan, 2008; Frankish, 2009, 2012; Graham, 2010a; Hamilton, 2007; Hohwy & Rajan, 2012; Murphy, 2012; Radden, 2010, 2013; Sass, 1994; Schwitzgebel, 2012; Stephens & Graham 2007; Tumulty, 2011, 2012; Young, 1999). The most discussed examples in the literature concern what the usual classification proposals label as monothematic delusions, i.e., a single belief-like state or small set of belief-like states that are held towards a single theme. Specifically, cases of

Capgras delusion (where the person asserts that a close person has been replaced by an identical impostor), Cotard delusion (where the person asserts that they are dead or disembodied) and mirrored-self misidentification (where the person does not identify themselves in the mirror and thinks instead that the person in the mirror is a stranger), are among the most widely discussed cases. On the contrary, polythematic delusions, such as the ones exhibited by people diagnosed with schizophrenia and other severe mental disorders, have not been so widely discussed, although it is a common assumption that a defense of doxasticism towards monothematic delusions will provide enough grounds for a defense of a similar account regarding polythematic ones (see Coltheart et al., 2011).

The reason why monothematic delusions have been more often analyzed is that, according to antidoxasticists, these are the ones that pose a major difficulty for doxasticism, since they cannot be easily accommodated within the functionalist or interpretivist theories of belief. Probably due to their conceptual resemblance, it is not always straightforwardly clear what precise theoretical framework supports the argumentative structure of many antidoxasticist proposals. Nonetheless, since these two theories of belief are not exactly identical, neither are the antidoxasticist arguments that might be drawn from each one respectively. For the sake of clarity, let's thus treat both arguments separately.

5.1.2.1. Delusions: Rationality outages or computer breakdowns?

On the one hand, according to those drawing from interpretivism, the patterns of actions and reactions of people with delusions usually fail to meet the above-mentioned rationality constraints. Therefore, delusions do not qualify as beliefs. Henceforth, we will refer to this first kind of objection as the objection from the Rational Constraint argument (RC) (see Bortolotti, 2010, 2012), whose argumentative structure could be construed as follows:

Rational Constraint argument

Premise 1: A can be truthfully ascribed the belief that p iff A's mental state meets certain rationality constraints (e.g., it is contentful, and procedurally, epistemically and agentially rational).

Premise 2: Delusional cases (at least many of them) fail to meet either one or all of these rationality constraints.

Conclusion: Thus, delusions (or at least many of them) do not count as beliefs.

On the other hand, according to *functionalist antidoxasticists*, at least many people with delusions systematically fail to display the behavioral, cognitive, or phenomenological patterns that one would expect if they really believed the content of their delusional

statements; in other words: delusions don't properly fit the stereotypical causal roles of belief. Thus, doxasticism about delusions is misplaced. Henceforth, we will refer to this second kind of objection as the objection from the Stereotypical Causal Profile argument (SCP) (see Miyazono & Bortolotti, 2014; see also Miyazono, 2019), whose argumentative structure could be stated as follows:

Stereotypical Causal Role argument

Premise 1: A can be truthfully ascribed the belief that p iff, A's mental state has certain belief-like causal roles (i.e., if certain contextual conditions are met, certain behavioral, cognitive or phenomenological patterns are observed).

Premise 2: Delusions (at least many of them) fail to play belief-like causal roles.

Conclusion: Thus, delusions (or at least many of them) do not count as beliefs.

Antidoxasticist approaches to delusions are thus defined in terms of a negative thesis: they deny a belief status to delusional phenomena on the grounds that either a) people with delusions do not behave, cognize or experience as it would be rational to expect of them if they really believed the delusional content; or b) delusional states do not display stereotypical belief-like causal roles. For some, the content of some delusional statements is itself bizarre enough to preclude an interpretation in literal terms (e.g., "My legs are stretching and shrinking at the same time"). This led Jaspers (1913/1963) to declare them as empathically "ununderstandable", in the sense that while they might be explainable in causal terms, they are nonetheless unintelligible from a rational point of view (see also Fulford & Thornton, 2017). Although the issue of content is indeed an important one, we'll leave it aside here to focus on the problems of epistemic, procedural, and agential irrationality, in the interpretivist terms, or the problem of the deviation from the causal stereotypical profile, in the functionalist jargon.

Along these lines, many authors have proposed that an important part of the problematic nature of delusions lies in the *bottom-up* or *input* side of their rational or causal story, i.e., in their being apparently formed on an insufficient or utterly bizarre evidential basis (Currie, 2000, Currie & Jureidini, 2001; see also Bortolotti, 2010; Fulford & Thornton, 2017). Claiming to know that one's parents have been replaced by aliens because they don't pay as much attention to one as they used to, or claiming that one's old client is a racist because of his condescending attitude towards a young person could be good examples of this.

However, some authors have pointed out that an anomalous evidential basis or causal input does not pose a serious threat to the conceptualization of delusions in doxastic terms

(e.g., Bayne & Pacherie, 2005; Bortolotti, 2010; Schwitzgebel, 2012; Wilkinson, 2013); after all, an insufficient or inappropriate source of evidence might be a problem for being attributed with *knowledge* of a certain state of affairs[45], but not for being attributed the belief that such state of affairs is the case. Thus, the problematic aspect of delusions seems to be mostly due to their their *top-down* or *output* features. As Schwitzgebel has put it:

> Beliefs can arise in any old weird way, but—if they are to be beliefs—they cannot have just any old effects. They must have, broadly speaking, belief-like effects; the person in that state must be disposed to act and react, to behave, to feel, and to cognize in the way characteristic of a normal believer-that-P. (Schwitzgebel, 2012, p. 14)

Specifically, antidoxasticists have focused on what has been called the *bad integration* and the *double-bookkeeping* objections to doxasticism about delusions (e.g., Bortolotti, 2010, 2011; Sass, 1994, 2014; Gallagher, 2009; Porcher, 2019; see also Bortolotti, 2018). On the one hand, bad integration refers to the commitment of "obvious mistakes in deductive reasoning, or fail[ing] to obey basic inferential rules governing the relations among beliefs and other intentional states" (Bortolotti, 2010, p. 62); i.e., displaying attitude-attitude inconsistencies. It thus compromises the assumption of procedural rationality or of the appropriate causal relations among beliefs and other intentional states. Claiming to believe that the vengeful Karmic force dictating what one should do is omniscient and then thinking that one can "cheat on" it to avoid one's Karmic duties, or claiming to believe that "trans women are women" and at the same time believing that "the Government's new 'trans law' might pose a threat to women's rights" would be good examples of badly integrated beliefs.

On the other hand, double-bookkeeping is a phenomenon where a certain agent, despite claiming to believe that *p*, behaves in ways inconsistent with what they claim to believe; in other words, the agent displays attitude-behavior inconsistencies[46]. This thus compromises the assumption of agential rationality or of the appropriate causal relations

[45] However, see Srinivasan (2020) for an explanation of why this argument does not apply either even with regard to certain knowledge claims (e.g., when the knower has been systematically exposed to situations of injustice and is thus reasonable to attribute them with the capacity to automatically detect discriminative behaviors towards them).

[46] The notion of double-bookkeeping might also be used for cases of attitude-attitude inconsistencies (see Bortolotti, 2010, p. 161-162). In fact, from our non-descriptivist approach to mental-state ascriptions, these would be very close concepts: bad integration would point to conflicts between the conceptual commitments acquired through two explicit mental-state ascriptions, whereas double-bookkeeping would point to conflicts between the conceptual commitments acquired through explicit mental-state ascriptions and those which one's actions and reactions seem to be conforming to. However, for the sake of clarity, we will here distinguish both cases.

between belief and action. Claiming to believe that one must collect every piece of garbage that one finds in the street to prevent one's partner for being attacked and then doing so only on some Friday and Saturday nights, or claiming to believe that LGBT+ people deserve our full support and then sharing trans-excluding content on the social media would be good examples of double-bookkeeping.

5.1.2.2. Not beliefs... but what then?

Antidoxasticist approaches to delusions thus converge on their defense of some variety of the negative thesis, which denies a belief status to delusions on the grounds of RC or SCP. What these approaches differ on is on the kind of positive thesis they advance with regard to the nature of delusions. Drawing from the issue of the bizarreness of some delusional statements, some authors have proposed non-assertoric approaches to delusions. For example, Berrios (1991) has defended the radical thesis that delusional statements are contentless (they are "empty speech acts", in his own terms). On a milder version of the non-assertoric approach, Sass (1994, 2004; see also Sass & Pienko, 2013) proposes to understand delusions in rather metaphorical terms; delusional statements are contentful, but their content is not to be determined by a literal interpretation of the statement. In line with Sass's phenomenological approach, other authors have criticized standard doxasticism on the grounds that delusions, more than doxastic deviances, are best characterized in terms of their specific experiential properties (e.g., Radden, 2013; Hohwy & Rajan, 2012). In a stronger version of this "experiential" variety of antidoxasticism, delusions would involve a whole experiential reality shift: they would constitute "alternative realities", in Gallagher's (2009) terms.

Although Sass's and Gallagher's phenomenological perspective introduces some interesting possibilities, we won't delve into a detailed discussion of it here. The main reason is that these approaches might not be properly characterized as "antidoxastic"; after all, claiming that some delusions involve certain alterations of the structure of experience is not incompatible with claiming that they are beliefs, taken as endorsements of such experiences (Sass, 2004, p. 77; see also Bayne & Pacherie, 2004b). Thus, we'll mainly focus here on those kinds of antidoxasticism which attempt to understand delusions in terms of other kinds of mental states different from belief. Among these, two main strands of antidoxasticism might be distinguished: a) *commonsensical antidoxasticism*, which rejects doxasticism about delusions, but not their interpretation in folk-psychological terms; and b) *non-commonsensical antidoxasticism*, which rejects both doxasticism and the folk-psychological conceptual framework altogether.

On the one hand, commonsensical antidoxasticists defend that delusions can be explained in terms of *propositional attitudes other than belief*. This kind of antidoxasticism thus

follows what we'll call a *reclassification* strategy (see Bayne & Hattiangadi, 2013), which rejects doxasticism in particular, but still assumes that delusions can be conceptualized in folk-psychological terms. One of the most renown examples of this approach is the meta-cognitive approach defended by Currie and collaborators (Currie, 2000; Currie & Jureidini, 2001), which states that delusions are not beliefs, but *imaginings*; specifically, imaginings that the person mistakes for beliefs. According to the author, the fact that many delusions are not properly acted upon can be rightfully accommodated if we think of delusions in terms of imaginings. When we imagine, instead of believing, that a certain state of affairs is the case, we might take the imagined content into action or not. If we just imagine that there is a Karmic force that might punish us if we don't comply with a series of obligations, we might sometimes act, reason or feel in accordance with the imagined content, but we're no longer rationally expected to do so. We might even be carried away by our imaginings, especially when they carry such an aversive content as Green's thoughts; however, we're no longer compelled to act in accordance with the imagined content, as we would be if we claimed to believe that such and such state of affairs is the case.

One major problem of this account is that it doesn't take into account that many people with delusions in fact display belief-like behaviors. For example, many people with delusions try back up their delusions with reasons, and they cannot but feel convinced and try to convince others about the truth of their thoughts. This feature is at odds with a purely imaginative account of delusions (Bayne & Pacherie, 2005; Bortolotti, 2010; Radden, 2013). Furthermore, even if they didn't, it wouldn't be clear whether the meta-cognitive approach solves the problems it aims to solve. On this account, people with delusions believe that they believe the delusional content (hence the "meta-cognitive" character of delusions), but in fact they just imagine it. Yet one might then wonder: why don't they act in accordance with what they believe to believe? Why wouldn't they be rationally compelled to act in accordance with what they believe to believe?

Alternatively, one might think, with Murphy (2012), that "[d]elusions are attributed [...] when we run out of the explanatory resources provided to us by our folk understandings of how the mind works" (p. 22), and thus reject their framing within the usual categories of folk psychology. Non- commonsensical antidoxasticists assume that folk psychology offers, at best, quite poor explanatory resources for understanding delusions. What if they don't fit our preferred accounts of belief -or imagination, for that matter? Why should our understanding of delusions be constrained by folk-psychological assumptions?

One possible way to implement this kind of non-folk antidoxasticism is to follow what we could refer to as a *rebranding* strategy; if folk-psychological resources are not good

enough for a proper scientific account for delusions, let's then craft a new, *sui generis* type for them (see Bayne & Hattiangadi, 2013). Along these lines, some authors have proposed to regard delusions, together with other strangely-behaved intentional states, as hybrid or 'in-between' states. For example, Egan (2008) has proposed to understand delusions as *bimaginations*, i.e., mental states that share the functional profile of both beliefs and imaginations, but do not fit any of them fully. Egan's (2008) "bimaginations" are precisely supposed to account for the observed deviances from (or only partial compliance with) both belief-like and imagination-like causal stereotypes. On this kind of rebranding response, scientific research is released from its folk-psychological grips; are delusions best understood as cases of "bimaginations"? Alright, let's then investigate what could be the biological bases of such hybrid states.

A yet more radical proposal can be drawn from Schwitzgebel's (2002, 2012, 2013) dispositionalism[47]. Schwitzgebel (2012) claims that in many delusional cases the person does not fully believe the delusional proposition, but instead just "fuzzy believes" it; delusions, thus, are best conceived of as "beliefs gone half-mad" (p. 13). In Schwitzgebel's account, delusions are not 'in-between' states in the sense that they are instances of some hybrid, middle-ground new mental type; what is 'in-betweenish' here is the truth of our folk-psychological belief ascriptions. In cases where the agent's dispositional profile does not fully meet the folk-psychological stereotypes that we associate with a certain belief, it will not be fully correct (nor fully incorrect) to describe someone as a proper believer of such content. But that shouldn't worry us, nor scientists for that matter. Instead of crafting a new mental type for delusions or similar quasi-doxastic phenomena, we might adopt some kind of *local eliminativist* perspective: if the behavioral, cognitive and phenomenological activity of people with delusions cannot be properly (i.e., truthfully) characterized in terms of beliefs, then let's just exhaustively specify their dispositional layout. We might choose to coin a new name for it or not, but that's inessential; once this dispositional profile has been fully specified, scientists are left free for determining its natural causes.

Non-commonsensical antidoxasticist proposals thus seem to draw from a similar argument to that which underlies non-reductive incompatibilism about the mind, i.e., that there's no principled reason why folk psychology should constrain scientific research about a particular set of behaviors, cognitions, or experiences (see Chapter 3, sections 3.1.3., 3.2.1.). In this sense, Schwitzgebel's approach would be the most radical; after all, Egan's

[47] In section 5.2.1.2. we'll come back to Schwitzgebel's dispositionalist proposal. As we'll see, although Schwitzgebel (2012) himself later endorsed an antidoxasticist approach to delusions, it was initially used by Bayne & Pacherie (2005) to defend doxasticism.

"bimaginations" are still partially framed by our folk-psychological understanding of what imaginations and beliefs are. But why should scientific theories retain any such residual commitment to a commonsensical view of delusions? Instead, scientists and clinicians might just specify the dispositional profile of each particular case, or alternatively attempt to describe the statistically typical dispositional profile of a given group of people. The latter option would be appropriate for a scientific approach that aimed at specifying common causal factors responsible for similar delusional cases, while the former would be the kind of scientific approach favored by individual-centered approaches, such as behavior analysis and its clinical application in Functional Behavioral Assessment-based interventions (see Chapter 8).

As we will see in the upcoming chapters, we are sympathetic to this kind of argument: scientists and clinicians should be let free to determine, by empirical methods, which are the natural causes of whatever pattern of interest, regardless of whether it fits our folk-psychological categories or not. Notwithstanding this consideration, however, we'll claim that there are still conceptual and pragmatic reasons to defend doxasticism, once the doxasticist understanding of delusions is understood along the lines of the non-descriptivist approach to the mind (see Chapter 6, section 6.3.). In the next section, we'll introduce the main pro-doxasticist contenders in the typology problem arena.

5.2. Doxasticisms

Several pro-doxasticist approaches have been developed in the last two decades to account for the challenges posed by antidoxasticism towards delusions. We will here distinguish between two major strands: a) *revisionist doxasticism*, which assumes that functionalism and interpretivism still grant a doxastic status to delusions when certain considerations are taken into account; and b) *non-revisionist doxasticism*, which assumes that functionalism and interpretivism are inadequate theories of belief and should thus be replaced by a different theoretical framework. Regarding revisionist approaches, we'll focus on two of the most widely discussed defenses of doxasticism about delusions: Bortolotti's (2010, 2011, 2012; Bortolotti & Miyazono, 2014) modest doxasticism, which entails an interpretivist defense of doxasticism, and Bayne & Pacherie's (2005; see also Bayne & Pacherie, 2004a, 2004b) dispositionalist defense, which constitutes a functionalist defense of doxasticism. Regarding non-revisionist approaches, we'll focus on Clutton's (2018) more recent defense of *scientific doxasticism*, which rejects functionalism and interpretivism and adopts instead a *cognitive-phenomenological theory of belief*. Finally, we'll present what we view as the core desiderata behind these proposals.

5.2.1. Revisionist doxasticism

Revisionist doxasticisms don't radically question the theoretical framework behind antidoxasticism -namely, interpretivism or functionalism; instead, they examine to what extent these two theories of belief really motivate antidoxasticism towards delusions and, to the extent that they do, revisionist doxasticisms recommend local revisions of the theoretical background. Although the two kinds of revisionist doxasticism that we will review here share many of the arguments in favor of doxasticism, we'll consider them separately depending on whether they put a greater emphasis on the articulation of a response to RC or to SCP.

5.2.1.1. Lisa Bortolotti's modest doxasticism

Bortolotti's (2010, 2011, 2012, Bortolotti & Miyazono, 2014; see also Bortolotti, 2018) modest doxasticism constitutes one of the most widely discussed defenses of doxasticism. If we recall the aforementioned RC argument against doxasticism,

> *Rational Constraint argument*
>
> Premise 1: A can be truthfully ascribed the belief that p iff A's mental state meets certain rationality constraints (e.g., it is contentful, and procedurally, epistemically and agentially rational).
>
> Premise 2: Delusional cases (at least many of them) fail to meet either one or all of these rationality constraints.
>
> Conclusion: Thus, delusions (or at least many of them) do not count as beliefs.

what Bortolotti proposes is to reject both of its premises, especially the first one: not only we can question the assumption that delusions do not meet the interpretivist's rationality constraints (since some of them in fact do), but we can also question whether these rationality constraints in fact reflect how our daily belief ascription practices work. In this regard, Bortolotti's strategy is twofold: firstly, she asks the empirical question as to whether delusions effectively fail to meet the standards of epistemic, procedural and agential irrationality[48]; secondly, she asks the conceptual question as to whether interpretivism is able to accommodate our straightforward understanding of many non-clinical phenomena in doxastic terms.

Regarding premise 2, Bortolotti points out that in fact many delusional cases fit well with the criteria imposed by interpretivism for an intentional state to qualify as a belief (see

[48] Regarding the content rationality constraint, Bortolotti (2010) takes it to overlap with the epistemic and procedural rationality constraints, hence she doesn't discuss it separately (see pp. 57-58).

also Bayne & Pacherie, 2005; Reimer, 2010). Firstly, as we have already seen, many delusions might be grounded on an insufficient evidential basis, but this does not seem to preclude our interpretation of them in doxastic terms. Secondly, regarding the bad integration and double-bookkeeping objections, it seems that many people with delusions do in fact reason and act upon their delusions (see Young, 1999). Bortolotti (2010, pp. 69-70, 164-165) gathers many examples of this from the clinical literature. For example, people with Cotard delusion (i.e., the delusion that one is dead) sometimes stop eating and bathing; in addition, they sometimes justify why, despite being allegedly dead, they are able to move and talk (for example, because they already are in Heaven).

Regarding premise 1, Bortolotti claims that the fact that interpretivism leaves some delusional cases out of what is interpretable in doxastic terms is a deficit of the theory itself; since it does not account well for how folk psychologists readily interpret some clinical -as well as non-clinical- cases in terms of belief, we should disregard classical interpretivism as an oversimplistic and overidealized model of belief ascription (see also Bayne & Pacherie, 2005; Clutton, 2018; Reimer, 2010; Rose et al., 2014). The gist of Bortolotti's defense lies in the following argument: if we let the interpretivist's rationality constraints be "too constricting", we must be ready to forego a doxasticist account of a vast amount of other non-clinical phenomena that we naturally interpret in terms of irrational beliefs (such as superstitious, contradictory or poorly acted-upon beliefs, etc.). Take Red, Blue and Green's cases. According the interpretivist account -at least on an stringent reading of it (see Reimer, 2012)-, all three would fail, to the same extent, to count as believers -or, at least, as good believers- of the contents they claim to believe [(1), (2) and (3), respectively]. Thus, although interpretivism seems to accommodate our probable response to Red's case, they fail to do so with regard to Blue's case and Green's case.

Consequently, Bortolotti (2010) and other doxasticists have argued that the standards imposed by the interpretivist's rationality constraints should not be seen as *constitutive* of our belief ascription practices, at the risk of failing to include a wide range of everyday irrational beliefs; instead, they must be seen as providing some *normative* criteria that partially guide our folk-psychological belief attributions. Procedural, epistemic and agential rationality are not seen as necessary conditions for an agent to be truthfully ascribed a certain belief, but just regulative ideals that are taken into account -though not exclusively-, when deciding whether someone merits a particular belief ascription. In addition, Bortolotti emphasizes the context-relative nature of belief ascription, i.e., the fact that the truth or falsity of a given belief ascription might be affected by contextual considerations at the moment of assessment (see section 5.2.1.3.; see also Chapter 6, section 6.1.1.). As she puts it, "The way in

which interpreters ascribe beliefs changes depending on the shared environment, on the subject and on the context of interpretation. We know from our own daily practice of interpretation that there are no golden rules" (Bortolotti, 2010, p. 262).

In this sense, Bortolotti argues that one fundamental flaw of classical interpretivism is that it fails to distinguish "between two notions of rationality—rationality as conformity or subscription to epistemic norms, and intelligibility of observed behaviour" (Bortolotti, 2010, p. 99). In the former, stronger sense of the term, to be a rational agent involves reaching a certain standard: a rational agent, in this sense, is one that draws systematically correct inferences from the available evidence and from their other intentional states, that systematically acts upon their self-professed beliefs, desires and intentions, and that is systematically able to ground their judgements in intersubjectively good reasons. In the latter, weaker sense of the term, none of this is necessary: a rational (i.e., intelligible) agent is just one whose behavior can be regarded as meaningful or purposeful; that is, that their behavior can be rationalized, made intelligible by connecting their doings with some *reason*, regardless of whether such reason is intersubjectively good or not, and regardless of whether the agent displays an overall rational (in the strong sense) pattern of activity or not.

> For the purposes of establishing intelligibility and proceeding to explaining and predicting behaviour intentionally, all we need is that the subject has a reason for reporting her attitudes or acting as she does that can be cashed in intentional terms. Whether her attitudes or actions meet standards of rationality is beside the point. (Bortolotti, 2010, p. 100)

> Intelligibility is a weaker notion than rationality. I can understand (sympathise with) behaviour that I do not regard as rational. If I expect people's behaviour to be *intelligible*, what I expect from them when they report belief states is that they are in a position to ascribe these beliefs to themselves and they have some relevant reason, some reason they regard as a good reason, for endorsing the content of their belief states. If I expect people's behaviour to be *rational*, I expect more [...]: for instance, I might expect from people reporting beliefs that they have reasons in support of the content of their beliefs that are intersubjectively acknowledged as good reasons. (Bortolotti, 2010, p. 264)

Classical interpretivism takes rationality (in the first and stronger sense) as a precondition for belief ascription; on this view, an agent must be already rational (i.e., systematically meet the standards of procedural, agential and epistemical rationality) in order to be granted beliefs, desires, intentions, and other mental states. Bortolotti's modest interpretivism, on the contrary, rejects this assumption and puts interpretivism upside down: we, as

folk-psychological interpreters of ourselves and one another, rationalize (i.e., make intelligible) each other's behavior *by means* of our mental-state ascriptions; as she puts it, "to interpret behaviour is to make it intelligible, to rationalise it in the weak sense I proposed" (Bortolotti, 2010, p. 102). It is in the same practice of interpreting one another in folk-psychological terms, in rationalizing each other's behavior, that we come to view each other as intelligible beings, whose behavior can be subject to normative consideration (i.e., assessed in terms of its correctness or incorrectness). Thus, ascribing beliefs and other mental states to an agent is the precondition for assessing if their behavior is rational (i.e., if it reaches the standards of rationality, in the stronger sense) or not, and not the other way around.

Finally, taking all this into account, Bortolotti proposes a relaxation of the interpretivist's requirements for a mental state to count as a belief. Firstly, instead of being fully procedurally rational, Bortolotti proposes that beliefs must just have *some* inferential connections with other beliefs and mental states. Secondly, regarding the standard of epistemic rationality, Bortolotti holds that beliefs need not be responsive to evidence (i.e., they need not change in light of contradictory evidence); they just need to be *sensitive* to it; in other words: all it takes for an intentional state to count as a belief is that it can potentially change in light of contradictory evidence, even if it does not in many occasions. Finally, an intentional state need not be agentially rational (i.e., action-guiding, in a strong sense of the term, and endorsed on the basis of intersubjectively good reasons); for Bortolotti, an intentional state might count as a belief if it is a) *behaviorally manifestable*, i.e., it must potentially lead to action in some of the relevant circumstances; and b) endorsed on the grounds of *subjectively* good reasons or, in Bortolotti's (2010, p. 264) words, on the basis of "some reason [the agents themselves] regard as a good reason".

5.2.1.2. Tim Bayne & Elisabeth Pacherie's dispositionalist defense

Bortolotti's (2010) thorough account thus aims to provide a response to RC, via the questioning of both its conceptual and empirical premises. On the other hand, Bayne & Pacherie's (2005; see also Bayne & Pacherie, 2004a, 2004b) approach can be viewed as an attempt to reject the SCP argument against doxasticism. Let's recall its argumentative structure:

> *Stereotypical Causal Profile argument*
>
> Premise 1: A can be truthfully ascribed the belief that p iff A's mental state has certain belief-like causal roles (i.e., if certain contextual conditions are met, certain behavioral, cognitive or phenomenological patterns are observed).
>
> Premise 2: Delusions (at least many of them) fail to play belief-like causal roles.
>
> Conclusion: Thus, delusions (or at least many of them) do not count as beliefs.

Specifically, Bayne & Pacherie (2005) follow a twofold strategy that is similar to Bortolotti's (2010) one. Firstly, the authors gather similar evidence showing that many people with delusions do in fact exhibit the behavioral, cognitive and phenomenological patterns that one would expect of a normal believer of the delusional content; thus, the second premise of SCP is not as empirically warranted as functionalist antidoxasticists take it to be (see also Bayne & Pacherie, 2004a, p. 6). Secondly, Bayne & Pacherie (2004a, 2005) also revise the first conceptual premise. In particular, they add a *ceteris paribus* clause to the functionalist understanding of beliefs, i.e., an "all things being equal, standard or normal" condition: for a person to be truthfully ascribed a given belief, they must display belief-like patterns of behavior, cognition, and experience *whenever certain standard conditions are met*[49]. According to Bayne & Pacherie, many of the attitude-attitude or attitude-behavior deviancies observed in the case of people with delusions might be satisfactorily accounted for in this way; in other words: even in cases where people with delusions systematically deviate from the stereotypical causal profile of belief, they argue, these deviations can be readily *excused* by appeal to some non-standard feature of the context of belief ascription.

In fact, rather than "adding" this *ceteris paribus* clause, they re-emphasize it, for it was already contained in some functionalist approaches to belief. Specifically, Bayne & Pacherie's defense of doxasticism builds up on Schwitzgebel's (2002, 2013) dispositionalist account, which explicitly contains this clause. According to Schwitzgebel, there is nothing more to having the belief that *p* than being truthfully ascribed the belief that *p*, and the yardstick that we employ to evaluate the truth or falsity of a certain belief ascription is a particular causal stereotype; specifically, a *dispositional stereotype*: we allegedly associate each belief with a stereotypical cluster of behavioral, cognitive and phenomenological dispositions, and then use such stereotype to determine whether the behavioral, cognitive or phenomenological activity (i.e., the functional profile) of a particular agent matches it sufficiently to merit a full belief ascription. Therefore, when a person manifests all the stereotypical dispositions commonly associated in a given community with having the belief that *p*, that person will be always truthfully ascribed the belief that *p*; on the contrary, if they do not manifest any of the stereotypical dispositions, they will never be truthfully ascribed the belief that *p*.

However, there are cases (probably most of the cases), where the person does not manifest all of the relevant dispositions, and thus might be difficult to discern whether

[49] As we'll see in Chapter 6 (section 6.1.1.) Bortolotti's defense of doxasticism also involves the invocation of a *ceteris paribus* clause to secure the conceptualization of delusions as beliefs (see Bortolotti, 2012).

someone can be truthfully ascribed the belief that *p* or not. Schwitzgebel's dispositionalism is thus a variant of what has been called a *sliding scale approach* to belief ascription, where having certain belief is a matter of degree; belief here is not an on-off or binary notion, but a graded or quantitative-like one (Schwitzgebel, 2012; see also Bortolotti, 2010, 2012). To illustrate this, take Schwitzgebel's comparison between attitude ascriptions and the assessment of personality traits:

> Compare, again, to personality traits. Few of us are 100 per cent extravert or 100 per cent introvert, 100 per cent high-strung or 100 per cent mellow. Rather, we match these profiles imperfectly and more closely in some respects than in others. If we match imperfectly enough, if we are stably prone to go sometimes one way, sometimes the other, often the best plan for describing us is to weasel out of any simple, overarching attribution and instead describe our patterns of splintering dispositions. In personality, there are gray, vague, in-betweenish cases. So also when our dispositions splinter away from neat alignment into attitudinal stereotypes. (Schwitzgebel, 2013, p. 86)

In these cases, we might still ascribe the belief that *p* or not depending on different features of the context of ascription. On the one hand, deviances from the causal stereotype might be *excused* by some non-standard situation (i.e., if the *ceteris paribus* clause is cancelled). If a proper excuse is found, then we might still claim that the agent in fact has the relevant disposition, but just fails to manifest it for some particular reason. In Schwitzgebel's account, this scenario would still allow for a fully true belief ascription. On the other hand, it could also be the case that the lack of manifestation of a certain disposition indicated the actual absence of such a disposition. In this scenario, according to Schwitzgebel, we might still ascribe the belief that *p* to the agent as an 'ascriptive shorthand', in order to facilitate communication with a given audience (see also Tumulty, 2011, 2012). Here, the ascriber's interests and those of their audience play a decisive role in determining whether the belief ascription is useful or not. For example, if Red's divergences from the folk-psychological causal stereotype associated to (1) are in fact indicative of the absence of some relevant disposition (e.g., the disposition to reject trans-excluding activists' arguments against a Trans Rights law), then one will decide to ascribe him (1) or not in the light of one's interests or standards or those of the audience. If we are addressing an LGBT+ audience that considers unconditional support to trans people to be a necessary condition to frame someone as a political ally, then we won't probably describe Red's attitude towards the LGBT+ movement as a belief proper; on the contrary, if addressing a Spanish center-left politician, we won't probably find any problem describing Red's attitude in doxastic terms.

Drawing from Schwitzgebel's dispositionalism, Bayne & Pacherie (2005) also characterize belief ascription as a context-relative enterprise: if deviations from the dispositional stereotype are readily excused or explained by certain non-standard feature of the context of evaluation (i.e., if the *ceteris paribus clause* is cancelled), then we might still truthfully attribute the belief that *p* to the agent. Furthermore, if "a deviation from the stereotype cannot be excused or explained in this way, whether or not the attributor ascribes the belief will depend on the context of the belief ascription and *what her interests are*." (Bayne & Pacherie, 2005, p. 181). And this is exactly what Bayne & Pacherie (2004, 2005) claim to be the case in many instances of delusions. They discuss several possible factors that could explain the apparent deviation from the stereotypical dispositional profile that people with delusions exhibit, mainly focusing on two: environmental pressures and non-standard or disrupted perceptual, motivational or affective conditions. People with delusions might not systematically engage in delusion-consistent behaviors to avoid the risk of being hospitalized or detained, or to avoid being labelled "crazy" and consequently stigmatized, etc. In addition, since people with delusions allegedly have anomalous perceptual and affective and motivational experiences, Bayne & Pacherie hold that it would not be strange that similar anomalous processes could be responsible for the agent's deviations from the stereotypical profile.

5.2.1.3. Four common assumptions

Although Bortolotti's modest interpretivist doxasticism and Bayne & Pacherie's dispositionalist approach are proposed in response to slightly different antidoxasticist arguments, they inevitably share some common features, as do their underlying theories of belief (i.e., interpretivism and functionalism). Specifically, both kinds of revisionist doxasticism share four central interrelated commitments:

1) *Folk-psychological belief*: Firstly, both rely on a folk-psychological notion of belief, which is taken to be at least partially individuated in terms of its normative force, or capacity to enter into normative or justificatory relations with action (see also Bayne & Hattiangadi, 2013).
2) *Ontologically non-committal approach to belief*. Secondly, both revisionist doxasticisms inherit the ontologically non-committal attitude to the truth of belief ascriptions from their underlying theories of belief. Thus, both hold that an agent *has* a belief if and only if they can be *truthfully ascribed* such belief.
3) *Context-relativity of belief-state ascriptions*: Adding to the previous commitment, revisionist doxasticisms emphasize, in one way or another, the context-relative nature of belief ascriptions. This implies that whether an agent can be truthfully attributed

the belief that *p* or not does not exclusively depend on whether the agent displays the right cognitive, behavioral or even phenomenological patterns, but also on diverse features of the context of ascription. This allows both kinds of revisionist doxasticism to introduce or emphasize contextual considerations regarding whether people with delusions believe the content of their delusions or not.

4) *Continuity*: Finally, this gives room for their defense of the *continuity thesis*, or the idea that there is no sharp divide between non-clinical irrational beliefs and delusional beliefs, since the presumption of rationality (in a strong sense) is not constitutive of belief ascriptions[50]; in this sense, revisionist doxasticisms assume that "the difference [between delusions and other irrational non-clinical beliefs], if there is one, is not in their epistemic features" (Bortolotti, 2010, p. 259).

As a result, both kinds of revisionist doxasticism manage to articulate a response to the antidoxasticist challenges. In turn, both claim that delusional phenomena, or at least most of them, can be neatly accommodated within the conceptual framework of the folk-psychological notion of belief; delusions can be rightfully interpreted as beliefs or, in other words, delusions *are* beliefs.

5.2.2. Non-revisionist doxasticism: The cognitive phenomenological approach

As mentioned earlier, both Bortolotti's and Bayne & Pacherie's defenses of doxasticism draw from the very same theories of belief that first gave rise to RC and SCP, interpretivism and functionalism, respectively. On the contrary, Clutton (2018) has recently advanced a different defense of doxasticism about delusions; in particular, a non-revisionist defense: instead of introducing local adjustments of the interpretivist or functionalist frameworks, Clutton (2018) has proposed to reject both altogether[51], due to their alleged "anti-realist tendencies" towards belief (p. 11).

The author is pointing here at the ontologically non-committal attitude of both interpretivist and functionalist theories of belief, i.e., their lack of a clear ontological individualization of beliefs as *real* entities, separated from the behavioral profiles that characterize

[50] This constitutes a departure from some forms of interpretivism, such as the above-mentioned Dennettian (1979/1987) proposal, where the assumption of rationality is taken to be a necessary condition for the adoption of the intentional stance towards the explanation of an agent's behavior.

[51] To be sure, Clutton rejects interpretivism and *dispositionalist* functionalism. However, given that Clutton's theory of belief entails the view that beliefs are dispositions to entertain occurrent phenomenal states before the eyes of the mind, and that these have causal roles, Clutton's proposal can be construed as a particular kind of functionalist approach; specifically, as a hybrid kind of functionalism, which exhibits features of both occurrentism and dispositionalism (see Nottelmann, 2013).

them. This attitude, according to Clutton (2018), is incompatible with what he calls *scientific doxasticism* about delusions, i.e., the "robustly realist" (p. 11) doxasticist approach that figures in prominent cognitive models of delusions (e.g., Alford & Beck, 1994; Coltheart, 2007; Coltheart et al., 2011; Ellis & Young, 1990; Freeman et al., 2002; Garety, 1991; Frith, 1992; Maher, 1974/2005) (see Chapter 7). On his view, the respectability of these scientific theories is in itself a very good reason to *assume* that delusions are beliefs, and that these are real entities with an independent ontological status. Thus, Clutton sets the task of providing a philosophical account of belief that can yield a "robust defense of doxasticism", i.e., one which accommodates the realist commitments at play in traditional cognitivist models of delusions.

This is the task that Clutton's alternative conceptualization of belief is purported to do. Specifically, he endorses Kriegel's (2015) cognitive phenomenological theory of belief, which Clutton describes as follows:

> On this view, beliefs are dispositions to have certain intentional, occurrent mental states whose phenomenal character is that of "judging that P." Specifically, on this view, S believes that P iff S is disposed to immediately judge that P when P-entertaining triggers obtain [...]. (Clutton, 2018, p. 4)

In addition, according to the cognitive phenomenological approach, when we immediately judge that *p* (i.e., consider *p* to be true), "we feel [...] a sense of mental affirming, as we entertain the proposition" (Clutton, 2018, p. 4). That particular feeling of assent, which might come is different degrees of conviction, is the distinctive phenomenological mark of belief (one that is absent when the agent is just entertaining the possibility that *p*, for example). According to Clutton, these *epistemic feelings* are analogous -though irreducible- to their sensory counterparts (such as seeing red or listening to Mozart's flute concerto, to use Clutton's examples), but distinct in their epistemic nature. They "capture a range of experiences that are cognitive rather than sensory in nature, like the experience of understanding (or entertaining, or believing) a proposition." (Clutton, 2018, p. 3).

Thus far, the cognitive phenomenological proposal would not be so different from Schwitzgebel's functionalist-dispositionalist approach, except that the latter individuates beliefs in terms of complex sets of dispositions, where not only phenomenological, but also cognitive and (overt) behavioral dispositions are considered. However, Clutton's construal of the cognitive phenomenological account of belief gives a particular reading of the dispositional element present in Kriegel's definition:

> On my account, to be disposed to have occurrent episodes is for one's neural system to be set such as to trigger the relevant occurrent state in response to the triggering conditions. It is for one's neural system to be "set" in such a way that when P is considered, suggested to one, etc., one is apt to immediately judge that P. This neural state is the "truth-maker" of the disposition, the categorical grounds in virtue of which a belief ascription can be true." (Clutton, 2018, p. 5)

Thus, contrary to interpretivism and functionalism, Clutton's cognitive phenomenological approach is strongly committed to realism about beliefs: beliefs are real entities, neural states whose triggering produces in the subject a distinctive kind of cognitive-phenomenological experience, i.e., that of judging that *p*, with its distinctive epistemic feeling of affirming the entertained proposition. Finally, in the cognitive phenomenological approach, the agent has a special epistemic access to their own mental states; while others can only access an agent's mental states through a mediate, inferential strategy (i.e., via the observation of the agent's behavior), the agent themself has direct or acquaintance knowledge of their own beliefs.

Importantly, these last two features of Clutton's proposal (i.e., the privileged conception of self-knowledge and the strict identification of the phenomenological disposition to judge that *p* with a certain neural state) introduce a radical departure from traditional defenses of doxasticism. The cognitive phenomenological approach still identifies *having* the belief that *p* with *being truthfully ascribed* the belief that *p*; however, on this account, there're just two ways for securing the truth of a particular belief ascription: a) from a first-person perspective, determining whether one systematically manifests the relevant phenomenological disposition to judge that *p* (something which only oneself has "direct access to", according to the cognitive phenomenological account); or b) from a third-person perspective, determining whether the person is in the relevant neural state -what Clutton (2018, p. 5) considers to be "the "truth-maker" of the disposition, the categorical grounds in virtue of which a belief ascription can be true". Not only the person's motor behavior, but also any cognitive and phenomenological activity other than that of "privately judging that *p*" are regarded as mere indirect and inconclusive sources of evidence about the person's beliefs. From a third-person point of view, the strongest indirect source of evidence would be the person's sincere assertion that they believe that *p*, since such assertion is taken to be an honest descriptive report of the person's subjective experience; according to Clutton (2018, p. 6), "we have good reason to accept such reports (we can take them at face value in the same way we would take a person's report of "seeing red" at face value)".

These features of the cognitive phenomenological view of belief are supposed to allow for a straightaway interpretation of delusions in doxastic terms. It does away with the SCP and RC objections in a straightforward manner; since the most reliable indicators of an agent's beliefs are their sincere self-ascriptions and, more importantly, their neural activity, whether they display attitude-attitude or attitude-behavior inconsistencies (i.e., bad integration or double-bookkeeping, respectively) is to some extent irrelevant. If Green sincerely asserts (3), or if his neurologist determines that such and such neural states correspond to his belief that if he doesn't comply with his Karmic duties, his partner will be attacked, then we should take Green's avowal or his neurologist's word at face value. This way, the cognitive phenomenological theory of belief provides a most pleasant accommodation for the realist assumptions behind scientific doxasticism and the cognitive models of delusions: beliefs are real entities, which cause belief-like patterns of behavior, cognition, and experience, and are supposed to be equivalent to certain states of the person's neural circuitry.

So far, we've seen how doxasticism towards delusions has been standardly assumed, then rejected and finally defended by various authors, drawing from different theoretical accounts of belief. After all these comings and goings, one might ask: but, why insisting so much on a defense of doxasticism? Why is it important whether delusions can be properly understood as beliefs or not?

5.2.3. The two desiderata of doxasticism

Defenses of doxasticism are usually motivated by a number of reasons, some of which have already come out. To begin with, doxasticism is typically defended on the grounds of its presumed scientific and clinical virtues. Firstly, there is an extended research framework in the field of cognitive neuropsychiatry (see Chapter 7, section 7.1.), that aims to develop a proper etiological account of delusions in doxastic terms (e.g., Coltheart et al., 2011; Ellis & Young, 1990). The rationale beyond the defense of doxasticism thus goes as follows: drawing from the assumption that we have a proper account of the subpersonal mechanisms underlying normal cognition, conceptualizing delusions as beliefs leaves us in a better position to study the deviations in those subpersonal mechanisms that could give rise to the development and maintenance of delusions (Bayne & Pacherie, 2005; Bortolotti, 2010, 2012; Bortolotti & Miyazono, 2014; Clutton, 2018; López-Silva, 2018; see Alford & Beck, 1994; Coltheart et al., 2011; Freeman et al., 2002; see also Currie & Jureidini, 2001; Tumulty, 2011, 2012). Clutton's non-revisionist defense especially emphasizes this motivation, but Bortolotti's (2010) and Bayne & Pacherie's (2005) revisionist approaches display a similar rationale when justifying their defense of doxasticism.

Relatedly, doxastic approaches aim to provide a sound framework to understand and promote the effectiveness of cognitive-behavioral interventions, and specifically the workings of cognitive techniques, e.g., the cognitive restructuring of the person's belief system via the Socratic dialog. As we saw in Chapters 1 and 2 (sections 1.3.2. and 2.3.3.), cognitive therapy is at least partially based on the assumption that what the person does and says, their problematic verbal and non-verbal behavior, is the result of their cognitive system, of how they represent the world they live in, other people and themselves (see also Chapter 7). Thus, the main therapeutic goal is to change the person's belief system through a process of questioning the evidential basis, consistency, or utility of certain beliefs. Given that some delusional cases are effectively treated through this kind of techniques, it seems to follow that the best way to characterize delusions is in doxastic terms (Bayne & Pacherie, 2005; Bortolotti, 2010, 2012). Otherwise, how is the effectiveness of these procedures to be explained? Furthermore, in providing the conceptual grounds for cognitive theories of delusions, doxasticism not only is supposed to be better equipped than antidoxasticist approaches to explain why cognitive techniques work, but also to promote their development by pointing to potentially relevant causal factors.

In addition, supporters of doxasticism also typically claim that doxasticism fits best with how our folk-psychological interpretative practices work in the case of delusions; in particular, they claim that doxasticism is better equipped to accommodate the fact that we, as folk psychologists, straightforwardly interpret cases like Green's in terms of belief. Rose et al (2013) have actually provided empirical support for this claim. The authors conducted a series of experiments where they presented participants with a story of a person with Capgras delusion (i.e., a delusion where the person claims that a beloved one is in fact an impostor) and controlled different variables to observe how that affected the probability that participants ascribed the belief in the claim's content to the person with Capgras syndrome. In their first experiment, the variable controlled by the authors was the presence or absence of assertion-behavior inconsistencies in the description of the delusional case. Specifically, participants were presented with either one of two possible endings: a) in the Typical (i.e., inconsistent) condition, the patient still treated his partner as if she were his wife (e.g., eating, sleeping and enjoying leisure time together); or b) in the Atypical (consistent) condition, the patient ceased to treat his partner as if she were his wife (e.g., ceasing to eat, sleep or going out with his partner). The authors observed no significant differences between the two conditions: no matter whether the Capgras patient displayed attitude-behavior inconsistencies or not, nearly a 100% of participants in both groups still ascribed to the patient the belief that her partner had been replaced by an impostor.

It thus seems that we, as folk psychologists, tend to interpret delusional cases in doxastic terms, regardless of whether the person displays belief-like behavior or not. Claiming that not only people with delusions, but also the vast majority of us are wrong about the kind of attitude that we should ascribe to people with delusions thus seems strained. In this sense, doxasticism is preferrable to antidoxasticism, since it does not need to provide any kind of error theory for why we folks are systematically misled towards understanding delusions as beliefs.

Besides this conceptual advantage, the capacity of doxasticism to reflect our actual folk-psychological interpretative practices is also connected to another major reason why doxasticism is usually endorsed: its presumed ethical and political virtues. These have to do with the conceptual link between our capacity to appropriately ascribe beliefs to an agent and our capacity to ascribe them with agency, autonomy, and responsibility for their actions (Bortolotti, 2010, 2012; Bayne, 2010; Bayne & Hattiangadi, 2013; see also Broome et al., 2010; Graham, 2010a; Sullivan-Bissett et al., 2016; Tumulty, 2012). As Bortolotti (2010, p. 1) puts it:

> [...] having beliefs is a necessary condition for autonomous agency. This means that to have a principled way to tell whether an individual can be ascribed beliefs and other intentional states does not only satisfy intellectual curiosity about the criteria for mindedness, but contributes to determining appropriate ethical stances towards other individuals. (Bortolotti, 2010, p.1)

Thus, doxasticism not only is the conceptualization of delusions that best reflects our folk-psychological interpretative practices; in addition, since belief ascription practices are intimately linked to our agency ascription practices, defending doxasticism leaves us in a better position to defend the adoption of an agential stance towards people with delusions. In the end, this would help us prevent cases of unjust and abusive treatment, e.g., cases of epistemic injustice in mental health encounters, violations of the person's right to informed consent, employment of dehumanizing methods like chemical or mechanical contentions etc. (Bueter, 2019; Carel & Kidd, 2014; Crichton et al., 2017; Drożdżowicz, 2021; Fernández-Costa et al., 2020; Miller-Tate, 2019; Ritunnano, 2022).

As we view it, the main motivations behind doxasticism can be cashed in terms of the following two desiderata:

a) A *scientific desideratum*, related to the claim that doxasticism leaves us in a better position to account for the causal processes involved in the acquisition, adoption and persistence of delusional states, which might thus in turn have implications for the

comprehension or development of our psychiatric and psychological intervention designs and techniques.

b) An *ethico-political desideratum*, related to the claim that doxasticism is best suited for defending the idea that delusional phenomena can be cashed in intentional or intelligible terms, thus safeguarding our attributions of agency, responsibility and autonomy to the person diagnosed; this, in turn, would provide a further barrier protection for their right to ethical treatment (including things like informed consent, avoidance of unjust or abusive assessment and treatment methods, etc.).

Although the commitment to these two general desiderata shows up throughout the work of many doxasticists, we think that the beginning of Bortolotti's (2010, p. 3) seminal work, *Delusions and Other Irrational Beliefs*, constitutes a prime example of this:

> First, agreeing on the belief status of delusions would provide justification for the approach of cognitive neuropsychology, which is explicitly grounded in the assumption that abnormal cognition can be explained as a deviation from the very same processes that characterise normal cognition. [...] Second, confirming the belief status of delusions would have consequences for diagnosis and therapy in psychiatry [...]. Fourth, characterising subjects with delusions as intentional agents capable of forming beliefs and acting on them would impact significantly on current debates about their ethical standing in clinical and forensic settings, and on the suitability of different types of psychiatric treatment. (Bortolotti, 2010, p. 3)

In the following chapter, we will try to show that neither revisionist nor non-revisionist doxasticisms are able to meet both of the above-mentioned desiderata at the same time. This, we'll argue, is due to the particular conceptualization of beliefs and belief ascriptions at play in each doxasticist approach. Against this background, we'll see how the pragmatist kind of non-descriptivism defended in Chapter 4 may provide some interesting insights in this debate, and how it allows for a different defense of doxasticism about delusions -specifically, one that emphasizes its ethical and political, rather than scientific virtues.

5.3. Conclusion

In this chapter we've laid out the main axes of debate regarding the standard conceptualization of delusions as irrational or strange beliefs, present both in traditional nosologies and cognitive models of delusions, as well as in the folk-psychological imaginary. After introducing this standard doxasticist approach to delusions, we've reviewed several ways in which it has been problematized. We've distinguished two main theories of belief from which anti-doxasticists draw to deny a doxastic status to delusions: interpretivism, according to which

beliefs are primarily individuated in terms of a series of rationality constraints (i.e., content, epistemical, procedural, and agential rationality), and functionalism, according to which beliefs are primarily individuated in terms of their particular causal roles, i.e., their particular causal connections with perception, other cognitions, and behavior. These two closely related conceptual frameworks give rise to the two main arguments against doxasticism: a) the Rational Constraint argument (RC), which states that delusions are not beliefs because people with delusions fail to act, cognize, and experience as it would be rational to expect of someone who really believed the content of the delusion; and b) the Stereotypical Causal Profile argument (SCP), which states that delusions cannot be adequately understood as beliefs because they fail to display the relevant belief-like causal roles.

Despite sharing this common negative thesis, antidoxasticist proposals diverge on their positive characterization of beliefs. In this sense, we've distinguished two main kinds of antidoxasticism. On the one hand, commonsensical antidoxasticists deny the doxastic status of delusions, but not the possibility to understand them with folk-psychological resources. These approaches thus attempt to explain delusions in terms of propositional attitudes other than belief, e.g., in terms of imaginings which the person mistakes for beliefs. On the other hand, non-commonsensical antidoxasticists reject both the doxastic understanding of delusions as well as their folk-psychological understanding. Instead, these approaches either opt for establishing a new, *sui generis* mental type for delusions (e.g., "bimaginations"), or rather for simply laying out the particular dispositional profile that characterizes delusions or subtypes of them.

After explaining the main criticisms against doxasticism, we've reviewed some its main defenses. We've distinguished here between two kinds of doxasticism: a) revisionist doxasticism, which assumes that, once certain conditions are take into account, interpretivism and functionalism still motivate a doxastic understanding of delusions; and b) non-revisionist doxasticism, which dismisses interpretivism and functionalism and proposes instead an alternative theory of belief. Regarding the former, we've seen how Bortolotti's revised interpretivism and Bayne & Pacherie's dispositionalist defense offer a twofold strategy to vindicate doxasticism about delusions. Firstly, both reject the empirical claim that people with delusions typically fail to conform to the rational or causal profiles of belief. Secondly, both introduce a series of revisions to the interpretivist and functionalist frameworks, e.g., offering some relaxed version of the rationality constraints or invoking a *ceteris paribus* or all-things-being-equal clause. In sum, these approaches inherit some of the similarities that hold between their mother theories of belief, namely the commitment to a folk-psychological notion of belief, the non-committal attitude towards the ontological status of beliefs, the

context-relativity of belief ascriptions, and the defense of the continuity thesis, i.e., the idea that delusions aren't categorically different from other non-clinical beliefs.

Regarding non-revisionist approaches, we've reviewed Clutton's recent cognitive-phenomenological defense of scientific doxasticism. This approach rejects the interpretivist and functionalist theories of belief altogether because of their "anti-realist tendencies" towards the notion of belief, which the author sees as incompatible with scientific doxasticism, i.e., the conception of delusions at play in prevailing cognitive models of delusions. Instead, he proposes the cognitive-phenomenological theory of belief, according to which beliefs are dispositions to "mentally assent" to *p*, whenever *p*-entertaining triggers obtain. According to this theory, there are two infallible sources of evidence regarding whether a person has a certain belief or not: the person's own cognitive-phenomenological dispositions, to which the person has some kind of direct epistemic access, and the neural states that presumably realize those cognitive-phenomenological dispositions. In addition, the most secure indirect source of evidence would be the person's sincere belief self-ascription. This way, Clutton provides a "robust defense of scientific doxasticism" (p. 2), which accommodates the realist assumptions of cognitive models of delusions.

Finally, we've explained which are the main reasons why doxasticism has been defended. In particular, we've highlighted what we see as the two main desiderata of doxasticism: a) the scientific desideratum, according to which doxasticism would leave us in a better position to understand the cognitive factors underlying the development and maintenance of delusions, hence contributing to the explanation and development of clinical procedures; and b) the ethico-political desideratum, according to which doxasticism leaves us in a better position to understand delusions in intentional terms, hence protecting the person's agency and providing some sort of conceptual barrier against unjust or abusive assessment and treatment practices.

In the upcoming chapters, we'll discuss the main merits and demerits of some of approaches reviewed here. The overarching goal will be to show what contributions our non-descriptivist approach to the mind can make to the debate on the typology problem, and what implications it might have for mental health assessment and treatment practices. In Chapter 6, we'll discuss whether current defenses of doxasticism can live up to their own scientific and ethico-political desiderata, and conclude that they cannot: specifically, we'll claims that a) revisionist approaches, if they are to provide a proper defense of doxasticism, cannot meet the scientific desideratum; and b) Clutton's non-revisionist approach, in its attempt to articulate a theory of belief that is akin to scientific doxasticism, fails to meet the ethico-political desideratum. We'll also discuss how our non-descriptivist approach to the

mind offers the kind of conceptual framework that revisionist doxasticisms need to provide a proper defense of doxasticism, which emphasizes its ethico-political virtues over its scientific implications. In Chapter 7, we'll also present various objections against Clutton's favored scientific doxasticism, concluding that the scientific desideratum doesn't provide us with enough good reasons to defend doxasticism about delusions. Finally, in Chapter 8 we'll introduce a non-cognitivist approach to the intervention with people with delusions, as well as the main contributions that non-descriptivism could make to such kind of interventions. On the whole, we'll see that our non-descriptivist approach to the mind provides us with the tools to accommodate a seemingly paradoxical claim: that (most) delusions are rightly conceptualized as beliefs, but that this needn't leave us in a better position to understand the causes of delusional phenomena nor to consequently design better interventions with people with delusions.

Chapter 6
A non-descriptivist defence of doxasticism about delusions

In the previous chapter we introduced the contending positions in the debate around the typology problem, i.e., the debate around how best to conceptualize delusions. Prevailing approaches to their nosology, scientific understanding, and treatment endorse a standard doxasticist conception of delusions, which conceptualizes delusions as strange or irrational beliefs. However, antidoxasticist approaches have questioned the conceptual viability of this definition; in particular, antidoxasticists claim that delusions cannot be adequately understood as beliefs because the particular behavioral, cognitive, and phenomenological patterns displayed by people with delusions fail to fit the causal or rational stereotypes that characterize beliefs, according to functionalist and interpretivist theories of belief, respectively.

Several authors have contested these criticisms and attempted to elaborate a proper defense of doxasticism about delusions. Bortolotti's (2010) and Bayne & Pacherie's (2005) revisionist doxasticisms, on the one hand, claim that, once certain considerations are taken into account, interpretivist and functionalism can in fact grant a doxasticist understanding of delusional phenomena. Clutton's (2018) non-revisionist defense, on the other hand, draws from a rejection of functionalism and interpretivism, endorsing instead an alternative, cognitive phenomenological theory of belief. Regardless of its various defenses, doxasticism has been typically motivated on the grounds of two main desiderata: a) the scientific desideratum, or the idea that doxasticism is better equipped to inform scientific research on delusions and the clinical intervention with people with delusions; and b) the ethico-political desideratum, or the idea that doxasticism provides a way to render delusional phenomena intelligible in folk-psychological terms, which in turn reinforces the agential status of people with delusions and prevents potential cases of unjust or abusive treatment.

The main goal of this chapter will be to show why current defenses of doxasticism cannot meet the aforementioned desiderata at once. In section 6.1., we'll delve into the reasons why revisionist doxasticisms fall short of the scientific desideratum. The reason, as we'll see, lies in the vagueness of the folk-psychological notion of belief. As we already pointed out in Chapter 3 (section 3.2.1.), folk-psychological concepts are just too vague in causal terms to constitute an optimal explanatory tool for a scientific account of delusions and their clinical intervention; the case at hand here will provide a more nuanced view of why this is so. In particular, we'll see that, to properly defend that delusions are cases of folk-psychological beliefs, revisionist doxasticists must ultimately assume that the truth of belief ascriptions can vary depending on the ascriber's evaluative framework. If that's the case, then their particular kind of doxasticism is of little use for scientific approaches to delusions.

In section 6.2., we'll discuss why Clutton's non-revisionist proposal fails to meet the ethico-political desideratum. We'll claim that his approach can be understood as a response to the problems of revisionist defenses of doxasticism; by means of his cognitive phenomenological theory of belief, which redefines the notion in terms more akin to cognitive-scientific models of delusions, his approach aims to provide a more homely account of the intuitions behind the scientific desideratum. However, it does so at the expense of the ethico-political benefits of doxasticism. The problem lies in its commitment to a dual kind of descriptivism, i.e., one which assumes that belief ascriptions are reducible to: a) the person's neural states; or b) the person's sincere report of their mental states. We'll argue that, for reasons similar to the ones discussed in Chapter 3 regarding the self-defeating character of both reductivist and non-naturalist approaches to mind, both readings of Clutton's proposal yield a notion of belief which lacks normative force (i.e., one which cannot rationalize a person's actions); consequently, the resulting kind of doxasticism is of little use for ethico-political purposes.

In section 6.3., we'll consider again whether, in the light of the previous discussion, doxasticism towards delusions should be defended at all. We'll argue that it does; not on the grounds of its scientific virtues though, but on the grounds of its ethico-political benefits. Specifically, we'll claim that revisionist doxasticisms, once viewed through the lenses of our non-descriptivist approach to the mind (see Chapter 4), are in a better position than its competitors to account for how our actual folk-psychological interpretative practices work in the case of delusions and why this conceptualization should be retained. We'll argue that doxasticism, conceptualized as a "by default" policy (i.e., "take the person's self-ascriptions at face value") can serve as a conceptual barrier against undue deagentializing practices and their concomitant risks in the field of mental health. However, we'll also argue that

revisionists' underlying commitment to the mindreading conception of folk psychology (see Chapters 2 and 3, section 2.2.1. and 3.1.1.) threatens to ditch their defense of doxasticism, because it makes them vulnerable to eliminativist arguments. We'll thus claim that our non-descriptivist approach to the mind offers a way to develop a more robust defense of doxasticism; one which is able to preserve its ethical and political benefits regardless of how far psychological sciences reach in the causal explanation of delusional phenomena.

Finally, in section 6.4., we'll conclude with an outline of the main conclusions of this chapter and present the topics to be addressed in the following ones.

6.1. The scientific desideratum: beliefs and their fuzzy causal roles

As we saw at the end of the previous chapter, one of the main reasons why doxasticism is typically endorsed is that it presumably offers certain advantages for research and intervention on delusional phenomena. Specifically, in framing delusions as beliefs, doxasticism allegedly leaves us in a better position to understand the causal processes involved in the development and maintenance of delusions. In this section, we'll explore whether the kind of notion of belief at play in revisionist defenses of doxasticism is compatible with this scientific desideratum.

To begin with, we might wonder how exactly do revisionist doxasticisms attempt to inform scientific or clinical research on delusions. As we saw in Chapter 5 (section 5.2.1.), both Bortolotti's (2010, 2012) modest doxasticism and Bayne & Pacherie's (2005) dispositionalist defense attempt to offer a better characterization of the folk-psychological notion of belief. Specifically, they propose several revisions to the interpretivist and functionalist frameworks that motivate antidoxasticism in order to enable the incorporation of delusional phenomena within the limits of this notion. One of their core shared commitments, however, is their ontologically non-committal attitude towards beliefs and belief ascription practices; there's thus no intention to provide a full ontological account of what beliefs are or what specific natural processes give rise to beliefs or their alleged pathologies. How, then, may this be informative for scientific or clinical accounts of delusions, which attempt to explain delusions in terms of the mechanisms and processes giving rise to them?

As Bortolotti (2010, 2012) herself recognizes, it's true that the kind of doxasticism she endorses remains silent about ontological and etiological questions; however, this doesn't mean that it's not useful for scientific or clinical models of delusions. According to her, the scientific appeal of revisionist doxasticisms lies in its ability to provide a proper *conceptual map of belief*; i.e., a proper characterization of the criteria that we as folk psychologists follow when we ascribe beliefs to one another, on the hope that "a good account of how belief

ascription works will impose constraints on the type of things that can play the role of beliefs" (Bortolotti, 2010, p. 2). To put it differently, the main goal of revisionist doxasticists is to delineate the conditions under which our folk-psychological belief ascriptions can be hold *true*; this, in turn, would provide a good roadmap for (neuro)scientists to establish the relevant mechanisms or processes involved in the causal production of belief-like phenomena[52].

Thus, for a doxastic account of delusions to be scientifically informative, revisionist doxasticists must be able to establish why belief ascriptions can be hold true in the case of people with delusions, despite their attitude-attitude and attitude-behavior inconsistencies. In addition, they need an account of belief ascriptions that yields their truth or falsity useful for the scientific investigation of the causal processes involved in the onset and maintenance of delusions.

Once this is clarified, we can ask whether revisionist doxasticisms can provide this kind of account of the truth conditions of our folk-psychological belief ascriptions. As we'll argue in what follows, they cannot. The problem lies in that, in order to accommodate delusions within their doxastic approaches, they end up letting the truth value of belief ascriptions depend -at least partially- on the *ascriber's standards.* Consequently, it is difficult to see how this could be of use for a scientific account of delusions: if a person's intentional state is -or is not- a belief depending on the ascriber's evaluation motives and standards, then how exactly is a doxasticist account of delusions going to inform research on the investigation of their causes? In other words: how are scientists to determine which are the causal mechanisms corresponding to a belief ascription that may vary *from ascriber to ascriber*?

6.1.1. Revisionist doxasticism and the elusive cartography of belief

Before delving into the reasons why revisionist doxasticisms are unable to meet the scientific desideratum, let's first recall what revisionist doxasticists take to be the criteria for true belief-ascription; in other words, let's take a look again at what their roadmaps of the concept of belief look like. Given that, despite some differences in emphasis, Bortolotti's modest doxasticism and Bayne & Pacherie's defense of doxasticism share a good deal of common ground, and given that our analysis will mainly delve with common features of these two approaches, we'll treat them here jointly.

Taken together, what both kinds of revisionist doxasticisms take to individuate the concept of belief, and thus the truth conditions of our folk-psychological belief ascriptions, is its *logical-causal profile*, i.e., its typical -and rationally intelligible- causes and effects. This

[52] In a sense, we could say that revisionist doxasticist approaches entail some sort of reductivism (see Chapter 2, section 2.2.2.1.), their role being the specification of what exactly is to be translated.

seems to provide a nice start point for the delineation of a useful roadmap for a scientific account of belief: all one has to do is to properly specify which are these causes and effects and then let scientists establish what kind of natural processes or mechanisms might be responsible for such logico-causal profile. In this story, revisionist doxasticists are like offender profilers, while scientists are the detectives in search for a culprit fitting, to some extent, the offender's profile. Or, abusing Ryle's metaphor, revisionist doxasticists would be like cartographers mapping the *logical* (or rather, *logical-causal*) *geography* of belief, while scientists would be in charge of finding the natural processes responsible for its particular type of landscape.

So far, so good. However, the problem comes precisely when trying to clearly specify how delusions, which many of us feel immediately inclined to interpret in doxastic terms, fit in the revisionist doxasticist's preferred map of belief. Delusional cases, such as Green's case, often fail to fit the belief glove, and sometimes strepitously. As we saw, revisionists respond to this challenge by loosening the glove steams, i.e., by revisiting the traditional interpretivist or functionalist frameworks and proposing a more laid-back account of belief and belief ascription practices.

Still, many delusional cases seem to meet neither of these less stringent criteria fully. Revisionists then fall back on another strategy: even in these cases, the truth of a belief ascription to people with delusions might be secured by invoking a cancellation of the *ceteris paribus* or "all-things-being-equal" clause, i.e., by resorting to some abnormal or non-standard feature of the context of belief ascription. As we saw, this is one of the core features of revisionist doxasticism: the conceptualization of belief ascription as a context-relative enterprise. Applied to delusions, revisionist doxasticists hold that their typical divergencies from the logical or causal stereotype of beliefs can be properly excused or explained away by non-standard features of the context of belief ascription which are common in cases of delusions (e.g., environmental pressures that preclude the manifestation of the disposition in question, anomalous affective or motivational states of the agent, etc.) (Bayne & Pacherie, 2005; Bortolotti, 2010, 2011, 2012). As we saw in section 5.2.1.2., Bayne & Pacherie (2005) draw upon Schwitzgebel's account of belief to argue that:

> [...] whether one should be ascribed the belief that P is not just a matter of whether the target manifests enough of the dispositions in the relevant cluster but also of whether his not manifesting some of these dispositions can be satisfactorily excused or explained by reference to some non-standard aspects of his situation.
>
> [...] in the case of many delusional patients an appeal to relevant non-standard factors can be made. [...] These non-standard perceptual and affective conditions may be thought to excuse

> the patient from manifesting the cognitive dispositions stereotypically associated with their belief. (Bayne & Pacherie, 2005, p. 148).

In addition, they also note that:

> [...] action is not caused by cognitive states alone but by cognitive states in conjunction with motivational states. As Stone and Young remind us (1997), deluded patients have disrupted affective and emotional states, and they know that acting on their beliefs might result in hospitalization. (Bayne & Pacherie, 2005, p. 185)

On the other hand, Bortolotti follows this same strategy in her reply to her critics. Specifically, in response to Schwitzgebel, she states that:

> [...] delusions are generally behaviourally effective, but can fail to guide action due to phenomena that are anything but rare in the psychiatric disorders that manifest with delusions, such as schizophrenia, dementia, and delusional disorders. These may include meta-representational deficits, conflicting attitudes, co-morbidity with depression, and fluctuations in motivation caused by changes in affect (e.g., poverty of action, avolition, flat affect, emotional disturbances). Action that would follow some delusions with bizarre content can also be inhibited by features of the physical and social environment surrounding the agent. (Bortolotti, 2012, pp. 47-48)

This is exactly where things start to go sideways for revisionist doxasticists. The problem, in short, is that this cancellation of the *ceteris paribus* clause cannot properly *excuse* certain attitude-attitude or attitude-behavior inconsistencies in the case of delusions. Tumulty (2011) has provided a full-fledged development of this counterargument. She focuses specifically on Bayne & Pacherie's use of Schwitzgebel's dispositionalism to preserve the doxastic character of delusions, but the results of her criticism also affect Bortolotti's modest interpretivist proposal. Let's thus see Tumulty's analysis in more detail.

6.1.2. Dispositionalism and doxasticism: a marriage of convenience?

On her critical analysis, Tumulty (2011) points out two inter-related issues with Bayne & Pacherie's interpretation of Schwitzgebel's proposal: a) their overestimation of the capacity of the dispositionalist account to safeguard the doxasticist approach through the cancellation of the *ceteris paribus* clause in the case of delusions; and b) their uptake of Schwitzgebel's idea of the context-relative nature of belief ascriptions. Let's see this in more detail.

Firstly, Tumulty argues that Bayne & Pacherie overlook Schwitzgebel's distinction between *excuses* and *explanations* of the functional deviances from the causal stereotype of belief. Excusing a dispositional deviance implies that the agent in fact *has* the relevant disposition, but fails to manifest it due to some non-standard situation. Let's assume that Green is taking online learning lessons from Monday to Thursday; if this was the case, then his only part-time trash-collecting activity could be due to the fact that he stays at home during those days and thus he doesn't get to see the garbage dumped in the road nearby his home. Or maybe he is so overloaded with work that he repeatedly forgets to comply with his Karmic duties during the workweek. These are proper excusing conditions. On the contrary, explaining a dispositional deviance, in Schwitzgebel's account, implies that the relevant disposition is in fact missing, due to whatever reason.

According to Tumulty, Bayne & Pacherie conflate these two ways of accounting for why an agent might fail to manifest a relevant disposition. On the one hand, Tumulty argues, it's not clear whether many of the examples that they present as excuses of the functional deviances are in fact excuses or, on the contrary, explanations of a dispositional absence. For example, Bayne & Pacherie argue that the fact that people with delusions often fail to act as expected is due to certain failures in their cognitive or affective 'circuits', so to speak. However, it's not clear whether the appeal to these 'circuit breakdowns' excuses the person's failure to act as expected or simply explains why some relevant dispositions are missing.

Secondly, according to Tumulty, this conflation of properly excusing and merely explanatory factors obscures the actual sense in which Schwitzgebel thinks that belief ascription is context-relative. Bayne and Pacherie understand that it's the *truth* of a belief ascription that is context-relative; thus, their notion of context-relativity not only encompasses the consideration of contextual variables that might preclude the manifestation of a relevant disposition, but also the *ascriber's interests and standards*; these, in their account, can also play a role in determining the truth value of a given belief ascription. To recall the example from section 5.2.1.2., this would mean that the standards of Red's audience could have an impact on the truth-value of his belief self-ascription in (1) (i.e., "I believe that the LGBT+ movement deserves our full support"): while it would be false for a pro-trans rights audience, it could nonetheless be true for a Spanish center-left politician.

However, in Tumulty's reading, the only way a belief ascription can still be properly *true* in the face of functional deviancies is by finding an appropriate excuse for such deviancies -not an explanation. Only if the relevant disposition is judged to be present, despite it not being acted upon, then the belief ascription will be true proper. On the contrary, if a relevant disposition is missing, then ascribing a certain belief will just be a linguistic

shortcut, whose convenience will depend on the ascriber's standards or their audience's; such ascription, however, will not be 'fully true' in a strict sense. According to Tumulty (2011, p. 603), "This is where context-relativity comes in on Schwitzgebel's account: with interpreters' pragmatic judgments as to whether a particular deviation is *important* given their communicative aims". Therefore, context-relativity is not invoked in relation to the *truth* of a given belief ascription, but in relation to its *convenience* in particular communicative exchanges. Sometimes it will be convenient to use belief-ascriptive language, while on other occasions it will be better to just specify the agent's dispositional profile exhaustively. In any case, the truth of the belief ascription falls entirely on the agent's particular dispositional profile and its matching or not the appropriate causal stereotype.

> [...] on Schwitzgebel's account, once a dispositional profile has been exhaustively specified, there is no further factual question as to whether or not a subject really (for example) believes that *p*. [...] No-further-fact dispositionalism means subjects with identical dispositional profiles can't have different mental states in distinct contexts. The introduction of belief-ascriptive language does not add information beyond that contained in a dispositional profile but refers to that information in a convenient way (Schwitzgebel, 2002, p. 252 n.6). Context comes in, on this account, when attributors must decide whether the use of ascriptive language (rather than a specification of profile) will be helpful to their audience. No interpreter wants to ascribe a belief if doing so will mislead her audience, by causing them to form inappropriate expectations about the behavior of the subject being discussed. (Tumulty, 2011, pp. 600-601)

6.1.3. Revisionists' context-relativity and the fuzzy causal roles of belief

Tumulty's criticism does not only have theoretical implications regarding the fitness of Bayne & Pacherie's defense of doxasticism within Schwitzgebel's dispositionalism. Beyond that, it highlights an inherent problem regarding the utility of folk-psychological belief ascriptions (as revisionist doxasticists construe them) in the scientific understanding of delusions. As matters stand, revisionist doxasticists are now left with two options. They might either: a) accept Schwitzgebel's idea that the agent's dispositional profile exhausts everything that we can say about the truth or falsity of a folk-psychological belief ascription; or b) subscribe to a kind of theory of belief similar to Bayne & Pacherie's construal of the dispositionalist stance, where the ascriber's standards are allowed to have an impact on the *truth* of a folk-psychological belief ascription. However, as we will now see, both options are problematic for the revisionist doxasticist.

First, if we strictly abide by the idea that the agent's dispositional profile exhausts everything that can be said of the truth or falsity of a given belief ascription, then the

doxasticist must abandon a substantive account of delusions in terms of beliefs. Several dispositionalists, including Tumulty (2012) and Schwitzgebel (2012) themselves, consider that delusions are indeed cases where it is not possible to resort to non-standard features of the context of evaluation so that we can understand delusional phenomena in doxastic terms proper. This leaves open the possibility of understanding delusions as "in-between cases": cases where the person manifests some, but not all, the stereotypical dispositions that we as folk psychologists associate with the belief that *p,* and where these deviations are not readily *excusable* by appeal to other variables precluding the manifestation of the relevant dispositions. Schwitzgebel himself (2012, p. 13) proposes to understand delusions as

> [...] cases of belief gone half-mad, cases in which enough of the functional role characteristic of belief is absent that the subject is in an "in-between" state regarding the delusive content, such that it is neither quite right to say the subject determinately believes the delusive content nor quite right to say that she determinately fails to believe that content. (Schwitzgebel 2012, p. 13)

On this scenario, the only way out for the doxasticist is to accept that belief ascriptions might not be *true* proper in the case of delusions, but counter that it might still be *convenient* in communicative terms to describe delusions as beliefs, for a number of pragmatic reasons. In this case, choosing to describe delusional cases in terms of beliefs will be at best a matter of communicative convenience.

Bortolotti (2012) gets close to concede this much when discussing the superiority of her modest doxasticist approach over Schwitzgebel's sliding scale approach. Specifically, she draws upon the ethico-political desideratum discussed in section 5.2.3. to justify, on pragmatic grounds, an on-off doxastic understanding of delusions even in cases where deviances from the folk-psychological stereotype of belief are not properly excused.

> The sliding scale might deliver good results in terms of allowing us to discriminate between mental states with more or less belief-like features, but it becomes *impractical* if we think that a lot hangs on whether an individual is ascribed beliefs. Suppose we think that only individuals with beliefs [...] are entitled to a certain form of moral consideration (e.g., because their possession of intentional states indicates that they also have morally relevant interests). Then, the ascription of partial beliefs does not help. [...] debates about whether people with delusions genuinely believe what they say informs claims about their capacity for autonomy and responsibility, about appropriateness of treatment, and about potential suspension of rights. (Bortolotti, 2012, pp. 45-46; emphasis added)

If we read Bortolotti as implying that our belief talk in the case of delusions might not be true proper, but nonetheless a most convenient way to describe people with delusions, the doxasticist claim that delusions *are* beliefs loses its strength considerably; it is reduced to a communicative recommendation, a proposal for a linguistic policy.

Alternatively, if revisionist doxasticists want to preserve a substantive doxasticist account of delusions (i.e., one where delusions can be *truthfully*, and not just conveniently, described in terms of belief), then they would be better off with some kind of theory of belief that allows for the ascriber's or their audience's standards to have an impact on the truth of belief ascriptions, such as the one that Bayne & Pacherie seem to have in mind. We could read Bortolotti's claim along these lines: here, pragmatic considerations related to the ethical and political perils of negating a doxastic status to delusional states should become part of what may determine the truth or falsity of our belief ascriptions.

In section 6.3. we'll defend this precise kind of approach to doxasticism about delusions in relation to the ethico-political desideratum. However, the problem now is that none of the two options we've just discussed (i.e., limiting the truth of belief-ascriptions to the specification of the person's dispositional profile nor letting pragmatic considerations form part of the truth-conditions of belief ascriptions) allows the doxasticist to satisfy the scientific desideratum. On the one hand, if understanding delusions in doxastic terms is a matter of communicative conveniency, it's hard to see how this linguistic policy can serve as a roadmap for a scientific account of delusions. For example, in some cognitive neuropsychiatric accounts of delusions (see Coltheart et al., 2011; see Chapter 7, section 7.1.), delusions are conceptualized as irrational or strange beliefs that result of a series of disrupted inner computational processes (allegedly identical to whichever patterns of neural activity). Consequently, if revisionist doxasticists concede that delusions are not *really* beliefs, but just conveniently described as such, then doxasticism loses its appeal for cognitive neuropsychiatric research.

On the other hand, revisionist doxasticists cannot meet the scientific desideratum either if they make the truth of our folk-psychological belief ascriptions dependent on the ascriber's interests and standards. This move opens a conceptual gap between the revisionist's notion of belief and the kind of notion of belief that would be of use for scientists. Several authors, at both sides of the typology problem arena, have pointed out this mismatch between what the folk-psychological notion of belief –at least as revisionists individuate it– *can* do for us and what it *should* do for it to be useful for scientific research (Clutton, 2018; Gerrans, 2004; Porcher, 2018; Tumulty, 2012). If doxasticists want the doxastic interpretation of

delusional phenomena to be useful for scientific and clinical practice, the concept of belief should be reducible or translatable to purely descriptive terms, since its role is to provide nomological explanations of delusional phenomena that enable us to causally explain, predict, and control them in a precise way. However, if the truth value of belief ascriptions is relative to the ascriber's standards -i.e., if it may vary significantly from ascriber to ascriber, independently of the facts- then the revisionist's notion of belief can hardly be informative for this nomological enterprise. To use our previous examples: if the offender profiling yields a profile that points to different offenders depending on *which detective analyzes it*, then how are they supposed to find the culprit? Or, if the cartography of a certain region represents different landscapes depending on *who uses the map*, then what is the map for?

Thus, revisionist proposals cannot meet the scientific desideratum that grounds their defense of doxasticism about delusions. Several authors have claimed that the problem lies in that the folk-psychological understanding of belief is just too *vague* for scientific purposes (e.g., see Clutton, 2018; Gerrans, 2004; Murphy, 2012; Porcher, 2018). As matters stand, one might either opt for one of three different strategies to account for delusions in a more *science-friendly* way: a) rejecting doxasticism about delusions, but not their interpretation in folk-psychological terms; b) rejecting both doxasticism and the folk-psychological conceptual framework altogether; and c) redefining the folk-psychological notion of belief so that it avoids the problems of revisionism. As we saw in Chapter 5 (section 5.1.2.), antidoxasticists have typically endorsed the first two options. Clutton's cognitive phenomenological defense of doxasticism, on the other hand, is a most recent example of the first kind of strategy. In the following section, we will consider whether Clutton's proposal is capable of avoiding the problems of revisionist doxasticisms. As we'll see, although his approach is designed to meet the scientific desideratum, it falls short of the ethico-political one. Finally, in the next section, we will claim that we should retain our commitment to this desideratum, and that we cannot thus take sides with antidoxasticists, for their proposals share the same problem of Clutton's doxasticism. We'll claim that doxasticism can and must be defended, though not on the grounds of its scientific appeal, but on the grounds of its normative import.

6.2. The ethico-political desideratum: *neurobeliefs*, private biographies, and their fuzzy normative roles

As we saw in Chapter 5 (section 5.2.2.) Clutton's (2018) proposal is thoroughly committed to what he calls "scientific doxasticism", or the default doxastic conception of delusions that one can find in prevailing diagnostic manuals and scientific models. Specifically, his intention is to reflect the conceptualization of beliefs at play in cognitive models of delusions (e.g.,

cognitive neuropsychiatry), where these are explained in terms of diverse breakdowns in the mechanisms and processes that are supposed to be involved in the normal or rational production and maintenance of beliefs. Clutton thus shares with Bortolotti and Bayne & Pacherie their commitment to the scientific desideratum of doxasticism; his pledge, though, acquires a more radical overtone. For Clutton, the fact that cognitive scientists characterize delusions as beliefs is in itself a most convincing reason to maintain that delusions are beliefs. In his own words:

> [...] one main reason for thinking that doxasticism is true is that multiple, prominent cognitive scientific theories classify delusions as beliefs. This is not the kind of thing we should ignore in defending doxasticism. Indeed, many who defend doxasticism start with this as a motivating factor, or at least as *prima facie* support for doxasticism [...]. I agree, and would argue that if we want to defend doxasticism generally, we should want to defend it in a way that actually defends *scientific* doxasticism. That would provide a welcome convergence of evidence between a philosophical view and prominent cognitive and clinical psychological theories. Clutton (2018, p. 14)

This is why Clutton's newest defense of doxasticism radically departs from its older revisionist siblings in the way it tries to deal with the SCP and RC arguments against doxasticism: while the latter propose local adjustments to the functionalist and interpretivist frameworks that first gave rise to SCP and RC, Clutton rejects that our understanding of the concept of belief and thus that of delusion should be framed by neither functionalism nor interpretivism. Nor, for that matter, by any philosophical theory that contradicts the cognitive view of delusions: if cognitive scientists and clinicians say that delusions are beliefs, then no conceptual analysis shall convince us otherwise. Instead, the role of philosophers should be to provide cognitive science with a convenient story of what beliefs are; a story that delivers a notion of belief amenable for scientific reduction, and which can also accommodate delusions neatly[53].

Specifically, Clutton draws from Kriegel's (2015) cognitive phenomenological approach to belief and proposes that believing a certain proposition is just a matter of being disposed to *judge* that *p* (i.e., to "mentally assent", with a certain degree and sense of conviction, to *p*). This way, according to Clutton (2018), his proposal can avoid the "anti-realist tendencies" (p. 11) he finds in interpretivist and functionalist accounts of belief; tendencies

[53] In this sense, Clutton's proposal could be understood as some kind of discourse eliminativism (see Chapter 2, section 2.2.2.2.).

which presumably are the reason why revisionist doxasticists cannot meet the scientific desideratum. The argument is similar to the one we've explained before: if "all" that revisionist doxasticisms can provide us is some kind of pragmatic license to speak of delusions in doxastic terms or some linguistic policy for doing so, then their doxasticism is not substantive enough -or robust enough, in Clutton's terms- to secure scientific doxasticism.

6.2.1. To the rescue: a custom-made theory of belief

Clutton's diagnosis of the problems of revisionism is two-fold. Firstly, according to the author, revisionist doxasticists make the truth value of a certain belief ascription excessively dependent on facts about the agent's dispositional profile other than their own sincere reports of their mental states. In addition, they allow for the possibility that, in cases where the agent's dispositional profile yields unclear conclusions regarding the doxastic status of the agent's intentional state, the decision to ascribe a belief to the agent would be a) just a matter of communicative convenience; or b) true proper, but only in virtue of the ascriber's standards or their audience's[54].

The author's recipe to overcome the alleged limitations of revisionism is quite straightforward. Basically, his strategy can be analyzed in two steps: firstly, he proposes a theory of belief that allows him to shift the focus from the agent's dispositional profile to their own mental-state self-ascriptions as sources of evidence about the agent's mental state; secondly, he articulates a purely descriptivist and ascriber-independent semantics of belief ascriptions, where their truth or falsity no longer depends on the ascriber's standards. Let's see this in more detail.

Firstly, Clutton equates having a certain belief that *p* with having a particular cognitive-phenomenological disposition: that of mentally assenting to *p*, or feeling a certain sense of "mental affirming" towards *p* while entertaining it "before the mind's eye". He concedes dispositionalists that an agent's other behavioral, cognitive, and phenomenological dispositions might be related to the agent's belief that *p*; however, according to Clutton, they do so only in an inessential or merely contingent way. On the contrary, it's the agents' "feeling of assent" towards their internally stored representations of the world that distinguishes beliefs from other mental states. This is the "phenomenological profile" of belief, its distinctive phenomenological mark: one for which, according to Clutton, the agent is supposed to have some sort of *knowledge by acquaintance* (i.e., direct knowledge, not mediated by any kind of inferential procedure). Consequently, he shifts the weight on the kind of evidential sources

[54] Clutton himself does not consider this second possibility. In any case, it can neither provide the kind of robust or substantive defense of doxasticism that he has in mind, as we've seen.

for determining the truth of a given belief ascription: now, the agent's sincere report of their mental states would be the best source of evidence for determining their mental state, while descriptions of the agent's patterns of actions and reactions would lose importance. In fact, according to Clutton, we should typically take the agent's sincere reports of their mental states at face value.

Secondly, Clutton explicitly reduces the cognitive phenomenological disposition that characterizes beliefs (i.e., the disposition to judge that *p* when *p*-entertaining triggers obtain) to a certain neural state. Recall that, in his construal of Kriegel's (2015) account, "to be disposed to have occurrent episodes [...] is for one's neural system to be "set" in such a way that when P is considered, suggested to one, etc., one is apt to immediately judge that P" (Clutton, 2018, p. 5). This enables him to articulate an outright rejection of the revisionists' context-relative semantics of belief ascriptions: according to Clutton (2018, p. 5), "this neural state is the "truth-maker" of the disposition, the categorical grounds in virtue of which a belief ascription can be true". In other words: no contextual considerations can influence the truth or falsity of a given belief ascription, and much less the ascriber's standards or interests; if and only if the agent is in the appropriate neural state, then they believe that *p*. Only two judges can unqualifiedly determine the truth or falsity of a belief ascription: the neuroscientist, by means of their neuroimaging methods, and the agent themself, by means of their privileged access to their cognitive-phenomenological experiences.

Drawing from these two assumptions (that believing that p is just a matter of having a particular cognitive phenomenological disposition and that this in turn is identical to a certain neural state), Clutton provides cognitive models of delusions with a most pleasant account of belief; one that has been tailored from the beginning to fit the glove of cognitive scientists and therapists. Not surprisingly, his proposal is in a better position than its revisionist counterparts to meet the scientific desideratum.

As it may be expected from our discussions in previous chapters, Clutton's cognitive phenomenological approach epitomizes the kind of conception of the mind that we find utterly flawed. It's main problem from our point of view is its almost full allegiance to the Cartesian theory of mind (see Chapter 2, section 2.1.); with the exception of dualism, the *way* in which he fleshes out his "realist" view of beliefs[55] is irrevocably committed to factualism,

[55] We intendedly refer to the *way* he deploys his mental realism because, as we saw in Chapter 4, one could perfectly be a "realist" about beliefs and still avoid factualism about the mind. In other words: one can say that beliefs and other mental states *exist* without implying that mental objects and events exist "out there", independently from our mental-state ascription practices.

causalism, intellectualism, representationalism, and, most importantly for our purposes here, descriptivism.

In particular, he seems to endorse a "two-headed" descriptivist approach to belief ascription practices -one which can be both described as an internalist and an externalist kind of descriptivism (see Chapter 3, section 3.1.2.), depending on which features of his proposal we focus on. Recall that, in order to safeguard the scientific appeal of doxasticism, Clutton denies that the truth or falsity of our belief ascriptions depends in any relevant sense on the context of belief ascription. Instead, he proposes that belief ascriptions are true or false in virtue of two kinds of state of affairs: a) from a first-person perspective, the agent's cognitive-phenomenological disposition to judge that *p*; and b) from a third-person perspective, either the agent's sincere report of their experience or whichever neural state that realizes the agent's cognitive-phenomenological disposition.

This two-headed descriptivist approach to the semantics of belief ascription poses some serious problems for Clutton's doxasticism. In general terms, we already saw in Chapter 3 that the so-called "dogma of descriptivism" ultimately leads to self-defeating kinds of naturalism and normativism, which render inadequate accounts of the close link between mind and normativity (see sections 3.2.2. and 3.2.3.). Here we want to focus on how these more general problems apply to the specific case of Clutton's defense of doxasticism. In the following sections, we'll focus on how the conceptual flaws of Clutton's approach render an ethico-politically useless kind of doxasticism. The problem, as we'll see, is that his cognitive-phenomenological theory of belief fails to reflect the normative character of our actual belief ascription practices, and thus fails to capture how belief ascriptions rationalize our actions and reactions.

6.2.2. The *neurophile's* utopia: neuroscience and the end of normativity

Recall the examples from the beginning of Chapter 5. Do Red, Blue, and Green really believe what they say they believe? If we abide by Clutton's perspective, it's not exactly clear what should we answer. Should we take their respective sincere assertions of (1), (2), and (3) at face value? Or should we wait for neuroscientists to tell us definitively? Imagine that neuroscientists managed to establish a nice and clean statistically significant correlation between Red, Blue, and Green's sincere belief self-ascriptions and certain neural states (say, the brain states ROT, BLU, and VERT). What would then happen if all three sincerely informed us that they no longer feel inclined to mentally assent to their previously believed contents, but nonetheless their brains still displayed ROT, BLU, and VERT patterns of activation in the relevant circumstances (e.g., those that used to trigger such "mental affirming" in the past)? What source of evidence would Clutton recommend us to trust? Should we take a person's

sincere belief self-ascriptions at face value or should we wait indefinitely for a more "mature (neuro)science of mind"? Both options, as we will now see, render an ethico-politically useless kind of doxasticism.

We'll first consider the latter option, which seems to be the one that Clutton (2018) himself would favor. Let's assume that the long-awaited arrival of a definitively mature cognitive neuroscience has finally become a reality. In this *neurophile*'s utopia, our current folk-psychological belief ascriptions would be in principle translatable to purely descriptive reports of an agent's neural circuitry: "S believes that *p*" would thus be reducible to "S is in such and such neural states". Politically, personally, and even clinically relevant questions like "does this British professor really believe that non-native English-speaking academics have the same research abilities than their native counterparts"?, "Do I really believe that there's an afterlife?", or "Does Green still believe that a revengeful Karmic force might punish him?" would no longer unsettle us: all we would have to do is to take a look at our brains to determine whether they are in certain neural state (e.g., VERT state in Green's case). We could just go to the neurologic clinics of the future and ask for a high-resolution, full-color 3D hologram of our current *neurobeliefs* and wind up these and other related debates.

This repeatedly invoked cyberpunk picture of our future belief ascription practices raises some doubts though. To begin with, at which point exactly would we cease to trust our own folk-psychological judgements of other's mental states and start taking neuroimages at face value? If, after repeatedly establishing a correlation between a given neural state and some belief-like behaviors, we started to observe discrepancies between both, when would we start to trust the *neurobelief* ascription over the regular, folk-psychological belief ascription?

Ultimately, one might also wonder whether the replacement of our belief-talk with some kind of technocratic *neurobelief* analysis would really ease up the kind of practical worries that we encounter in cases like the ones we've discussed here. As we saw in Chapter 4 (section 4.2.2.), when we wonder whether Red really believes that the LGBT+ movement deserves our full support, whether Blue really believes that her client is displaying a racist attitude towards her, or whether Green really believes that a vengeful Karmic force might punish him, we're engaging in a primarily evaluative and regulative, not descriptive or causal-explanatory practice; we are not primarily interested in causally explaining what they have *in fact* done -or predicting what they will in fact do-, but in establishing what they *should* have done -or what they should do from now on. Can Red be considered a proper ally? Should Blue refer her client to another psychologist? Is Green accountable for his actions and decisions? These are the kind of questions that we're really interested in, for a number

of practical purposes: weighting Red's opinions in certain LGBT+ matters, counseling Blue on whether she refers her client to another psychologist, or assessing Green's decision-making capacities.

However, in the *neurophile*'s utopia, these practical questions would no longer make sense. If our belief talk is reducible to a pure description of our brain states, then these practical questions would have a straightforward answer: let's consult neuroscientists. But what exactly could neuroscientists tell us? How would they determine the normative character of our actions? There are no "correct" nor "incorrect" brain states, no moral nor epistemic merit in having our brains configured in one way or another. Neural states are not reasons upon which we decide to take a certain course of action; they're just one part of a series of natural events that cause our behavior. Thus, all that these future neuroscientists could tell us is that we're in some given neural states and that, with a certain probability, we would probably behave, cognize, or experience in certain ways. Their *neurobelief* ascriptions would totally lack the normative force of our 'commonsensical' or folk-psychological belief ascription practices.

This is why, on this first interpretation of Clutton's proposal, the cognitive-phenomenological theory of belief yields an ethico-politically useless kind of doxasticism. Recall that this desideratum draws from the idea that belief ascriptions render an agent's actions and reactions intelligible; they allow us to see them as the result of a series of *reasons*, and not just a series of *causes* -hence why they inform our judgements about the agency and autonomy of a person. By contrast, under the cognitive-phenomenological defense of doxasticism, conceptualizing delusions as beliefs doesn't render them intelligible nor unintelligible. *Neurobelief* ascriptions just lack this normative force; they don't rationalize an agent's actions, but just picture them as the result of a series of causal processes.

6.2.3. The omniscient self-biographer: sincerity and belief ascription

If we shall retain the ethico-political value of doxasticism -and we think we should-, then a theory of belief that holds belief ascriptions reducible to descriptions of neural states is not worth endorsing. However, Clutton leaves the door open for an alternative interpretation of his semantics of belief ascription, as we previously saw. Apart from neuroscientists, the other ultimate judge on which we could rely to determine the truth of a given belief ascription would be the agent themself.

Let's now consider this latter option. On this alternative, belief ascriptions are also taken to be purely descriptive statements; in this case, it's not a neural state that is described from a third-person perspective, but a recurrent inner experience that the agent themself describes: their cognitive-phenomenological experience of judging that *p* in certain relevant

occasions. On this construal of Clutton's account, the agent's sincere report of their mental states is all we need to consider when assessing the truth or falsity of their self-ascriptions. Clutton here endorses a particularly strong version of the first-person authority thesis (see Chapters 2, 3 and 4, sections 2.1.3., 3.1.1., and 4.2.4.), which in the end renders "S believes that *p*" reducible to "S sincerely claims 'I believe that *p*'".

This option, at first sight, seems to have a virtue over the *neurophile*'s account of belief: after all, we do take a person's sincere belief self-ascriptions as a normative indicator of what the agent should think, experience, or do. When we want to know what others think about a certain matter, one straightforward way to assess that is to simply ask them what they believe; looking inside their skulls will not give us that kind of information. However, do we (or should we) always and exclusively rely on our fellow's sincerity to decide whether they believe what they claim to believe? Is their sincerity all it takes to determine the truth or falsity of their belief self-ascriptions? In other words: does it always make sense to take an agent's sincere belief self-ascription at face value?

The main problem with the strong version of the first-person authority thesis that Clutton endorses is that it portrays us as some kind of omniscient narrators of our own mental lives. As we saw elsewhere, this kind of commitment leads to two inter-related problems: a) it assumes that we are infallible judges of our own mental states (see Chapter 2, section 2.1.3.); and, relatedly, b) it yields a flawed conception of normativity (see Chapter 3, section 3.2.3.).

Firstly, according to the strong version of the first-person authority thesis, it would not make sense to say that people may be sometimes misguided about what they in fact believe about a certain matter: one would always be an infallible authority regarding one's own mental states. However, this doesn't reflect our actual folk-psychological interpretative practices. Let's again consider our starting examples. On the one hand, it doesn't seem problematic to take Blue and Green's sincere belief self-ascriptions at face value; in fact, we agree with Clutton in that in these cases we should. But take Red's example instead. Is really his sincere assertion of (1) enough to grant him the belief that "the LGBT+ movement deserves our full support"? Recall that, as Bortolotti (2012, p. 45) stated, "a lot hangs on whether an individual is ascribed beliefs"; in this case, for example, Red's recognition as a proper LGBT+ political ally. Should LGBT+ activists grant him such a recognition? Many would probably disagree. And we needn't question Red's sincerity to do so: maybe he's just insufficiently educated to appropriately reckon the kind of duties or *commitments* that come with self-ascribing the belief that "the LGBT+ movement deserves our full support"; maybe he's just insufficiently trained to automatically detect or respond to situational clues that signal an

opportunity to take an appropriate action (or inhibit an inappropriate one), and still lacks the capacity to recognize situations where he or others may be in fact hindering LGBT+ people's access to equality[56].

As we said in previous chapters, it thus seems sensible to assume that, on many occasions, we might be wrong about what we think we believe -sometimes, others will be in a better position than us to know what we really think or how we feel about a certain matter. Therapy provides a good example of this: as part of the effort to make our own values or *evaluative* beliefs explicit (i.e., how we think in evaluative terms about ourselves or the world us), therapists often try to help us recognize the potential contradictions between what we claim to believe and what we in fact believe, in the light of our particular patterns of actions and reactions (see also Chapter 9, section 9.2.1.). And, if we might be wrong about what we really believe, then our sincere self-ascriptions may not always be the only nor the most important factor to determine our attitudes towards some matter. Sometimes, or even typically, this factor might be enough, such as it might be the case in Blue's and Green's examples; at other times, like in Red's case, it might be not. Clutton's cognitive-phenomenological theory of belief, in assuming the strong version of the first-person authority thesis, is unable to capture this.

But this commitment not only is conceptually flawed because it can't accommodate how our belief ascription practices in fact work. Secondly, and most importantly for our purposes here, it also leads to the same kind of self-defeating normativism that we saw in Chapter 3 (see section 3.2.3.). This is the reason why this alternative construal of Clutton's proposal neither renders an ethico-politically useful kind of doxasticism: in its commitment to the strong version of the first-person authority thesis, the cognitive-phenomenological approach is unable to account for the normative force of belief ascriptions.

In a nutshell, the problem with Clutton's approach is that it allows for the possibility of private rule-following (see Chapter 3, section 3.2.3.; Wittgenstein, 1953/1958; see also Kripke, 1982). As we've seen, interpreting an agent's activity in intentional terms allows us to view their doings as a result of a series of reasons, thus making it evaluable in terms of different correctness criteria. This, in turn, can inform our judgements regarding a person's agency, accountability, and decision-making capacities. However, for something to count as

[56] This also works the other way around: sometimes, we refuse to withdraw a given belief ascription in the light of the ascribee's *denial* of a certain belief ascription. And this is especially true in clinical settings: for example, would we stop ascribing Green the belief that an unforgiving Karmic force might punish him if he sincerely denied believing such thing already, but still complied with his "Karmic duties"? We are pretty sure that most clinicians would not. We'll come back to this issue in Chapter 8 (section 8.4.) when we discuss the "superficiality objection" against behavior analytic approaches to delusions.

a reason for acting in a certain way (regardless of whether it's a good or a bad reason), it must be *sanctionable* as correct or incorrect under a certain evaluative framework. If, as Clutton suggests, we're some kind of omniscient self-biographers, then we could *never* be wrong about our own mental states; whatever we decide that we believe is in fact what we believe.

This is to say that whatever norm an individual thinks they're following when they act is in fact the norm they're following. And *vice versa*: whatever course of action that the agent takes can be made in accordance with any norm. If we got to decide how the pawn should be moved around the chessboard, then we could call "chess" whatever game we decided to play; accordingly, whatever move that we wanted to do with the pawn would be "correct" or justified according to the ever-fluctuating rules of this strange and solipsistic chess game. Likewise, if Red decided that what logically follows from believing that "the LGBT+ movement deserves our full support" is that one should share LGBTphobic content in the social media, then that's what would follow from that belief. If at other time he decided otherwise, then the opposite would hold: believing that "the LGBT+ movement deserves our full support" could be a valid reason for both sharing and not sharing LGBTphobic content on his social media. Both courses of action could be equally "correct" or "justified" in the light of his belief self-ascription.

In the end, there would be no correct nor incorrect courses of action, since there would be no *shared* standards for determining what beliefs justify or rationalize what actions. Belief ascriptions have the capacity to inform judgements about a person's agency precisely because they point to the *intersubjective* norms that one should abide by, i.e., to the evaluative frameworks that, within a given community and form of life, determine which of one's actions are correct or incorrect. If the rules of the practice of belief ascription were settled by one and oneself alone then, properly speaking, there would be no practice nor rules that we could sensibly talk about; relatedly, if no one could ever sanction one's belief self-ascriptions, nor one's actions in the light of such self-ascriptions, then one's actions would be, by definition, intelligible or unintelligible *exclusively for oneself* -in other words: they would be neither intelligible nor unintelligible at all.

In sum, if we want to retain a concept of belief such that belief ascriptions (and self-ascriptions) are viewed as proper rationalization devices -and thus informative regarding further judgements of someone's agency or accountability- then they cannot be reduced to a mere description of what an agent sincerely claims to believe. Nor, for that matter, to *any* kind of purely descriptive statement. This is the fundamental problem of Clutton's cognitive phenomenological theory of belief: that our folk-psychological belief ascriptions, once tools

of normative inquiry, are reduced to merely descriptive reports of some given state of affairs: from a first person-perspective, a description of some inner, private ongoings; from a third-person perspective, a description of their alleged neural basis or of the agent's own sincere report. This constitutes an example of the is-ought fallacy, as we saw in Chapter 3 (section 3.2.2.). In this two-headed descriptivist account of the semantics of belief ascriptions, their normative force is either completely washed away, as in the *neurophile*'s preferred construal of Clutton's proposal, or diluted in a boundless, inescapable, and incorrigible knowledge of oneself, as in the omniscient self-biographer's interpretation. This is why Clutton's doxasticism cannot meet the ethico-political desideratum: once devoid of their normative force, belief ascriptions cannot longer inform our judgements about a person's agency or decision-making abilities.

To be fair, Clutton's main goal is not to provide an ethico-politically relevant kind of doxasticism; in fact, providing a theory of belief that accommodates the actual workings of our folk-psychological interpretative practices is just a secondary goal for him (see Clutton, 2018). As we saw at the beginning of this section, his main goal is to provide a theory of belief that accommodates scientific doxasticism, i.e., the conception of delusions at play in prevailing cognitive models of delusions. His cognitive-phenomenological theory of belief can thus be seen as part of a discourse eliminativist project (see Chapter 2, section 2.2.2.2.): one which doesn't want to reflect our actual folk-psychological practices, but to shape the notion of belief to make it fit with that of the cognitive sciences. So, the question now is: why should we care about the ethico-political desideratum in the first place? Why not just be contempt with Clutton's *redefinition* of the notion of belief, if it presumably yields a better characterization of delusions for scientific and clinical purposes?

In Chapter 7 we'll call into question the assumption that cognitive models of delusions actually are the best possible scientific and clinical approach to delusional phenomena. In fact, we'll give several reasons why retaining such commitment to scientific doxasticism may in fact hinder progress in the scientific understanding and clinical treatment of delusions. However, in the next and last section we'll first delve into why we think it's important to have an ethico-politically relevant kind of doxasticism, and how our non-descriptivist approach to the mind can help in this regard.

6.3. So… are delusions beliefs? The non-descriptivist defence

As matters stand now, it seems like maintaining a doxasticist approach about delusions is a doomed project. On the one hand, the revisionist defenses of doxasticism offered by Bortolotti (2010, 2012) and Bayne & Pacherie (2004, 2005) fall short of the scientific

desideratum; neither yields a theory of belief that can be of use in tracking down the possible causes of delusional phenomena. On the other hand, Clutton's non-revisionist, cognitive-phenomenological defense of doxasticism falls short of the ethico-political desideratum; since his revised notion of belief lacks normative force, his doxasticism cannot inform our judgements about the agency or autonomy of people with delusions.

Why, then, not take sides with antidoxasticists? As we saw in Chapter 5 (section 5.1.2.) antidoxasticists typically opt for either one of two possible strategies to account for delusional phenomena: a) rejecting doxasticism about delusions, but not their interpretation in folk-psychological terms (i.e., what we referred to as "commonsensical antidoxasticism"); or b) rejecting both doxasticism and the folk-psychological conceptual framework altogether (i.e., what we referred to as non-commonsensical antidoxasticism). As an example of the former, we focused on Currie's (2000; see also Currie & Jureidini, 2001) meta-cognitive approach, which characterizes delusions as imaginings that the person mistakes for beliefs - or, to put it differently, as *mistaken meta-beliefs*. As examples of the latter, we focused on Egan's (2008) conceptualization of delusions as *bimaginations* (i.e., as states in between beliefs and imaginings) and Schwitzgebel's (2012) more radical "in-between" approach, which characterizes delusions as cases of *fuzzy-believing* or "beliefs gone half-mad".

A first problem with some these approaches, as doxasticists have pointed out (Chapter 5, section 5.2.3.), is that they don't actually reflect our folk-psychological interpretative practices: we do seem to understand delusions in terms of beliefs (see Rose et al., 2014). But why is this relevant at all? At the end of section 6.1.3., we saw that many authors, at both sides of the doxasticism/antidoxasticism debate, have questioned that folk psychology should have anything to do with our *scientific* models of delusions (see Clutton, 2018; Gerrans, 2004; Murphy, 2012; Porcher, 2018). To be sure, we agree with this claim: there's no principled reason why folk psychology should constrain our scientific understanding of the natural causes of delusional phenomena (see also Chapter 3, section 3.2.1.). As we'll see in Chapter 8, we think that a better scientific and clinical strategy follows from Schwitzgebel's recommendation: regardless of whether delusions can be understood in terms of beliefs or not, when it comes to psychological intervention, we should just specify, for each particular case, which are the relevant patterns of action and reaction (i.e., in Schwitzgebel's terms, their particular "dispositional profile") and which causal factors maintain them.

That said, we nonetheless think that there are still reasons to defend doxasticism. Specifically, we think that a genuinely folk-psychological kind of doxasticism, such as the one proposed by revisionist doxasticists, can and should be defended. As we view it, we needn't renounce doxasticism itself, but just doxasticists' traditional claim that

understanding delusions in terms of beliefs leaves us better equipped to understand their etiology. We think that this is fully consistent with defending that at least many delusions are beliefs (or, to put it differently, that when we interpret someone as believing the contents of their delusion, these belief ascriptions are true in most cases). The kind of non-reductivist, yet compatibilist approach to the mind afforded by non-descriptivism (see Chapter 4) can help us see why these two claims (i.e., that delusions are beliefs and that this needn't be informative about the actual causes of delusional phenomena) are compatible; once we stop viewing folk psychology as some kind of pre-scientific attempt to causally explain behavior, it doesn't seem paradoxical to claim that delusions are best interpreted as beliefs and that this doesn't imply that delusions are caused by certain disruptions in some kind of otherwise well-functioning inner belief-producing mechanisms.

In this last section we'll discuss why we think that, despite the flaws of traditional defenses, the idea that delusions are beliefs is still worth endorsing. Our main goal will thus be to present a non-descriptivist defense of doxasticism; in particular, we'll defend that revisionist doxasticism, when viewed through the lenses of our non-descriptivist approach to folk psychology, is preferrable to antidoxasticism for two reasons: its ability to accommodate how our belief ascription practices actually work in the case of delusions and, most importantly, its ethico-political benefits. We'll highlight three main contributions of our non-descriptivist approach: a) that it provides revisionist doxasticisms with the kind of context-relative account of belief ascriptions that they need to secure doxasticism about delusions; b) that it offers a nuanced way to flesh out the ethico-political desideratum; and c) that it protects revisionist doxasticisms from certain self-defeating assumptions, which make it vulnerable to eliminativist tendencies.

6.3.1. Non-descriptivist doxasticism and the context-relativity of belief ascription

One main reason why we think that revisionist doxasticisms are in better shape than both antidoxasticism and non-revisionist doxasticism is that they approach a more accurate characterization of how our belief ascription practices *really* work. In this sense, we think that revisionist doxasticisms are more in line with the Wittgensteinian urge not to "think", but to "look", i.e., not to discipline our actual belief ascription practices on the grounds of some ideal theory, but to reflect how these practices in fact work (see Chapter 4, section 4.1.2.).

Let's see this in more detail. As we view it, both classical interpretivism and functionalism, on the one hand, and the cognitive-phenomenological theory of belief, on the other, yield somewhat idealistic models of our belief ascription (and self-ascription) practices. Construed as attempts to establish the exact rules that govern these practices, these

approaches seem to advance one or another *golden rule* for interpretation, e.g., rationality constraints, accordance with some folk-psychological stereotype, first-person authority, etc. (see Chapter 4, section 4.2.4.). However, as we've seen in Red, Blue, and Green's cases, we don't always *weight* the facts about a person's patterns of action in the same way. Sometimes the agent's sincere assertion about their mental states seems sufficient for most interpreters to grant a belief ascription (as it seems to be the case with Blue and Green); at other times, though, interpreters seem to demand much more, as it appears to be the case with Red. An appropriate theory of belief should thus be able to accommodate this variability.

In their revision of the interpretivist and functionalist frameworks, revisionist doxasticists get closer to do so. The key insight of revisionist proposals resides in the adoption of a context-relative approach to the semantics of belief ascription. Specifically, as we saw in section 6.1.3., what they need to secure a doxasticist understanding of delusions is a specific kind of context-relativity: one which allows for the truth value of a belief ascription to vary across ascribers, i.e., to vary according to different ascriber's standards.

From our perspective, our non-descriptivist view of the mind can provide such an approach. As we saw in Chapter 4 (section 4.2.2.), the kind of non-descriptivism that we favor draws from the idea that folk-psychological mental-state ascriptions are moves in primarily evaluative, as opposed to descriptive, language games. When we assess whether a person merits a given belief ascription, what we're doing is evaluating whether they are sufficiently compliant with the norms that determine the courses of action to be expected from someone whom is ascribed such belief. In this sense, we're like examinators that decide whether someone's patterns of action "pass" the exam, i.e., whether the agent's patterns of action reaches certain thresholds to be granted a belief ascription[57] However, contrary to idealistic models of belief ascription, these thresholds, these examining criteria, are not *given,* nor ultimately fixed for life -otherwise, belief ascriptions would still be reducible to descriptions of whatever patter of action falling under such fixed rules. By contrast, from our approach, the relevant norms to consider when deciding whether a belief state ascription is true or false may vary across contexts of assessment and, crucially, *across ascribers*. In other words: the truth value of belief ascriptions also depends on the ascribers' or interpreters' standards

[57] Contrary to sliding scale theorists though, and in Bortolotti's line, belief ascription here is not seen as "a matter of degree" (Schwitzgebel, 2012, p. 15), but as a binary decision: as examiners, we don't *rate* how much someone's performance fits certain criteria and then decide whether it's pragmatically convenient to talk of belief; we just decide whether someone passes the "doxastic exam" or not. Truly enough, this will often be a difficult decision, and depending on how high the situational demands for a clear ascription are, we might just decide to suspend our judgement; however, we won't attribute a "fuzzy-belief", which has no understandable normative force.

and evaluative frameworks, grounded on the particular "forms of life" of the communities they belong to.

Our Wittgensteinian kind of non-descriptivism thus implements the particular kind of context-relativity that revisionists need to secure doxasticism about delusions. Here, context-relativity is construed as a kind of *pluralism* about the different criteria that competent speakers follow when ascribing beliefs to each other (see Chapter 4, section 4.2.4.). On some occasions, certain features of the context will lead us to emphasize first-person authority, understood as a social, interpersonal norm (Borgoni, forthcoming, see also Almagro-Holgado, 2021; Villanueva, 2014); in other words, we might grant the truth of a given belief ascription on the sole basis of the person's sincerity. This fits Blue's and Green's cases, where it seems weird to deny that they believe what they claim to believe, no matter how "deviant" their overall patterns of action are. By contrast, on other occasions, the situation might lead us to emphasize other factors, e.g., the person's overall consistency. This fits more with Red's case, whose belief self-ascription seems to be more questionable. Villanueva's (2014) "expressivist strategy" and Almagro's (2021) notion of "contextual authority" (Chapter 4, section 4.2.4.) allow us to accommodate both cases.

We'll later come back to the benefits of this kind of contextualist approach when assessing the viability of the ethico-political desideratum of doxasticism. As of now, it offers us a more realistic picture of how belief ascription practices in fact work. In doing so, they provide us with a way to understand why we tend to interpret many cases of delusion as beliefs that doesn't rely on some "privileged access" approach to self-knowledge (such as the one that Clutton endorses). Instead, it assumes that, in cases like Green's, most people view the person's sincerity as the most relevant criteria to determine the truth of their belief self-ascription, overriding other concerns about the overall consistency of the person's actions or other possible criteria. Probably, one reason for this is that it's often difficult to find an alternative intentional (i.e., rationalistic) explanation for the person's claims: what other reason could Green have for claiming that a Karmic force might punish him if they didn't believe it? Why would he expose himself to the many possible negative repercussions of saying that (e.g., hospitalization, stigmatization, etc.) if he didn't believe what he says? This, as we'll now see, is the second major benefit of revisionist doxasticists precisely: that, unlike their competitors, they offer a way to understand delusions in intentional terms, which comes with certain ethico-political benefits.

6.3.2. 48% believing, 48% deciding? Non-descriptivist doxasticism and the ethico-political desideratum

From our perspective, the ability of revisionist doxasticisms to accommodate our folk-psychological interpretative practices in cases of delusions is not, in itself, the most important benefit of these approaches; rather, we think that the primary reason why this ability is worth preserving lies in its ethico-political strengths. As various doxasticists have stressed (Bortolotti, 2010, 2012; Bayne, 2010; Bayne & Hattiangadi, 2013), antidoxasticism leaves us in an unsettling situation regarding the agential status of people with delusions (see also Graham, 2010a; Tumulty, 2012). The problem is that the antidoxasticists' alternative characterizations of delusions in terms of folk-psychologically unfamiliar kinds of mental states -e.g., "mistaken meta-beliefs", "bimaginations", "beliefs half-gone-mad", etc.- lack normative force, and hence cannot rationalize a person's actions and reactions. This is especially so in the case of those kinds of antidoxasticism that characterize delusions as "in-between" cases. As Bayne and Hattiangadi (2013, p. 140-141) have pointed out, when we ascribe beliefs to someone,

> [...] we have some grip on the kinds of theoretical and practical behaviours it would be rational for them to engage in. But we don't have any grip on the kinds of theoretical and practical behaviours in which it would be rational for a subject to engage if their attitude to *p* is not that of belief but is merely belief-like. Talk of belief is not only in the business of providing causal explanations of behaviour, it is also –and perhaps even more fundamentally– in the business of making ourselves intelligible to each other as rational creatures. (Bayne and Hattiangadi, 2013, pp. 140-141)

This, as doxasticists have pointed out, can ultimately have undesirable ethical and political consequences regarding the way people with delusions are treated. The problem here is similar to the one we found in the *neurophile*'s reading of Clutton's cognitive-phenomenological approach (see section 6.2.2.); "bimaginations" or "half-way belief profiles" (e.g., say, -say, "48% believing") don't render a person's doings any more intelligible than "neurobeliefs". Once again, we're left wondering how to answer certain practical questions on the grounds of such unfamiliar ascriptions: should Blue refer her client to another psychologist if she just "bimagines" that her client is being racist towards her? Should Green search counseling if he just "48% believes" that a cosmic force might punish him? Ultimately, these bizarre half-way mental-state ascriptions don't serve as reasons for action, nor thus can inform judgements about a person's agency or decision-making capacities; after all, it doesn't make sense to speak of someone's "becision" to refer one's client nor someone's "48

% decision" to undergo therapy. This is where the ethico-political strengths of doxasticism reside: in that, contrary to antidoxasticists, it does provide a way to try to understand delusional experiences within the logical space of reasons (i.e., within the realm of meaning and intelligibility).

Although we're sympathetic to this argument, we think that the conceptual link between belief ascriptions and judgements about a person's status as an agent needs further development. This connection is key to motivate doxasticism on the grounds of its ethico-political virtues; however, doxasticists haven't fully fleshed it out (see Bayne & Pacherie, 2004b; Bortolotti, 2010, 2012). Some might thus want to question such conceptual link: does withdrawing a particular belief ascription to someone necessarily compromise their agential status? Take Red again, for example. If we stopped ascribing Red the belief that "the LGBT+ movement deserves our full support" on the grounds of his inconsistencies, we wouldn't need to stop interpreting him as a competent believer of many other things; we could deny his belief self-ascription of (1) without ceasing to see him as someone *generally* prone to act upon his beliefs, capable of endorsing them with intersubjectively good reasons, and hence capable of making their own decisions.

The claim that ceasing to ascribe beliefs to someone jeopardizes our capacity to view them as an intelligible and autonomous agent thus stands in need of further qualification. In reality, this is only the case when there's a *massive* breakdown in our capacity to view the agent as abided by the same norms as us, i.e., when the possibility to interpret them in folk-psychological, intentional terms is *completely* taken off the table. But this isn't what antidoxasticists are recommending to do in the case of people with delusions; what they say is just that we shouldn't view the person as a (full) believer of the content of the delusion *in particular*, not that we should massively stop ascribing beliefs to them.

Notwithstanding this qualification, we still think that doxasticists' ethico-political considerations can still be warranted by some kind of slippery slope argument. It's true that withdrawing a particular belief ascription is not the same as ejecting someone from the logical space of reasons and thus ceasing to view them as proper agents of their actions; however, it might be one first step in that direction. This is especially the case if we understand antidoxasticism as instantiating a policy according to which we should *by default* deny the truth of the person's belief self-ascriptions regarding the delusional content if systematic inconsistencies are observed. As we view it, such kind of policy would be particularly dangerous in the case of people who are already vulnerable to suffer from deagentializing practices and discourses (e.g., people suffering from structural inequalities), as it's the case of

people with mental health problems -and especially so in the case of people with psychotic experiences.

Let's see this in more detail. The point here is that assuming by default that someone is systematically wrong about their mental states -even if we circumscribe this policy to some of their mental states and to cases in which systematic inconsistencies are observed- might be one way of reinforcing these deagentializing practices, which in turn might contribute to legitimize certain forms of unjust or abusive treatment (Almagro-Holgado et al. 2021; Borgoni, 2019; Roessler, 2015). The increasingly studied phenomenon of *epistemic injustice* -i.e., a "wrong done to someone specifically in their capacity as a knower" (Fricker, 2007, p. 1)- provides a case in point (see also Kidd et al., 2017). In cases of epistemic injustice -in particular, of its *testimonial* variety-, an agent's word on diverse matters is systematically and wrongfully put into question due to some prejudice regarding their social identity, i.e., due to their class, gender, race, ethnicity, or other intersecting axes of oppression (Fricker, 2007; Kidd et al., 2017). Examples of this form of injustice are provided by women's recurrent experience of mansplaining (Manne, 2020), trans people's experience of having their trans identities negated due to cisgender and gender-binary expectations (Cocchetti et al., 2020), or the less well-off (e.g., racialized, trans, working-class, etc.) women's experiences of having their particular demands unheard or questioned by mainstream cisgender white feminisms (Srinivasan, 2021). This systematic questioning ultimately hinders the intersubjective recognition of the agent's status as a giver of knowledge -which is tantamount to damaging the agent's capacity as a knower proper, once we assume the non-descriptivist idea that there's a constitutive link between "having a mental state" (e.g., knowledge) and "being ascribed that mental state by others" (see Almagro et al., 2021). In the end, this unjust questioning of an individual's epistemic capacities also entails a questioning of their agency and decision-making abilities; after all, someone who is systematically taken to fail at knowing things can't be trusted to make reasoned and reasonable decisions (see Almagro et al., 2021; Borgoni, 2019, Fricker, 2007; Roessler, 2015; see also Kidd et al., 2017).

In our case at hand, the problem concerns the agent's *self-knowledge* abilities; by analogy, systematically denying a person's authority over their own mental states damages their capacity as "self-knowers", which in turn dampers their agential status and reinforces deagentializing practices that may be already at play. The wrongful character of this default questioning of a person's authority over their own mental states can be better appreciated when we take into account Borgoni's (2019, forthcoming) conceptualization of first-person authority as a social, interpersonal norm (see Chapter 4, section 4.2.4.): if "saying that "someone has first-person authority" [...] means that she has the right to be deferred to when it

comes to communicating her mind" (Borgoni, forthcoming, p. 16), then systematically denying a person's authority amounts to denying them that deferential treatment -the kind of treatment that we owe to those whom we consider our equals. Borgoni (2019) illustrates this particularly insidious kind of testimonial injustice with two examples. The first concerns a woman who believes that her male colleague's decision on a hiring process is misguided. In reply, his male colleague denies that she really believes that, and comments that "women just like to fuss with men's decisions". The second example concerns a slave society where slaves' mental state self-ascriptions (e.g., their desire for housing and freedom) are systematically ruled out as expressions of self-ignorance (e.g., "in reality, they don't know what's best for them"). Cartwright's creation in 1851 of a brand-new mental illness, drapetomania, to explain Black slaves' desire to flee captivity provides a case in point in this regard.

Once again, this questioning of a person's authority over their own mental states can damage the person's agency and decision-making capacities, and even more profoundly so: if people who fail systematically at knowing things can't be trusted to act reasonably, those who allegedly fail at knowing things *about themselves*, for which others normally enjoy a presumption of authority, come out as even less trustworthy in terms of their ability to act autonomously[58]. Roessler (2015) fleshes out this link between self-knowledge and autonomy as follows:

> [...] for a person to be autonomous, she needs to have a sense of self-worth and she needs to have self-knowledge. She has to fundamentally value her projects, which must also be recognized by (significant) others—this is what self-worth means. And

[58] Borgoni (2019, forthcoming) herself advocates for separating issues concerning first-person authority from questions about an individual's self-knowledge. This allows her to point out that, in her examples, women and slaves retain their self-knowledge abilities even if they're denied the deferential treatment required by the norm of first-person authority: they *know* what they believe or want, even if others wrongfully deny their authority. We think Borgoni (2019) is partially right on this point. However, this seems to put pressure on the idea that there's something specifically "epistemic" about the particular kind of unjust treatment they're subjected to (see Almagro et al. 2021). In addition, we disagree with her claim that only first-person authority, not self-knowledge, "has an attributional element" (p. 297); as we view it, there's also a constitutive link between "having (self-)knowledge" and "being ascribed (self-)knowledge by others" (see Almagro et al., 2021) and, what is more, between denying someone's authority and denying them self-knowledge abilities -after all, one cannot sensibly deny someone's authority *and* still insist that such person knows what they believe, desire, feel, etc. A way to reconcile both positions, as well as for making room for what's specifically epistemic about the kind of harm that certain people suffer, is the following: we can still recognize that those who suffer from a systematic questioning of their authority still know what their mental states are because there're *others* (members of the oppressed community, significant allies, etc.) who recognize them as authoritative and ascribe them such self-knowledge abilities. But that doesn't mean that they're not suffering a distinctively epistemic harm: insofar as their social status as self-knowers is being *generally* undermined (i.e., with regard to at least certain groups of people), their self-knowledge abilities are hindered.

> she has to know in broad terms what she wants, believes, intends, desires in order to be able to reflect on her beliefs and desires in order to find out what she autonomously believes and wants to do—this is what self-knowledge involves. As will be noted, the capacity to act autonomously on reasons of one's own is dependent on self-knowledge and self-worth. Both of these aspects play at least some role in all of the different and more substantial concepts of autonomy. (Roessler, 2015, p. 69).

In fact, Roessler (2015) points to the questioning of one's self-knowledge as one path by which standard cases of testimonial injustice (those targeting the person's knowledge of the world) produce a harm in the agency of those who suffer it; as she puts it, "the link between epistemic injustice and autonomy is to be found in the fact that epistemic injustice damages and unsettles a person's relation to herself -to her self-worth as well as her self-knowledge, both of which are prerequisites for autonomous action" (p. 68).

We aren't claiming that antidoxasticism is in itself a form of epistemic injustice; in order to claim that, the questioning of the individual's first-person authority should be based on prejudicial ideas about their social identity (Fricker, 2007); although these prejudices surely are at play in mental health contexts (e.g., Carel & Kidd, 2014), antidoxasticism is motivated by certain general theories of belief -which, as we've seen are rather idealistic and don't actually fit our interpretative practices. However, in its wrongful assumption that we should deny an individual's first-person authority by default when it comes to their delusions, antidoxasticism could unfoundedly promote an unjust deagentializing attitude towards them via the progressive undermining of their self-knowledge abilities; thus, it could end up being a potential source of injustice, even if it doesn't fit the standard definition of "epistemic injustice".

To be fair, a deagentializing attitude can sometimes come with short-term benefits, e.g., when it helps reduce blaming responses to the person's actions in their social environment. Drawing from such cases, some antidoxasticists have noted that questions about the agential status of people suffering from psychological distress are not straightforward to answer (hence, for instance, the "insanity defense" in legal contexts). Schwitzgebel (2012), for example, argues that "in many cases of delusion it *shouldn't* be straightforward to assess intentionality, and the ethical and policy applications *are* complicated, so that a philosophical approach that renders these matters straightforward is misleadingly simplistic" (p. 15). Although we agree with Schwitzgebel on this point, we think that the defense of doxasticism on ethico-political grounds is still reasonable. After all, doxasticism needn't entail that we should *always* and in every possible case attribute agency to people with delusions; it just

helps to prevent situations where such attributions of agency may be *unwarrantedly withdrawn* -for instance, by a policy consisting in denying their first-person authority by default. This is particularly important given that people with mental health problems -and specially those with psychotic experiences- are already vulnerable to unwarranted deagentializing practices. And the history of mental health institutions itself is replete with grim reminders: from the pathologizing of merely non-normative aspects of their lives to the employment of brutal treatment methods, or even the denial of their consideration as full human beings, mental health patients have been subject to multiple kinds of abuse on the grounds of their allegedly diminished decision-making capacities. Although mental health institutions have made considerable progress in this sense, unjust or abusive practices remain a sadly frequent phenomenon; the common employment of coercive measures (e.g., mechanical restraints; e.g., Fernández-Costa et al., 2020) or the many ways in which users are themselves subject to diverse forms of epistemic injustice in mental health encounters (e.g., Bueter, 2019; Carel & Kidd, 2014; Crichton et al., 2017; Drożdżowicz, 2021; Miller-Tate, 2019; Ritunnano, 2022) attest to this. To be sure, there're myriad reasons why this kind of abuses have occurred; however, it's not unreasonable to think that disowning people with mental health problems from their authority over their own minds might have a precipitating or reinforcing role.

In this sense, we think that a conceptualization of delusions that avoids this is worth preserving, and our non-descriptivist reading of doxasticism might help in that regard. From this perspective, it needn't be the case that all and every possible cases of delusions will be straightforwardly interpretable as cases of belief. If we take the notion of "contextual authority" seriously (Almagro, 2021; Villanueva, 2014), then we must proceed on a case-by-case basis: it may result that while some -even most- cases of delusions are properly understandable as beliefs, others won't be; in other words, there might be cases where the presumption of first-person authority is (rightfully) overridden by considerations about the overall consistency of the person's behavior -or the strangeness of the delusional claim. However, this is still compatible with defending doxasticism as the position that we should hold *by default*, given its ethico-political merits. The point here is that exhibiting attitude-attitude or attitude-behavior inconsistencies doesn't (nor shouldn't) always imply that we must deny the person's authority over their own minds; as Bortolotti's (2010) counterexamples of non-clinical irrational beliefs show, in many occasions we don't withdraw our belief ascriptions to people who fail at reasoning or behaving in ways in which we would expect them to. Taking this into account, and following again Borgoni's (2019, forthcoming) understanding of first-person authority, we could then rephrase doxasticism as the following

policy: "by default, don't question the person's authority over their own minds"; or, to put it differently, "by default, take the person's belief self-ascriptions at face value".

So far, we've seen how our non-descriptivist approach to the mind can contribute to defend doxasticism about delusions on the grounds of its ability to accommodate our actual interpretative practices and, most importantly, its ethico-political benefits. In the next and final section, we'll see why it can provide a better, more robust defense of doxasticism than the one afforded by revisionist approaches. The key insight will lie in its rejection of the causal-explanatory conception of folk-psychology that seems to underplay revisionist defenses.

6.3.3. Towards a more robust defense of doxasticism

In light of the above, we agree with Bortolotti (2012, p. 39) in that "the doxastic view of delusions does not tell us everything we want to know about delusions, but [...] it is at least as good as the alternative views and, in some respects, preferable". Specifically, we have defended that though it might not provide an accurate route map for research on the etiological basis of delusions, doxasticism is preferrable to other alternatives because it provides a homely account of why we straightforwardly interpret delusions in doxastic terms, and because it provides a protective policy regarding the agential status of people with delusions. What we want to argue here is that, if we're to provide a most robust defense of doxasticism, we must forego the descriptivist framework underlying its revisionist defenses.

From our perspective, the commitment to descriptivism is shared by all the contending approaches to the typology problem. As we've seen, this is particularly obvious in the case of Clutton's non-revisionist defense (section 6.2.), as well as in the case of some anti-doxasticist approaches (section 6.3.2.). But revisionist doxasticists, in framing their accounts as some sort of offender profiling enterprise, aimed at providing a nice route map for cognitive neuropsychiatric research on delusions, are also framed in a descriptivist approach to the mind. In particular, they retain a strong commitment to the standard image of folk psychology, i.e., its conceptualization as a pre-scientific, causal-explanatory theory of action (McGeer, 2007, 2021; Zawidzki, 2008; see Chapters 2 and 3, sections 2.2.1. and 3.1.1.).

This underlying conceptualization of folk psychology is expressed in an oft-repeated mantra, endorsed by both supporters and detractors of doxasticism: namely, that, since neuroscience and scientific psychology are not mature enough yet to provide a full-fledged account of the causal underpinnings of belief-like behavior, providing a nuanced account of how folk-psychological belief ascriptions work is the best thing we can do (see Bortolotti, 2012, p. 50; Clutton, 2018, p. 16; see also Schwitzgebel, 2013, p. 94). Bortolotti's (2012, p. 50) answer to Murphy's (2012) criticisms provides a case in point:

> The [...] question is what we ought to do while we wait for scientific psychology and neuroscience to hand us a new mental vocabulary. As I suggested earlier, folk-psychological notions such as beliefs, desires, and intentions pervade our understanding of minded beings and are central to the systematization of our moral intuitions. It is essential to flag inconsistencies in these notions and challenge some of their uses, but it would be difficult to do without them altogether at this stage. Murphy says: "Bortolotti's arguments [...] may not serve as a foundation for a developed science of abnormal intentional states" [...]. True. But what shall we say about delusions, self-deception, etc. while the science of abnormal intentional states reaches maturity? And how are we going to develop such a science if not by gradually revising our existing conceptual framework? (Bortolotti, 2012, p. 50).

We think that this a gravely mistaken assumption to make; one which threatens to jettison the whole doxasticist project. The problem is that it leaves the door open for eliminativist arguments like the ones we saw in Chapters 2 and 3 (sections 2.2.2.2. and 3.2.2.): in a nutshell, if our "folk-psychological notions [...] are central to the systematization of our moral intuitions", as Bortolotti (2012, p. 50) defends, but there's still a chance that these notions are nonetheless eliminated in future and more "mature" scientific accounts of behavior, then our current defense of doxasticism on ethical and political grounds is in fact a pretty weak one.

From our perspective, a more robust defense of doxasticism requires us securing its ethical and political benefits -nowadays, and in the face of any future development of the cognitive and behavioral sciences. Doxasticists would thus be better off if they definitively abandoned the underlying commitment to the causal-explanatory view of folk-psychology; it's only when we think of belief ascription practices as exercises of pre-scientific theorizing that we're tempted to think of actual scientific advancements as incompatible with folk-psychological assumptions. Our non-descriptivist, evaluativist and regulativist view of belief ascription opens up an alternative, compatibilist approach. Once we understand folk-psychological interpretation as a primarily evaluative and regulative enterprise, through which we assess each other's actions in normative rather than nomological terms, the eliminativist threat disappears: no current *nor future* description of the many causes of action could yield a plausible candidate to replace our folk-psychological mental-state ascriptions -at least as long as it doesn't convey the normative force of our interpretative practices. And so doxasticism about delusions can be properly secured: no matter how much the psychological sciences advance, as long as our folk-psychological interpretative practices don't change, the

understanding of delusions in terms of beliefs can still be warranted on ethico-political grounds.

6.4. Conclusion

In the previous chapter we saw that the main motivations behind the defense of doxasticism about delusions were two: a scientific desideratum, according to which doxasticism leaves us better equipped to develop better scientific theories and treatments of delusions, and an ethico-political desideratum, according to which doxasticism provides a way to rationalize the behavior and experiences of people with delusions, hence establishing a further barrier against undue questioning of their status as agents. In this chapter we've first focus on showing why existing defenses of doxasticism about delusions fail to meet both desiderata at once.

On the one hand, we've seen that revisionist doxasticisms fail to meet the scientific desideratum. These approaches take it that, if we succeed at providing a proper account of our belief ascription practices, we might gain a better understanding of what kind of things beliefs are; this, in turn, could be used by cognitive scientists to determine the specific causal factors involved in the production and maintenance of beliefs and belief-related behaviors and experiences. The problem lies in that, to provide a proper defense of doxasticism about delusions, revisionists must ultimately adopt a specific kind of context-relative understanding of belief ascription practices -in particular, one according to which the truth value of belief ascriptions is made partially dependent on the ascriber's evaluative framework. The resulting account of belief is therefore of little use to the science of delusions: if what counts as an instance of believing a certain content might vary across ascribers, then how is this supposed to provide a common ground for the analysis of the causes of belief-related phenomena (e.g., delusions)?

On the other hand, we've claimed that Clutton's non-revisionist approach fails to meet the ethico-political desideratum. We've first seen how Clutton advances his cognitive phenomenological theory of belief precisely in response to revisionist's alleged "anti-realist" tendencies, and in order to best serve the interests of scientific doxasticism, i.e., cognitive scientific models of delusions. On this account, to believe that p amounts to having a disposition to mentally assent to p whenever p-entertaining triggers obtain. As we've seen, this theory commits Clutton to a dual form of descriptivism; specifically, one according to which belief ascriptions are reducible to either a) an individual's description of their own cognitive-phenomenological experiences; or b) the neural states that realize them. We've seen that none of these options renders an ethico-politically useful kind of doxasticism: on the one hand, mere descriptions of a person's neural states lack the normative force necessary to

make sense of the intelligibility of a person's actions and reactions; on the other, Clutton's depiction of individuals as "omniscient self-biographers" of their own inner lives ultimately leads to a flawed conception of normativity and, again, of the normative force of belief ascriptions. Therefore, we've concluded that neither reading of Clutton's proposal yields the kind of defense of doxasticism that could inform judgements about a person's status as an autonomous agent.

Finally, we've discussed why, after all, doxasticism is worth maintaining, and how our non-descriptivist approach to the mind can promote a better defense of it. In particular, we've claimed that revisionist doxasticism, once reformulated in non-descriptivist terms, stands in better shape than its competitors; not because it provides some accurate route map for the scientific research on the etiology of delusions, but because it yields a more solid proposal in conceptual and ethico-political terms. Firstly, we've seen how Villanueva's (2014) expressivist strategy, implemented through Almagro's (2021) notion of contextual authority, enables us to accommodate why in some cases the presumption of first-person authority prevails over considerations about the consistency of an individual's overall patterns of action (e.g., cases like Blue's or Green's), and why in some other cases it doesn't (e.g., cases like Red's). This way, it offers a way to accommodate why we tend to interpret cases of delusions in doxastic terms, while avoiding endorsing a "privileged access" account of self-knowledge. Secondly, we've seen why this not only reflects how our interpretative practices work in cases of delusions, but also why they should continue to work as they do. The reason is that antidoxasticism, read as a recommendation to deny an individual's first-person authority by default when it comes to the delusional content if systematic inconsistencies are observed, might promote unwarranted deagentializing practices against people with delusions. By contrast, doxasticism can be read as a more desirable policy; one according to which we should respect an individual's authority over their mental states and thus take their mental-state ascriptions at face value by default. To conclude, we've viewed how revisionist doxasticisms, in retaining a commitment to the causal-explanatory view of folk psychology, leave doxasticism at the mercy of eliminativism. On the contrary, our non-descriptivist and compatibilist approach to the mind offers a more robust defense of doxasticism; one whereby the ethico-political benefits of doxasticism can be retained, no matter how far psychological sciences reach.

To sum up, throughout this chapter we've explored where defenses of doxasticism stand with respect to their own desiderata. In doing so, we've generally proceeded as if these desiderata were themselves justified. In the last section we've presented several reasons why we think that the ethico-political desideratum is indeed justified, why it does provide the

grounds for a proper defense of doxasticism, and how non-descriptivism can help us attain it. But what about the scientific desideratum? Is understanding delusions as beliefs necessary or beneficial for our comprehension of the causal processes involved in the production and maintenance of delusional experiences and behaviors? If it were, then some might want to ditch everything we've said in the last section and endorse instead Clutton's (2018) defense of scientific doxasticism. Forget about the ethico-political benefits of revisionist proposals; if their folk-psychological notion of belief is too vague for scientific purposes, why retain such notion? Why not embrace instead a different theory of belief -one which sidelines folk-psychological assumptions in favor of accommodating the notion of belief at play in cognitive scientific models?

The underlying premise here is that these models offer the only or the best possible approach to the science and intervention on delusions. In the following chapters, we'll challenge this assumption. In Chapter 7 we'll see that, despite their many virtues, the kind of cognitive scientific models of delusions that Clutton has in mind -i.e., those framed by traditional cognitivist assumptions- face a series of inter-related conceptual and empirical issues. Specifically, we'll point out certain limitations of the implementation of these models in clinical practice, showing how these could be partially due to their intellectualist commitments (see Chapter 2, sections 2.1.2. and 2.1.4.). In Chapter 8, we'll see how non-cognitivist approaches to mental health (in particular, behavior analytic approaches) offer certain ways to improve the quality of interventions with people with delusions.

Chapter 7
Scientific doxasticism and cognitivist approaches to delusions

In the previous chapters, we've seen that doxasticism about delusions has been traditionally endorsed on the grounds of two main desiderata: a) a scientific desideratum, related to the claim that conceptualizing delusions as beliefs presumably leaves us in a better position to account for the causal processes involved in the origin and maintenance of delusions; and b) an ethical-political desideratum, related to the claim that doxasticism leaves us in a better position to account for the intelligibility of the experiences and behaviors of people with delusions, thus promoting a further barrier against dehumanizing assessment and treatment practices. Two main lines of defense of doxasticism have been implemented. On the one hand, revisionist approaches like Bortolotti's (2010) modest interpretivism and Bayne & Pacherie's (2005) dispositionalism have attempted to provide a defense of doxasticism via a revision of the theoretical framework which antidoxasticist approaches draw from (i.e., classical interpretivism and dispositionalist functionalism, respectively). By contrast, Clutton's defense constitutes a non-revisionist approach, which rejects interpretivism and functionalism and endorses instead an alternative cognitive-phenomenological account of belief. Furthermore, his approach is not primarily a result of a conceptual analysis of the folk use of belief ascriptions in everyday contexts; instead, it constitutes an attempt to mirror the kind of notion of belief at the core of cognitive models of delusions in scientific and clinical practice -hence Clutton's insistence that his defense is not of doxasticism *per se*, but of *scientific doxasticism* in particular.

As we saw, revisionist approaches fall short of their professed scientific desideratum, since they're ultimately forced to endorse some kind of relativist approach to belief ascription, according to which the truth value of belief ascriptions may vary across assessors and contexts of assessment. On the other hand, although Clutton's non-revisionism is

specifically designed to solve this problem, his cognitive-phenomenological defense ends up falling short of the ethical-political desideratum of doxasticism, since it renders a normatively inert conception of belief. Alternatively, we've argued that our pragmatist and non-descriptivist approach to folk-psychological interpretation can provide a better framework for defending doxasticism about delusions. In particular, it allows us to a) accommodate the fact that delusions are typically understood in terms of beliefs despite the attitude-attitude and attitude-behavior inconsistencies displayed by people with delusions; b) retain doxasticism's ethico-political value, i.e., that it provides a conceptual barrier against undue de-agentializing practices; and c) protect doxasticism from potential eliminativist arguments while at the same time respecting the autonomy of scientific psychology in relation to folk psychology.

At this point, however, advocates of the kind of scientific doxasticism that Clutton has in mind might reply something along the following lines: why should we care if cognitive scientific models of delusions provide an ethico-politically informative kind of doxasticism or not? Why shouldn't we just do away with folk psychology, even if the resulting kind of doxasticism is ethico-politically useless? If what science tells us about delusions doesn't match our folk-psychological assumptions, then so much the worse for folk psychology; and if our cognitive scientific models of delusions doesn't provide what we want it to provide in ethico-political terms, then well, why shouldn't science prevail over "feelings"? We might just accept Clutton's scientifically-informed *revision* of the folk notion of belief -as a discourse eliminativist might suggest (see Chapter 2, section 2.2.2.2.)-, and let go the ethical-political desideratum in favor of the scientific desideratum.

A possible response -probably foreseeable by now- is the following: from our non-descriptivist point of view, as we've seen, there's no necessary tension between "what science tells us about delusions" and our folk-psychological interpretative practices; such tension only arises when one thinks of folk psychology as a kind of pre-scientific theory which scientific psychology is to accommodate, refine, or abandon. If cognitive scientists can make use of some sort of revised notion of belief (e.g., one redefined in terms of whatever functional state they find useful to appeal to in their causal explanations of behavior), then so be it; what non-descriptivists just point out is that such revised notion, devoid of all normative force, won't amount to a "belief" proper, i.e., a belief in its genuine, folk-psychological, ineliminable, and irreducible sense (see Curry, 2020 for a similar argument).

But here we want to go further than that. To begin with, it's questionable that the notion of belief at play in many cognitive scientific theories can be told apart from the folk-psychological one. If it were possible, then the truth-conditions of folk-psychological belief

ascriptions and those of "cognitive-scientific" belief ascriptions should be independent from each other; in other words: whether one can be ascribed a certain folk-psychological belief (e.g., the belief that one's dissertation will never come to an end) should be independent from whether one can be ascribed some particular cognitive-scientific belief (e.g., a hypothetical brain state mediating one's perception of the laptop's screen and one's feelings of despair). Rather, what scientific doxasticists like Clutton seem to have in mind is that when folk psychologists understand some delusions in terms of beliefs, such belief ascriptions must ultimately refer to some fact, internal to the person, which the cognitive (neuro)scientist is in a best position to investigate. Hence "beliefs" and similar notions (e.g., cognitive schema), as they're used in prevailing cognitive scientific models, are best understood as modelled on an intellectualist, representationalist, and computationalist construal of the folk-psychological notion, rather than pointing to some fact independent from our folk-psychological interpretative practices.

What we want to stress here is that this *traditional cognitivist* construal of the notion of belief, as we'll refer to it, is not only unable to meet the ethical-political desideratum, but may also be pernicious from a scientific point of view. The reason, as we've repeatedly pointed out, is that modelling scientific theories on mindreading conceptions of folk psychology can hinder progress in the analyses of the causal processes underlying target behaviors. The main goal of this chapter will be to show that this might be precisely the case of cognitive models of delusions. In particular, we'll argue that intellectualist assumptions at the core of these models might unduly restrain the range of variables analyzed and intervention methods considered, hence impairing progress in the intervention with people with delusions.

The structure of the chapter is as follows. In section 7.1., we'll introduce two main cognitivist approaches to the understanding of delusional phenomena: cognitive behavioral therapy for psychosis (CBTp) and cognitive neuropsychiatry. We'll assess their theoretical framework, mainly focusing on the putative causal factors that these approaches posit to explain delusions. In sections 7.2. and 7.3., we'll analyze some of the main empirical and conceptual shortcomings of scientific doxasticism, respectively. We'll first conduct a narrative review of the empirical evidence on the efficacy of CBTp, paying special attention to the available data of its efficacy in the case of delusions. Then, we'll see which are, from our point of view, the main theoretical problems of these approaches. We'll contend that some of the problems of cognitive models of delusions may be anchored in their underlying intellectualist view of the mind. This intellectualism poses a groundless, pre-empirical assumption about what must be the actual causes of delusional phenomena, which may constrain

research and intervention and hence preclude the development of more efficacious procedures. Finally, in section 7.4., we'll draw the main conclusions of this chapter.

7.1. Traditional cognitivist approaches to delusions

Cognitive models have long shaped the understanding of delusions and the range of possibilities considered regarding their assessment and treatment. In particular, cognitive behavioral therapy (hence CBT; see Chapters 1 and 2, sections 1.3.2. and 2.3.3.), in the field of clinical psychology, and cognitive neuropsychiatry, in the field of psychiatry, have set the standard cognitivist understanding of delusional phenomena.

These approaches can be seen as somewhat complementary cognitivist accounts of delusions: the former draws from and contributes to the development of cognitive models of delusions to understand the psychological processes underlying their development, maintenance, and treatment; the latter, in turn, tests these cognitive models and investigates the neural etiology of the information processing deficits that they posit to explain delusional phenomena (e.g., Bell et al., 2006; Blackwood et al., 2001; Coltheart et al., 2011; Garety et al., 2007). In any case, both can be jointly characterized by their *traditional cognitivist* understanding of the mind. The term "traditional cognitivism" is sometimes used by postcognitivist scholars (see Chapter 2, section 2.2.2.1.) to identify the more standard approaches to cognitive science that they target (i.e., the representationalist, computationalist, and internalist information-processing theories of cognition that rose during the 1970's) (e.g., see Menary, 2010; Newen et al., 2018). Here we'll use this term to characterize the kind of cognitive models of delusions at play in both cognitive neuropsychiatry and CBT. To be sure, there are some differences between the conceptual foundations of these approaches. For instance, while cognitive neuropsychiatry has clear roots in cognitive neuropsychology and cognitive neuroscience (e.g., Coltheart, 2007; David & Halligan, 1996, 2000; Ellis & Young, 1990), the development of CBT was only partially informed by information-processing models of cognition; psychoanalytic theories were another major pillar at the beginning -although these were subsequently abandoned or reintegrated in information processing accounts of psychopathology (Beck, 1952, 1963, 1964; Ellis, 1958; see Chapter 1, section 1.3.2.). Nonetheless, as we'll see, contemporary cognitive models of delusions result from an integration of early CBT theories (e.g., Alford & Beck, 1994) with cognitive models that highlight specific information processing deficits as the key cognitive factors in the development and maintenance of delusions (e.g., Freeman & Garety, 2006); these, in turn, constitute the roadmap used by many cognitive neuropsychiatry researchers to determine the hypothesized underlying

neural etiology (see Coltheart et al., 2011; Langdon et al., 2008). Therefore, we've deemed the label "traditional cognitivism" appropriate to jointly refer to these approaches.

On the one hand, as we saw in Chapters 1 and 2 (sections 1.3.2. and 2.3.3.), a core premise of CBT is that mental health problems are ultimately due to certain failures or impairments of normal or psychologically-supportive modes of cognitive processing (e.g., Dobson & Dozois, 2010; Knapp & Beck, 2008); on this view, the individual's representations of reality play a fundamental mediational role in the origin and maintenance of psychological distress. Against this background, delusions have been traditionally conceptualized as "beliefs [involving] severe cognitive dysfunction which leads to negative (harmful) consequences; simply put, delusions are maladaptive cognitive constructions of internal or external phenomena" (Alford & Beck, 1994, p. 370).

According to Kingdon & Turkington (1991), Beck's (1952) intervention with a person with a persecutory delusion constituted the first attempt to use "reasoning techniques" in the treatment of delusional thought. During the 1980's and the 1990's, the employment of cognitive theories and procedures, which had been previously developed for the treatment of "maladaptive" modes of reasoning in other mental health problems (e.g., depression; see Beck, 1979), began to be also applied to the treatment of delusional beliefs and other psychotic phenomena, thus establishing the grounds for the subfield of cognitive behavioral therapy for psychosis (CBTp) (see Alford, 1986; Alford & Beck, 1994; Chadwick & Lowe, 1990, 1994; Garety et al., 1997; Kingdon & Turkington, 1991; Kuipers et al., 1997; Naeem et al., 2015; see also Johns et al., 2014; McLeod, 2009; Upthegrove, 2018). The general procedure involved in the implementation of CBTp is briefly described in Naeem et al. (2015) as follows:

> CBT for psychosis focuses on establishing links between thoughts, emotions and behaviours and by challenging dysfunctional thoughts. It challenges delusions using Socratic dialogue and dealing with hallucinations and beliefs underlying the hallucinations. It also uses normalisation techniques as well as behavioural techniques to reduce distress and improve functioning. The key elements of CBTp include: engaging the patient, collaboratively developing a problem list, and deciding on a clear goal for the therapy session. Once the goal had been decided on, a CBTp technique would be used (e.g., guided discovery and Socratic questioning) to identify distortions in thinking style. This would be followed by an agreed task (homework) for the patient to complete by themselves before the following appointment (e.g., attempting to identify these distortions over the next week and trying to correct them). Regular feedback and asking the patient to provide a capsule summary (i.e. personal understanding) of the session are also crucial elements. A formulation (narrative of the person's history) is jointly generated

> to make sense of the emergence and maintenance of the problem at hand. (Naeem et al., 2015, p. 4)

Although a shared general premise of CBTp is that "dysfunctional" reasoning patterns are at the core of delusional experiences, early CBTp interventions didn't draw from a clear set of hypotheses regarding the exact kinds of cognitive deficits presumably at play in cases of delusions, nor thus they targeted key cognitive mechanisms specific of this kind of mental health problem (Alford & Beck, 1994; see also Mehl et al., 2015[59]). At most, early CBTp approaches assumed that delusions were dysfunctional beliefs which, like other dysfunctional cognitions (e.g., depressive beliefs), resulted from certain cognitive distortions and maladaptive schemas -i.e., negatively biased ways of thinking about oneself, the world, and the future (e.g., negative self-schemas; see Alford & Beck, 1994). In fact, CBTp has often been advanced in defense of the continuity thesis (Chapters 1 and 5, sections 1.3.1. and 5.2.1.3.), i.e., the idea that the difference between clinical and non-clinical cognitions, behaviors, and experiences is more a matter of degree rather than sharp categorical difference (see Bentall, 2003; Chadwick & Lowe, 1990, 1994; Mehl et al., 2015).

By contrast, throughout the 1990's and the 2000's, more specific cognitive models of delusions began to appear. These models hypothesized specific information processing deficits to play a key role in the formation and maintenance of delusions (e.g., Freeman et al., 2002; Garety et al., 2001). Freeman & Garety (2006) provide an overview of possible factors involved in the development of delusions which includes over 20 different variables. Here we might distinguish between non-specific and specific factors. The former typically include environmental conditions and disruptive perceptual, affective, and interpersonal experiences that may have a role in the development of delusions, but which are not themselves specifically cognitive (i.e., information processing) deficits; these might include sleep disturbances, stressful hallucinations, excessive worry and ruminative thoughts, interpersonal sensitivity, etc. (e.g., Alford & Beck, 1994; Freeman et al. 2002; Freeman & Garety, 2006, 2014; Garety, 1991; Garety et al., 2001, 2007; Garety & Freeman, 1999). By contrast, specific factors typically involve properly cognitive mechanisms or *reasoning biases* that are hypothesized to be specifically disrupted in people with delusions. Traditionally, these have included: a) a heightened tendency to "jump to conclusions", i.e., to prematurely endorse certain beliefs on the grounds of limited or inconclusive evidence; b) attributional biases, in particular a bias towards greater externalization and personalization of the causes of negative events,

[59] This article originally contained inconsistencies in the data reported and conclusions drawn. For a criticism of the original article and a corrigendum by the authors, see Laws (2016) and Mehl et al. (2019), respectively.

which would presumably explain why some people with delusions tend to misattribute disruptive internal experiences to external agents; and c) disrupted Theory of Mind (ToM) abilities, i.e., disruptions in the person's ability to make correct judgements about other's mental states (see Bentall, 2003; Brakoulias et al., 2008; Freeman, 2007; Freeman et al. 2002; Freeman & Garety, 2006, 2014; Frith, 1992; Garety, 1991; Garety et al., 2001, 2007; Garety & Freeman, 1999, 2006; Kaney & Bentall, 1989; Kinderman & Bentall, 1997; Mehl et al., 2018). Although inconsistencies in the evidence supporting the presence of jumping-to-conclusions, externalizing attributional style, and ToM biases in people with delusions have been pointed out (e.g., see Brakoulias et al., 2008; Diez-Alegría et al., 2010; Freeman & Garety, 2014; Mehl et al., 2014, 2018), it's still widely accepted that at least some of these reasoning biases play a major role in the origin and maintenance of delusions.

Finally, in relation to the negative schemata originally hypothesized by CBTp approaches to underlie delusions, Garety et al. (2001) assume that these reasoning biases "occur against a conducive social-cognitive background", whereby disruptive events in one's social environment "may create an enduring cognitive vulnerability, characterized by negative schematic models of the self and the world (e.g. beliefs about the self as vulnerable to threat, or about others as dangerous) that facilitate external attributions and low self-esteem" (p. 190); in other words, these heightened cognitive biases would both derive from and in turn feed back to the negative cognitive schemas hypothesized by CBTp to be at the root of delusional thinking. Along these lines, more recent CBTp interventions, referred to as "causal-interventionist approaches to CBTp" (Lincoln & Peters, 2010; Mehl et al., 2015; see section 7.2.), have been designed to target these hypothetical specific and non-specific factors in order to enhance therapeutic effectiveness and efficiency (e.g., Foster et al., 2010; Freeman et al., 2014, 2015).

On the other hand, drawing from the same traditional cognitivist theoretical framework, cognitive neuropsychiatry has focused during the last three decades on determining the neural underpinnings of the information processing deficits allegedly involved in different mental health problems (see Coltheart, 2007; Coltheart et al., 2011; David & Halligan, 1996, 2000; Ellis, 1998; Ellis & Young, 1990; Frith, 1992; Halligan & David, 2001; Hohwy & Rosenberg, 2005; Langdon et al., 2008; McKay, 2012; Stone & Young, 1997; Young, 1999). In particular, cognitive neuropsychiatric approaches to delusional phenomena have had a prominent role in the emergence and establishment of the "psychiatry-as-applied cognitive neuroscience" motto that characterized the origins of third-wave biological psychiatry during the decade of the brain (i.e., 1990's). As we saw in Chapter 1 (see section 1.5.1.), the grounding assumption of these approaches is that models of normal cognitive and neural

functioning could be deployed to understand the specific information processing deficits involved in mental health problems, which in turn could help to establish the specific neural biomarkers behind such deviations. As David & Halligan (2000) put it:

> What marks [cognitive neuropsychiatry] as distinctive within medical specialisms is the explicit move beyond diagnosis and classification, toward offering a *cognitive explanation* for the disorder and, as an important second, *location of the brain systems* responsible. The characteristic feature of cognitive neuropsychiatry, however, remains its theoretical approach: using patterns of impaired and intact cognitive performance found in patients to inform and revise current models of normal cognitive functioning. (David & Halligan, 2000, p. 506)

In the case of delusions, the main goal of cognitive neuropsychiatry thus is to develop causal explanations of delusional experience and behavior by importing the conceptual framework and tools from cognitive neuroscience and cognitive neuropsychology. Typically, cognitive neuropsychiatric research has primarily focused on the analysis of monothematic delusions (i.e., delusions involving one specific belief content or a relatively small and closely inter-related system of beliefs, such as Capgras delusion, Frégoli delusion, mirrored-self misidentification, etc.), arguably due to their more straightforward neurobiological origin (e.g., see Coltheart 2007); however, the field has later expanded to the analysis of polythematic delusions (i.e., delusions about diverse and not necessarily related topics, like many paranoid delusions) (e.g., Langdon et al., 2008).

Ellis & Young (1990) were among the first to provide a cognitive neuropsychiatric theory of several delusional misidentification syndromes. To take just one example, these authors suggested that the case of Capgras delusion could be properly analyzed and understood as "a mirror image of prosopagnosia" (p. 244); specifically, whereas cases of prosopagnosia are typically characterized as involving an impaired "ventral route" to facial recognition (causing an inability to semantically or cognitively recognize others' faces) paired with an intact "dorsal route", involved in the affective recognition of others' faces (which would explain why people with prosopagnosia fare better in recognizing familiar faces), cases of Capgras delusions would presumably involve the reverse pattern of impairment: i.e., an intact ventral or semantic-recognition route and an impaired dorsal or affective-recognition route, which would explain why people with Capgras delusions are unable to identify familiar faces as belonging to affectively close people, consequently adopting "some sort of rationalisation strategy in which the individual before them is deemed to be an imposter" (p. 244).

Since these foundational analyses, cognitive neuropsychiatric theories have been proposed for different delusions, although controversies remain as to the number of factors that are assumed to be involved in the development of delusional thinking. In parallel to the aforementioned distinction between non-specific and specific factors, there's a common distinction in the literature between *one-factor* and *two-factor* cognitive neuropsychiatric theories or delusions (see Coltheart et al., 2011; Maher, 1974/2005). On the one hand, one-factor theories typically assume that delusional beliefs are the result of just one kind of anomaly; typically, an anomaly in the perceptual or affective processing of incoming sensory or sensory-motor information that results in bizarre or disturbing phenomenological experiences (Maher, 1974/2005). Thus, the characteristic assumption of one-factor approaches is that delusions just involve an impairment in the experiential, not in the reasoning domain; delusions do not result from impaired reasoning processes, but instead are rational responses to abnormal or bizarre experiences. In this sense, Maher's (1974/2005) work was often cited by early CBTp approaches (Alford & Beck, 1994; Chadwick & Lowe, 1990, 1994; see also Bentall, 2003; Mehl et al., 2015) to defend the continuity thesis, or the idea that psychotic experiences and beliefs are on a continuum with more typical non-clinical phenomena.

By contrast, two-factor theories typically assume that some or a combination of the above-mentioned reasoning biases (e.g., increased tendency to jump-to-conclusions, attributional biases, disrupted ToM, etc.) reflect the presence of specific impairments in cognitive or metacognitive processing within the clinical population (e.g., a diminished ability to manage one's cognitions in a self-regulatory manner) (e.g., see Coltheart et al., 2011; Davies & Egan, 2013; McKay, 2012). Supporters of the two-factor approach typically claim that one-factor theories can only account for the emergence of delusional contents, but not for the endorsement of such contents or the maintenance of the delusional belief over time. In other words, two-factor theories assume that while the experiential factor may partially account for why certain ideas might come to one's mind, a second, specifically cognitive factor is needed to account for why the person systematically fails to update their beliefs in the face of overwhelming amount of counterevidence or seemingly obvious contradictions between the delusional claim and other beliefs endorsed by the person. Thus, two-factor theories assume that delusions are the result of a dual impairment: one affecting the processing of sensory information, and another one affecting "normal" reasoning processes. Nonetheless, two-factor supporters differ with regard to the specific kind of cognitive or meta-cognitive impairments that are supposed to originate and maintain delusions, e.g., a failure in the initial evaluation of the plausibility of the delusional claim, a failure in the update of existing beliefs in the face of new contradictory evidence, a failure in the meta-cognitive processes

regulating the good integration among existing beliefs, some particular combination of these, etc. (see Davies & Egan, 2013).

It's easy to see how the cognitive-phenomenological theory of belief (Clutton, 2018) fits the cognitivist picture behind CBTp and cognitive neuropsychiatric approaches to delusions. As we saw in Chapter 5 (see section 5.2.2.), this theory states that "beliefs are dispositions to have certain intentional, occurrent mental states whose phenomenal character is that of "judging that P" and that "S believes that P iff S is disposed to immediately judge that P when P-entertaining triggers obtain" (Clutton, 2018, p. 4). On the one hand, this definition perfectly matches the theoretical assumptions of many CBT approaches, according to which a major factor in the origin and maintenance of psychological problems would be the presence of a series of underlying cognitive schemas (i.e., dispositions) that would prompt repetitive negatively-valenced thoughts (i.e., the P's) and their automatic judgement as true, due to the presence of certain cognitive distortions or reasoning biases that would in turn feed back to the maladaptive schemas. On the other hand, Clutton's (2018) identification of this cognitive-phenomenological dispositions with certain neural realizers -the "truth-markers" of the disposition (p. 5)- matches well the main assumption of cognitive neuropsychiatry, i.e., that the cognitive architecture of mental health problems, as informed by traditional cognitivist models, can be mapped out on the "relevant brain structures and their pathology" (Halligan & David, 2001, p. 209). In the next sections, we'll see some of the main empirical and conceptual problems of such a conception of belief and the traditional cognitivist assumptions it draws from.

7.2. Current evidence for CBTp

In Chapter 5, we saw why Clutton's (2018) cognitive-phenomenological theory of belief was unable to accommodate the ethical-political desideratum of doxasticism. Nonetheless, one might still want to side with the cognitive-phenomenological view, and insist that its scientific virtues -namely, its ability to neatly accommodate traditional cognitivist approaches to delusions- override other concerns regarding its ethical or political benefits. Here we want to stress two critical considerations against this view, namely: a) that the presumed scientific virtues of this kind of doxasticism are not so clear, as the available evidence suggests; and b) that the conceptual framework of scientific doxasticism and the cognitive-phenomenological theory of belief may unnecessarily constrain the possibilities of intervention with people with delusions, obscuring potentially relevant variables and partially explaining the ambiguous evidence regarding the efficacy of CBTp interventions. In this section, we'll focus on the former, leaving the discussion of the latter for section 7.3.

Since the first long-term randomized control trial assessing the effectiveness of CBTp (see Garety et al., 1997; Kuipers et al., 1997), multiple randomized control trials and meta-analyses have been carried out to study the absolute and relative efficacy of CBTp (see Jauhar et al., 2014; Jones et al., 2018a, 2018b; Lincoln & Peters, 2018; Lynch et al., 2010; Mehl et al., 2015, 2019; Stiko et al., 2020; Turner et al., 2014, 2020; Van der Gaag et al., 2014; Wykes et al., 2008; Zimmerman, 2005). The efficacy of this intervention has been assessed both in isolation and in combination with pharmacotherapy, as well as compared to both passive control groups (e.g., Treatment-As-Usual or standard care control groups), and active controls (e.g., pharmacotherapy, other psychological interventions, etc.). In addition, several studies have analyzed the absolute and relative efficacy of CBTp for different symptom groups (i.e., positive vs. negative symptoms), as well as for specific symptoms within each group (e.g., hallucinations vs. delusions). Additionally analyzed variables have included the different forms of delivery (e.g., individual vs. group CBTp, whether the intervention was case formulation-based or not, whether it employed a causal-interventionist approach or not, etc.), the methodological quality of the trial designs (e.g., whether they controlled for blindness allocation or not), etc.

There remains substantive controversy regarding whether CBTp alone or in combination with other forms of care is effective or not. On the one hand, several meta-analyses have yielded positive results regarding the efficacy of CBTp (Jauhar et al., 2014; Mehl et al., 2015, 2019; Naeem et al., 2016; Stiko et al., 2020; Turner et al., 2014, 2020; Van der Gaag et al., 2014; Wykes et al., 2008; Zimmerman, 2005). This led in 2009 to the recommendation of CBTp as a first-line treatment for people with a diagnosis of schizophrenia in the NICE (2009) guidelines, subsequently remaining as an add-on treatment in the NICE (2014) recommendations. However, effect sizes have been typically found to range from small to moderate, as well as to vary widely depending on the kind of control procedures employed, with low-quality methodological designs (i.e., those not controlling for allocation blindness and with a high risk of bias) typically yielding higher effect sizes (e.g., Jauhar et al., 2014; Turner et al., 2014; Wykes et al., 2008; see also Lincoln & Peters, 2018; Mehl et al., 2018; Stiko et al., 2020). Furthermore, some meta-analyses have yielded negative results regarding the overall efficacy of CBTp. For example, Lynch et al. (2010) found no significant differences between CBTp and non-specific interventions in relapse prevention and reducing symptoms of schizophrenia when allocation blindness was controlled for. In addition, two Cochrane reviews (Jones et al., 2018a, 2018b), didn't found CBTp in combination with standard care to be more efficacious than standard care alone (Jones et al., 2018a) nor standard care in combination with other psychosocial treatments (Jones et al., 2018b).

Although these critical meta-analyses have been highly contested on the grounds of certain methodological limitations (e.g., Hutton et al., 2014), the controversy regarding the efficacy of CBTp interventions remains. This controversy is especially acute in the specific case of the efficacy of CBTp for the treatment of delusional phenomena (see Freeman, 2011). In this sense, the small-to-moderate effect sizes of CBTp interventions seem to be primarily due to its efficacy in the management of hallucinations, where the evidence seems to be more stable, rather than delusions, which case is far less clear (see Lincoln & Peters, 2018, Turner et al., 2014, 2020; van der Gaag et al., 2014; see also Upthegrove, 2018). On the one hand, some studies converge in claiming that CBTp has a stable small-to-moderate effect on delusions (Lincoln & Peters, 2018; Mehl et al., 2015, 2019; Naeem et al., 2016; Turner et al., 2014, 2020; van der Gaag et al., 2014), and some have even found that effect size has increased with time (Stiko et al., 2020). A recent meta-analysis by Turner et al. (2020) yields an even more optimistic view. The authors carried out both conventional and cumulative meta-analyses to assess the efficacy of CBTp for hallucinations and delusions as well as its evolution over time to determine whether the evidence base of CBTp was already sufficient. They concluded that "the existing evidence base for the effect of CBTp on hallucinations and delusions is both statistically stable and sufficient", suggesting that "further RCTs repeatedly testing CBTp are unlikely to have a significant impact on the magnitude or significance of treatment effects or to alter our conclusions in any substantive way" (p. 10).

However, on a closer look, many of these studies reveal a not-so-rosy picture. Turner et al. (2020) themselves note that in their conventional meta-analysis "CBTp did not demonstrate superiority for delusions compared with active controls in the context of low power" (pp. 11). This negative result added to those of previous studies where they had found that the effect size of CBTp for positive symptoms in general lost statistical significance when researcher allegiance was controlled for (Turner et al., 2014), and that the effect size of individually tailored formulation-based CBTp for delusions in particular was only significant when compared to standard care, but not when compared to active treatment (Van der Gaag et al., 2014). Similarly, Mehl et al. (2015, 2019) found no statistically significant increase in the effect size of CBTp interventions on delusions when compared with standard care at follow-up, nor when compared with other interventions at either end-of-therapy or follow-up.

Acknowledging these inconsistencies, some have suggested different ways to enhance the efficacy of CBTp interventions overall and for delusional phenomena in particular. In this sense, several authors have insisted on the necessity of adopting a *symptom-based*, *formulation-based* and *causal-interventionist* approach to the implementation of CBTp procedures (see Freeman, 2011; Garety & Freeman, 2013; Lincoln & Peters, 2018; Mehl et al. 2015,

2019; Morrison et al., 2004; Van der Gaag et al., 2014). In other words, these authors encourage CBTp practitioners to: a) address specific problems rather than broad diagnostic categories; b) tailor intervention to the specific case, basing "therapeutic judgements on a careful appreciation of their patients' history and circumstances" (Morrison et al., 2014, p. 6) and c) target the above-mentioned specific and non-specific processes putatively involved in the development and maintenance of delusions (see section 7.1.). Freeman (2011) sums up his approach as follows:

> How can CBT for psychosis move forward? The approach recommended here is to follow three principles: reduce the heterogeneity of psychosis by focussing on single symptoms; use developments in the theoretical understanding to guide therapy; and show that putative causal processes have been changed, in what has been termed an interventionist–causal model approach. (Freeman, 2011, p. 135).

Some initial indicators of the enhanced therapeutic power of this kind of strategies have been provided by Lincoln & Peters (2018) and Mehl et al. (2015, 2019), who noted that more recent CBTp trials taking a causal-interventionist approach seem to yield increased effect sizes. These promising initial results encourage the adoption of the kind of precision-based and case-centered approach suggested by these authors.

However, it's not at all clear that the enhanced efficacy of causal-interventionist implementations of CBTp is actually due to the targeting of the specific mechanisms or processes hypothesized by cognitivist theories to mediate the development and maintenance of delusions (e.g., reasoning biases, negative cognitive schemas, etc.). Several studies show that the effect size of CBTp interventions on delusions doesn't seem to be mediated by change in these putative proximal factors. Brakoulias et al. (2008), for instance, assessed jumping-to-conclusions, attributional biases, and ToM both before and after therapy with CBTp. They found that, although CBTp had a significant effect in reducing preoccupation and conviction, it didn't have any effect over jumping-to-conclusions nor attributional biases, with the results for ToM being inconsistent. Later on, Garety et al. (2015) conducted a "proof-of-concept experiment" to assess whether a brief computerized reasoning training intervention targeting jumping-to-conclusions and belief flexibility would improve these reasoning biases and delusional thinking, and whether the improvement in the latter was mediated by the former. Although they found a significant effect of the intervention on both paranoia and reasoning, they found no significant evidence of a mediating effect once they controlled for potential baseline confounders. Similarly, regarding the putative causal role of negative self-schemas,

Freeman et al. (2014) didn't find any significant effect of CBTp neither on paranoia nor on negative cognitions about the self.

More recently, Mehl et al. (2018, p. 153) have pointed out that the putative causal roles of neither attribution nor ToM biases seem to be supported by the available evidence, leaving only jumping-to-conclusions and negative self-schemas as possible cognitive mediators - although, regarding the latter, they state that "there is no study that demonstrates that reducing negative self-schemas has an effect on delusions, despite several attempts to find this" (p. 153). In order to assess whether therapeutic change on delusions was mediated by change on these three reasoning biases (i.e., jumping-to-conclusions, externalizing attributional style and ToM bias) and on negative self-schemas (i.e., implicit and explicit self-esteem), Mehl et al. (2018) conducted a secondary mediation analysis. Although they found individualized CBTp to be effective for the treatment of delusions, they found no significant difference between pre- and post-treatment measures for any of the reasoning biases nor for explicit self-esteem; only implicit self-esteem changed throughout therapy, but neither this nor the rest of hypothesized causal factors mediated the effect of the CBTp intervention on delusions.

As we can see, the evidence supporting the cognitivist understanding of delusions and their intervention is rather equivocal (see also Upthegrove, 2018). Although there seems to be some degree of consensus regarding the overall efficacy of CBTp, its efficacy on the particular case of delusions is more controversial. Moreover, the traditional reasoning biases and negative self-schemas traditionally hypothesized by cognitive models to explain the development of delusions don't seem to mediate the efficacy of CBTp. As we view it, the equivocal evidence basis of traditional cognitivist models of delusions (i.e., scientific doxasticism) casts serious doubts on the presumed virtues of these models and thus on the claim that doxasticism can or must be defended on the grounds of its scientific virtues. In the following section, we'll see how these problems may at least partially arise from the kind of intellectualist view of the mind characteristic of traditional cognitivism.

7.3. Intellectualism, the straitjacket of psychological intervention

Adding to the empirical ambiguities of traditional cognitivist approaches to delusions, we think that scientific doxasticism also has a number of conceptual difficulties, mainly due to its Cartesian commitments. Let's assume Clutton (2018) is right in that his cognitive-phenomenological theory of belief provides an accurate description of the picture of the mind behind many cognitivist approaches to delusions. If that is so, then cognitivist models of delusions can be primarily characterized by a dual commitment to both internal and external

descriptivism, since having a certain belief is a matter of both a) entertaining certain mental objects before "the eyes of the mind", and b) having one's neural system set up in certain ways; see Chapters 3 and 6, sections 3.1.2. and 6.2.). In addition, these models buy almost all of the Cartesian framework; except -arguably- the commitment to dualism, they fully embrace a) factualism (i.e., that beliefs are some kind of factual entity); b) mental causalism (i.e., that beliefs cause certain types of behavior); c) intellectualism (i.e., that to have a certain belief or to act in accordance with it is a matter of entertaining certain "regulative propositions" or "inner instructions" in the mind -either understood as the brain or one's "private-phenomenological realm"- and then acting accordingly); and d) representationalism (i.e., that our knowledge of the world and of other minds is necessarily mediated by a "veil of representations" of the world, and that we have some kind of immediate or privileged access to our own mind).

As we've seen throughout the preceding chapters, these philosophical commitments are problematic because they lead to a series of conceptual puzzles, e.g., self-defeating forms of naturalism, untenable views of the knowledge of the external world and of others' mental states, etc. (see Chapters 2, 3, and 4). In the particular case of doxasticism about delusions, we've already seen that a doxastic approach drawing from these conceptual commitments is unable to live up to the ethical-political desideratum. Here we want to point out that these conceptual commitments are also pernicious from a clinical and scientific point of view, and might partially explain some the ambiguous empirical support for the efficacy of CBTp interventions on delusions and the cognitive models attempting to explain their development.

From our perspective, the main problem with the conceptual framework of cognitivism, and with the commitment to intellectualism in particular, is that it constrains research about the causes of target behavioral patterns of interest and neglects the causal role of different environmental sources of control. As we saw in previous chapters, the intellectualist legend amounts to the idea that any given instance of intentional or goal-directed behavior owes its normative character to some anterior internal operation (Ryle, 1949/2005, p. 20). The entire traditional cognitivist framework is built around this idea: that performing an action voluntarily, in a goal-directed manner, always involves the anterior manipulation of inner regulative propositions representing the different steps to successfully complete whatever task at hand, and that the correct or incorrect character of such action is given in terms of how these inner rules were managed and applied.

In the clinical field, this is expressed in an unwarranted over-emphasis on verbal sources of control, especially in hypothesized forms of inner verbal-cognitive control. Cognitive models of delusions assume that what is wrong about delusional experiences and

behaviors lies in the hypothetical presence of certain maladaptive representations of the world, oneself and others (i.e., negative cognitive schemata) maintained by certain information processing biases or vicious meta-cognitive operations (see Chapters 1 and 2, sections 1.3.2. and 2.3.3.). Thus, if two people present the exact same pattern of non-normative behavior (e.g., claiming things like "I believe a vengeful Kharmic force is blackmailing me into ordering my shelf in an exact rectangular manner", picking up every cigarette bump and throwing it to the bin, etc.), cognitive models assume that these patterns must ultimately be due to the same inner cognitive causes (i.e., inner disrupted or negative self-schemas, reasoning biases, etc.). Specifically, CBTp draws from the assumption that certain inner maladaptive schemas give rise to repetitive and automatic negative thoughts via certain cognitive distortions or information processing deficits, and that these automatic negative thoughts are the primary proximal cause of verbal and non-verbal delusional behavior. In turn, cognitive neuropsychiatric models draw from this same assumption to try to map out these internal deficits to whatever atypical pattern of brain activity.

To begin with, this assumption entails a certain degree of circularity: it explains problematic covert and overt behavior (e.g., delusional statements, repetitive negative thoughts, performance in different tasks measuring reasoning biases, etc.) in terms of certain internal cognitive or meta-cognitive deficits, while at the same time taking those same behaviors as evidence of the presence of such internal deficits. Furthermore, even if we assume that certain reasoning styles (e.g., jumping-to-conclusions) are significantly more common in people with delusions than in other populations, that still doesn't mean that these alleged factors explain the origin and development of delusions; they are themselves part of what ought to be explained (see Stewart et al., 2016).

A more important objection to the intellectualist assumption at the core of cognitive models of delusions comes from the multiple realizability argument (see Chapters 2 and 3, sections 2.2.2.1. and 3.2.1.), or the idea that the same observed patterns of behavioral, inferential, and phenomenological activity can be realized via different causal processes in different species, individuals, or even moments of time for a given individual. As we'll see in more detail in the next chapter, functional analytic interventions with people with delusions based on a Functional Behavioral Assessment (hence FBA) provide empirical evidence for this in the clinical field (see Froján-Parga et al., 2019). These approaches have consistently shown that topographically similar patterns of verbal behavior (e.g., those related to the same type of delusions) may be maintained by different environmental contingencies in different people (e.g., contingencies of positive reinforcement, negative reinforcement, Pavlovian conditioning, etc.) or even different moments in time. In other words: what is typically

understood as explained by common internal variables (e.g., an "irrational" belief, a jumping-to-conclusions bias, a negative self-schema, etc.) was shown to be maintained by completely different environmental factors in different cases.

This has important implications for the assessment of the efficacy of psychological interventions, in particular of CBTp. As we view it, one of the factors that may explain the ambiguities regarding the efficacy of CBTp interventions on delusions, as well as regarding the mediational role of the hypothetical proximal mechanisms proposed by cognitive models, is its traditional overemphasis on hypothetical common internal factors and its neglect of differential, directly testable sources of environmental control. Specifically, this might be due to two possible reasons; namely, that on the assumption that delusional experiences and behaviors are originated and maintained by maladaptive schemas and reasoning biases, CBTp practitioners a) restrain the range of intervention tools to verbal-cognitive techniques (e.g., cognitive restructuring via Socratic dialog, etc.) dismissing the potential therapeutic efficacy of alternative behavioral techniques (e.g., reinforcement of alternative responses, exposure, behavioral activation); and b) overlook the importance of assessing the environmental contingencies maintaining target behaviors in each individual case, which leads to a manualized application of therapeutic techniques that isn't guided by considerations regarding the actual maintaining variables in each case (Froxán-Parga, 2020).

The latter case is even more worrying than the former. Consider Green's example again. Imagine that after conducting a functional assessment of Green's claims about the vengeful Karmic force extorting him, we find out that it's primarily maintained by others' attention -the most common positive reinforcer found in the studies reviewed in Froján-Parga et al. (2019). If that were so, then initiating a collaborative reason-giving exchange to test the truth or coherence of Green's thoughts (see Alford & Beck, 1994), might not constitute the best intervention strategy (provided that the intervention goal were "problem reduction", which might not always be the case; see Chapter 8, section 8.3.). In fact, it could well be counterproductive, for the reason-giving exchange itself may become a source of positive reinforcement for Green's claims. It could also be the case that the factors controlling Green's atypical claims are in fact different from those maintaining Green's non-verbal behaviors (e.g., trash-collecting and shelf-organizing); maybe these are controlled by negative reinforcement contingencies, such as the removal of Green's own automatic negative thoughts about his partner when he complies with his Karmic duties. If that were the case, then focusing exclusively on modifying Green's thoughts and claims on the assumption that these necessarily are the root cause of his non-verbal problematic behaviors would clearly be a mistake.

Finally, this has obvious implications for research on the neural basis of mental health problems; when we assume that topographically similar patterns of behavior are necessarily due to similar internal processes, we are overlooking the possibility that they are actually maintained by completely different environmental contingencies in different cases, which will probably map out to different patterns of neural activity sustaining such behaviors. In this sense, the seemingly circular character of cognitivist explanations of delusions is not only conceptually flawed, but may also dampen the nomological power of scientific explanations of delusional phenomena and of interventions with people with delusions. This is a clear example of what Skinner (1974) viewed as "the major damage wrought by mentalism", i.e., that "when what a person does i[s] attributed to what is going on inside [them], investigation is brought to an end" (p. 19).

From our perspective, the grounding mistake of scientific doxasticism lies in the assumption that the psychopathological character of certain behaviors and experiences must be necessarily explained by reference to the allegedly maladaptive functioning of some otherwise well-functioning hypothetical inner mechanism. As we view it, neither intelligent performances "inherit all [their] title to intelligence from some anterior internal operation of planning what to do" (Ryle, 1949/2009, p. 20), nor disruptive, bizarre, challenging, or "mad" doings inherit their "mad character" from a disruption in some putative internal computation process. Our categorical, inherently normative distinctions between "psychopathological" and "non-psychopathological" behaviors and experiences are ultimately grounded on particular, not always shared evaluative frameworks (Fulford 2011; Fulford & Van Staden, 2013; Thornton, 2007, 2014); as such, they may be useful for certain ethical, political, clinical and scientific purposes, but they do not necessarily carve the nature of mental health problems at its joints, nor thus map out the causal processes that may explain them in different cases.

Relatedly, from our non-descriptivist approach to mental-state ascriptions (see Chapter 4), neither does our assessment of delusional phenomena in terms of beliefs. Once again, to assess a relevant behavioral pattern in terms of beliefs is not merely to describe those same patterns using some "ascriptive shorthand", but neither it is to describe some inner ethereal or neural cause of such behaviors; it is to view them under the light of certain social norms. In this sense, our non-descriptivist approach also avoids the problem of circularity that cognitivist models face: when we infer that someone has a certain belief from their overall patterns of behavior (broadly construed), we're not pointing to some hidden extra fact for which we have no other evidence than the observed patterns of behavior themselves; we're just assessing the person's behavior as logically (not causally) connected with

certain conceptual commitments: those that determine how the person should have behaved or should behave from now own. In other words: the connection between behavior and belief is not causal, but logical or "grammatical". Finally, as we saw in previous chapters, our non-descriptivist approach also respects the autonomy of scientific psychology in relation to folk psychology (see Chapters 3, 4, and 6, sections 3.2.1., 4.2.3., and 6.3.). Thus, whether we decide that delusions are correctly understood as beliefs or not needn't have a relevant impact on our capacity to establish their biological and environmental causes. Our non-descriptivist approach leaves it open for scientists and clinical practitioners to determine which are the relevant causal factors involved in the origin and maintenance of mental health problems, delusions included, regardless of whether these fit our folk-psychological interpretative practices or not.

7.4. Conclusion

In the previous chapters, we saw that the cognitive-phenomenological defense of doxasticism about delusions (Clutton, 2018) didn't allow for a proper defense of the ethical and political desiderata of doxasticist approaches. In this chapter, we've seen that its underlying claim to the scientific and clinical superiority of traditional cognitivist approaches to delusions might not hold either. As we saw, the cognitive-phenomenological theory of belief was explicitly designed to fit scientific doxasticism, i.e., the actual doxastic conception of delusions at play in traditional cognitivist approaches to delusions, such as traditional CBTp and cognitive neuropsychiatry. The cognitive models of delusions underlying these approaches draw from the assumption that delusions are beliefs somehow gone-wrong which ultimately result from a series of negative cognitive schemas representing the agent, the world around them, and the future. These negative cognitive schemas foster and in turn are fed back by a combination of a) non-specific factors, such as excessive worrying, and b) specific factors, including a number of cognitive distortions or information processing deficits like jumping-to-conclusions (i.e., a tendency to arrive to a conclusion on the light of insufficient evidence), attributional biases (e.g., an externalizing and personalizing attributional style for negative events) and ToM deficits (i.e., an inability to form accurate representations of others' minds).

Here we've seen that these cognitive models face a number of empirical and conceptual objections. Firstly, the evidence supporting the absolute and relative efficacy of cognitive interventions with people with delusions is ambiguous. In addition, we've seen that there's evidence against the putative causal role of the specific cognitive factors that are supposed to explain both the development of delusions and the efficacy of CBTp interventions; of the above-mentioned hypothetical cognitive causes, only jumping-to-conclusions and

negative self-schemas seem to be significantly increased in people with delusions, and none of the relevant factors has been consistently found to mediate the efficacy of CBTp interventions on delusions, not even when these interventions are specifically designed to target those reasoning biases and negative cognitive schemas.

Secondly, there are also conceptual reasons to reject scientific doxasticism. Other than the general problems associated with the kind of descriptivist, factualist, causalist, intellectualist, and representationalist commitments underlying traditional cognitivism in general, there are conceptual problems that specifically attain to cognitive models of delusions. These mainly derive from its commitment to intellectualism, or the idea that to believe that *p* or to act in accordance with such a belief is a matter of previously entertaining such proposition in the mind (or brain) and then acting accordingly. To begin with, cognitivist explanations of delusions exhibit certain degree of circularity, since the evidence for the existence of the hypothetical cognitive deficits involved in delusional thinking comes from the very same patterns of behavior that these putative causal factors were supposed to explain. In addition, cognitive models face the problem of multiple realization, i.e., that different causal realizers in different individuals or even moments in time may cause the exact same behavior topographies. FBA-based interventions with people with delusions (see Froján-Parga et al., 2019) constitute a clinically relevant example of this. Drawing from this, we've suggested that the intellectualist view of the mind at play in cognitive models of delusional phenomena might partially explain the ambiguous evidence supporting CBTp interventions on delusions. The core problem would be that intellectualism leads to an over-emphasis on the role of verbal and covert behavior in the explanation of delusional phenomena, which would lead to a) an excessive focus on the use of cognitive techniques (e.g., Socratic dialog) at the expense of other procedures; and, more importantly b) the "manualization" of psychological interventions, ultimately dismissing the importance of assessing the environmental contingencies maintaining target behaviors in each individual case.

We can thus conclude that the cognitive-phenomenological theory of belief, as well as the scientific doxasticism it stands for, are not only a dead-end if our goal is to offer an ethico-politically informative conceptualization of delusions; they may be also pernicious for scientific and clinical purposes. We agree with those in favor of adopting an individually-tailored, formulation-based, and causal-interventionist approach to psychological intervention with people with delusions; however, we think that cognitive models, as well as the kind of view of the mind underlying them, may not provide the best framework for that purpose. In the face of these limitations, we think that an alternative, non-cognitivist (i.e., non-computationalist, and non-representationalist) approach is worth exploring. In Chapter 8, we'll

see how early and contemporary functional-analytic approaches (see Chapter 1, sections 1.3.1. and 1.5.2.) may provide a more appropriate framework for a different kind of individually-tailored and causal-interventionist approach to interventions with people with delusions.

Chapter 8
Functional analytic approaches to delusions

In the previous chapter we've seen how scientific doxasticism, or the conception of delusions at play in traditional cognitivist approaches, may negatively impact assessment and treatment. Specifically, it might lead us to overlook how environmental contingencies maintain target behaviors and to neglect the employment of therapeutic techniques aimed at modifying such environmental contingencies. The main problem with these approaches is that they adopt some variety of what Hurley (2001) called the "sandwich model of cognition", according to which the interaction between an agent and the environment is necessarily mediated by information processing mechanisms (i.e., hypothetical cognitive mediators of the relation between perception and action), charged with the task of forming, storing, and manipulating accurate representations of the outer world (see Chapters 1 and 2, sections 1.3.2. and 2.2.2.1.). Consequently, environmental variables are seen as distal causes of behavior, conceding explanatory primacy to hypothetical proximal cognitive mediators.

As matters stand, alternative *non-cognitivist* approaches to mental health are worth exploring. Drawing from a common rejection of the representationalist and computationalist view of the mind, the most important common features of this kind of approaches are a) *non-reductivism*[60] (i.e., a focus on the organism-environment system as the core unit of analysis in psychological research, which is assumed to be inexplicable by appealing to hypothetical inner states); and b) *non-representationalism* (i.e., the assumption that it's not necessary to posit inner hypothetical representations of an outer world to explain behavior)

[60] As we saw in Chapter 2 (see section 2.2.2., footnote 19), the term "reductivism" and its counterpart "non-reductivism" have been used in a number of different ways. Here we're using "non-reductivism" in the sense that these approaches reject the possibility of reducing descriptions of the dynamic interactions between an organism and the environment to the language of hypothetical inner functional or neural mechanisms.

(e.g., Chiesa, 1994; de Haan, 2020a, 2020c, 2021; Hayes, 2016; Moore, 2008; Nielsen & Ward, 2018, 2020; Skinner, 1963, 1974, 1977; Sturmey, 2020; see also Barrett, 2019). Historically, the most representative exemplar of a non-cognitivist approach to the philosophy of psychology has been radical behaviorism (Skinner, 1945, 1953, 1974), as well as similar or derived philosophical views related to the experimental and applied analysis of behavior (see Hayes, 2016, 2021). More recently, certain strands of postcognitivism within the cognitive sciences (namely enactivism -see Chapters 1 and 2, sections 1.5.3., 2.3.6.), have encouraged a somewhat similar approach to the study of cognition and behavior (e.g., de Haan, 2020a, 2020b, 2021; Nielsen, 2021; Nielsen & Ward, 2018, 2020). However, while behavior analysis has a long history of applications and innovations in the clinical field and has yielded some of the most effective methods of assessment and intervention, enactive approaches to mental health are still in their infancy. These approaches have made more emphasis on core conceptual issues (e.g., the integration problem) than on pointing out yet unnoticed relevant causal variables or yielding new specific treatment procedures (although see, for example, Röhricht et al., 2014).

Therefore, functional analytic approaches to mental health (Chapter 1, sections 1.3.1. and 1.5.2.) constitute the most solid non-cognitivist option in the clinical field to date. The main goal of this chapter will thus be to provide a narrative review of functional analytic approaches to the intervention with people with delusions, as well as to analyze their conceptual underpinnings and related strengths and weaknesses. In particular, we'll focus on the two main "strands" or "tendencies" within the functional analytic approach to mental health that we saw Chapter 1 (sections 1.3.1. and 1.5.2.): a) "traditional" behavior analysis (see Hayes, 2016), which draws from a more "orthodox" or prevalent understanding of the main tenets of radical behaviorism; and b) Acceptance and Commitment Therapy (ACT), a "post-Skinnerian" functional analytic approach, whose scientific and philosophical framework is now based on Contextual Behavioral Science (CBS) and functional contextualism.

The structure of the chapter is as follows. In section 8.1., we'll introduce some of the main shared features and differences between these two functional analytic strands. In sections 8.2. and 8.3., we'll review the conceptual framework and efficacy of each approach to the intervention with people with delusions and other psychotic experiences. In section 8.4., we'll see how these approaches, despite their explicit rejection of the Cartesian view of the mind, still seem to carry with them a somewhat residual endorsement of some of its core defining commitments -namely, intellectualism- which may limit their efficacy and perceived utility. In consequence, we'll explore how our pragmatist kind of non-descriptivism may help to overcome these limitations and attain a sounder functional analytic approach to

the intervention with people with delusions. Finally, in section 8.5. we'll present the main conclusions of this chapter.

8.1. Common features

Let's first recap the main features of functional analytic approaches to mental health. As we saw in Chapter 1 (sections 1.3.1. and 1.5.2.), these approaches share a series of core conceptual commitments, both of a general character and specific to the clinical field. To begin with, they are primarily characterized by a certain understanding of pragmatism and the adoption of a pragmatic attitude towards the science of behavior (see Cooper et al., 2019; Hayes, 2021; Moore, 2008). Beyond theoretical considerations, functional analytic approaches prioritize the effective intervention upon the behavior of individuals and groups -hence they set as their ultimate goals the prediction and control of behavior. In addition, rather than assuming a mechanistic, *structuralist* view of behavior (such as the one deployed by traditional cognitivist approaches), functional analytic approaches adopt a *functionalist* perspective on psychological phenomena (Sturmey et al. 2020). Recall that, in this context, "functionalism" and "structuralism" have a different meaning from their use in the cognitive and social sciences, respectively (see Chapter 1, section 1.3.1.); in a nutshell, what "functionalist" conveys here is the assumption that the basic unit of analysis is the "organism-environment system", as enactivists put it, and that behavior is to be primarily explained in terms of the three types of variation and selection: a) natural selection, operant selection, and cultural selection (Skinner, 1953, 1981, 1990; see also Alonso-Vega et al., 2020, Cooper et al., 2019; Sturmey, 2020).

Thus, what is of interest to functional analytic researchers is not the topography of a given pattern of behavior itself (i.e., its physical properties, frequency, duration, etc.), no matter how bizarre it may be, but the functional relation between a given pattern of activity and the environmental contingencies that control it. "Behavior" is thus understood as a relational notion, which refers to such functional relations, and which encompasses both "overt" and "covert" responses (e.g., motor activity, but also mental imagery, inner speech, and so on). Overall, a properly psychological analysis involves the explanation of behavior in contextual terms -although different approaches differ as to what the "context" might include. What explains a given responding pattern is how it's functionally related to the antecedent environmental events that may elicit or evoke it and how it may in turn be selected by the environment (Skinner, 1953, 1957, 1974, 1981, 1990; see Chiesa, 1994; Cooper et al., 2019; Froxán-Parga, 2020; Hayes, 2016, 2021; Moore, 2008; Sturmey, 2020).

Applied to the clinical setting, functional analytic approaches share three core assumptions: firstly, that mental health problems essentially are constellations of behaviors,

broadly construed, which, due to their scant or excessive frequency, their contextual inappropriateness, or their association with individual or social distress, are negatively valued within a social-linguistic community of reference; secondly, that the variables that originate and maintain such non-normative[61] behaviors are fundamentally the same that originate and maintain any other kind of behavior -no "deficit" or "internal mechanism-gone-wrong" view of psychopathology is assumed; and thirdly, that psychological assessment and intervention should generally proceed on a case-by-case basis, analyzing the exact environmental variables that may control each behavior of interest in each particular case and arrange the intervention accordingly. In this sense, functional analytic approaches to mental health endorse a strong view of the continuity thesis (see Chapters 1 and 5, sections 1.3.1. and 5.2.1.3.). For functional analytic practitioners, there's no need to appeal to any kind of deficit or internal process gone-wrong to account for psychopathological behavior. Psychopathological behaviors do not necessarily reflect an underlying disruption of otherwise well-functioning cognitive or neurobiological processes, nor information processing failures which are presumably ubiquitous across both clinical and non-clinical populations; rather, functional analytic approaches assume that psychopathological behaviors can be produced and maintained by the exact same operant and classical conditioning processes that produce and maintain non-clinical behaviors (see Ayllon & Haughton, 1964; Ayllon & Michael, 1959; Ferster & DeMyer, 1962; Froxán-Parga, 2020; Hayes et al., 1999, 2001; Layng & Andronis, 1984; Lindsley, 1963, 1964; Rosenfarb, 2013; Skinner, 1953, 1977; Sturmey, 2020; Wong, 1996, 2006, 2014; Wilder et al., 2020).

It follows from these three core characteristics is that diagnostic labels and other general mental health categories are deemed as inappropriate tools for analyzing the actual causes of each individual's mental health problems. To be sure, these categories may be useful for a number of other important purposes: among others, they might help foster interprofessional communication, organize the distribution of administrative resources, provide people with hermeneutical tools to understand their experience, or even help to articulate political struggles against the diverse oppressions suffered by people with non-normative psychological make-ups (e.g., see Chapman 2020; Singer, 1999). The core point of functional

[61] "Atypical" has been more commonly used. We've preferred to use "non-normative" for a number of reasons. Firstly, because, as many authors have pointed out, psychotic experiences are actually more common among non-clinical populations than it's widely assumed (see Bentall, 2003). Secondly, because we think that although "atypical" aims to express a non-pathologizing attitude, it still conveys the "statistically deviant" motto behind certain self-styled naturalist approaches to mental health (e.g., Boorse, 2014; Kendall, 1975). By contrast, we think that "non-normative", while capturing the non-pathologizing attitude that "atypical" intends to convey, still hints at the essentially value-laden nature of mental health judgements.

analytic approaches is just that diagnostic labels are poor guides towards the analysis of the actual causes of each individual's mental health problems. No matter whether one receives this or that diagnosis, the analysis of the causes of mental health problems should ideally stem from an idiosyncratic analysis of the environmental contingencies maintaining each person's problems; in other words, intervention should ideally be preceded by a *functional analysis* or *functional assessment* of behavior (see Chapter 1, section 1.5.2.1.), the characteristic functional analytic method for case formulation.

Functional analytic approaches thus offer a non-cognitivist variety of the kind of individually tailored, formulation-based, and causal-interventionist approach to psychological intervention encouraged by some CBTp researchers, as we saw in Chapter 7 (see section 7.2.). However, instead on hypothesizing particular cognitive deficits in each individual case, functional analytic practitioners focus on the analysis of the particular environmental conditions in which each individual lives and how these may be maintaining target patterns of interest. Stewart et al. (2016) provide a case on point of this kind of approach to the analysis of delusions:

> (…) [Contextual Behavioral Science (CBS)] makes no appeal to mediating mental mechanisms. Instead it explains behavior in terms of environmental events and does so by identifying functional relations between (past and present) environment and behavior. To illustrate, consider [jumping-to-conclusions (JTC)] and [persecutory delusions (PD)]. Although the correlation between JTC and PD is interesting, CBS researchers would not stop there. They would also investigate what this or other related patterns of behavior are a function of (i.e., what history of learning and environmental factors give rise to and maintain JTC). Once identified, these environmental variables can be manipulated to exert influence over the behavior of interest. For instance, by reinforcing efforts to seek out additional information before drawing conclusions we may improve performance on probabilistic reasoning paradigms and influence JTC in (real-life) situations. Hence, in CBS the study of delusional beliefs deemphasizes mental mediators (e.g., JTC, attentional biases) and instead searches for environmental moderators (e.g., antecedents and consequences that give rise to and maintain delusional behaviors). Stewart et al. (2016, pp. 238-239)

Within this broad common framework, we can distinguish two main "strands" of functional analytic approaches, as we saw in Chapter 1: "traditional" behavior analysis and the post-Skinnerian, functional-contextualist approach that characterizes Acceptance and Commitment Therapy (ACT). Although both share a lot in common, the differences between more "orthodox" radical behaviorists and the more "heterodox" functional contextualists

has eventually leaded to a split between them two (see Hayes, 2016, 2021; de Rose, 2021). The origin of these differences can be traced back to Hayes et al.'s (2001) formulation of their "post-Skinnerian" theory of human language and cognition, Relational Frame Theory (RFT). As we saw, RFT aims to certain complex behaviors (e.g., symbolic behavior) in terms of verbal or relational responding and the formation of "relational frames" (i.e., equivalence and non-equivalence relations); more traditional behavior analytic approaches reject the explanatory necessity of relational responding and its consequences for the analysis of behavior.

However, to really grasp the difference between both approaches one needs to understand the different clinical contexts and problems that each was primarily applied for. Traditional behavior analytic approaches were historically circumscribed to the sphere of in-patient settings (e.g., Ayllon & Michael, 1959; Ayllon & Haughton, 1964; Lindsley, 1956, 1964), and still today seem to be most widely applied in contexts where the control over environmental contingencies is maximal and where the problematics addressed are clearly operationalizable (see Beavers et al., 2013). By contrast, one of the main motivations behind RFT was to account for clinical change following psychotherapy or "talk therapy" in outpatient contexts, whereby behavioral changes are: a) mostly produced through verbal interaction inside the clinical context (vs. direct manipulation of environmental contingencies in the extra-clinical context); b) transferred from in-session to extra-clinical contexts; and c) often sustained in the face of competing contingencies. In this sense, as we saw, the core idea behind the post-Skinnerian strand was to reformulate in behavioral terms key cognitivist notions (i.e., mental representation), developed in the first place to fill the explanatory "gaps" in behavior changes following psychotherapy (see Chapter 1, sections 1.3.2. and 1.5.2.2.).

Due to these important differences, we'll here consider them separately. In the next sections, we'll see how these functional analytic approaches have been applied to the intervention with people with delusions. We'll first review traditional behavior analytic interventions in section 8.2., leaving the discussion of ACT interventions for section 8.3.

8.2. Traditional behavior analytic interventions with people with delusions

As we saw in Chapter 1, (sections 1.3.1. and 1.5.2.2.), traditional behavior analytic approaches to mental health primarily explain clinical changes in terms of respondent and operant conditioning processes, as classically defined. In this sense, "traditional behavior analytic approaches to mental health" encompasses those applications of the experimental analysis of behavior to the clinical field that are grounded in the philosophy of radical behaviorism (see Skinner, 1974; see also Chiesa, 1990, Moore, 2008) or in a seemingly orthodox understanding

of it (see Hayes, 2016, 2021). Among other characteristics, this more traditional approach is typically characterized by a) a stricter preference for single-case experimental designs (vs. group designs) in maximally controlled settings; and b) a sharper preference for the direct analysis of behavioral change following the manipulation of environmental contingencies vs. theorizing and modelling of hypothetical mediational variables (see Skinner, 1950; see also Staddon, 2021).

These stricter methodological standards characterize the work of both traditional experimental and applied behavior analysts. In the field of mental health, traditional behavior analytic interventions have typically taken place in inpatient settings and other clinical contexts that allow for a maximum control over the environmental contingencies potentially controlling the individuals' behavior (see Hayes, 2016). They are also characterized by the employment of single-case designs (preferably experimental) to assess both the functions of target behaviors and the efficacy of the intervention (see Beavers et al., 2013). Contemporary forms of this kind of behavior analytic intervention ideally draw from a Functional Behavioral Assessment (FBA; see Chapter 1, section 1.5.2.1.) of target behaviors, whereby the possible or actual environmental contingencies maintaining them are laid out, and which constitutes the basis of the subsequent intervention procedures; when this is the case, the intervention is called an "FBA-based intervention" (e.g., Hurl et al., 2016). In this sense, some take this more traditional variety of functional analytic approach as constituting a most profound rejection of the medical model; one which not only rejects the assumption that psychological problems are medical phenomena in the stronger sense (i.e., neurobiological diseases at root), but also the characteristic methodology of the medical model, i.e., the employment of group comparisons (e.g., randomized control trials) and meta-analyses of the results of such group comparisons to measure the efficacy of psychological interventions (see Wilder et al., 2020; Wong, 1996, 2014) Instead, traditional behavior analytic practitioners take seriously the need for tailoring the intervention to each individual case and to assess the efficacy individually, relating it to the control of the basic classical and operant conditioning processes that presumably explain therapeutic change (Cooper et al., 2019; Froxán-Parga, 2020; Sturmey, 2020).

Drawing from these philosophical and methodological principles, traditional behavior analysts operationalize most psychological problems as patterns of behavior that, for some reason, are negatively evaluated within a given social and cultural context (Sturmey, 2020). In addition, they assume that such behavioral patterns are primarily maintained by environmental contingencies and at least partially modifiable by introducing changes in those environmental contingencies. Delusions and other psychotic phenomena are no

exception. From this framework, psychotic experiences have been typically conceptualized as patterns of non-normative verbal behavior (whether overt or covert) which are maintained by the same operant and respondent processes that explain any other kind of behavior (see Burns et al., 1983; Layng & Andronis, 1984; Lindsley, 1963, 1964; Mace et al., 1988; Mace & Lalli, 1991; Rosenfarb, 2013; Salzinger et al., 1964; Skinner, 1936, 1957; Sturmey, 2020; Wilder et al., 2020; Wong, 1996, 2006, 2014). In Lindsley's (1964) terms, "to a behaviorist a psychotic is a person in a mental hospital. If psychosis is what makes, or has made this person psychotic, then psychosis is the behavioral deviation that caused this person to be hospitalized, or that is keeping [them] hospitalized" (p. 232).

This traditional behavior analytic conception of psychotic phenomena draws from the kind of straightforward eliminativist approach that characterizes behavior analysis -at least in its more orthodox understanding (see Chapter 2, section 2.3.2.). From this perspective, delusions and other psychotic phenomena are operationalized in terms of non-normative verbal behavior because this enables their inter-rater register, their experimental control, as well as the direct observation of their variation following the rearrangement of environmental contingencies. In this sense, it has been repeatedly observed that these rearrangements produce the expected variations, and that psychotic behaviors and experiences can be both experimentally induced and modified by means of behavioral procedures (see Ayllon et al., 1965; Burns et al., 1983; Layng & Andronis, 1984; Lindsley, 1963, 1964; Mace et al. 1988; Oswald, 1962; Rosenfarb, 2013; Skinner, 1936, 1957).

Moreover, this conceptualization also relies on the kind of sociological data that is often put forward in defense of psychosocial approaches to psychotic experiences (Rosenfarb, 2013; Wong, 1996; Wilder et al., 2020). As the literature on the social risk factors of schizophrenia shows, there are myriad socio-economic factors that have been observed to increase the risk of having a psychotic mental health problem: having a low socio-economic status, living in urban areas, having had adverse or traumatic experiences (e.g., child abuse), belonging to groups oppressed on the basis of racialization, ethnicization, gender, etc. Rosenfarb (2013) has attempted to explain these social risks from a functional analytic perspective. According to him, "in examining the social environmental factors that have been associated with the onset of schizophrenia, the common denominator appears to be an early environment that is socially isolating and leaves an individual feeling socially defeated" (p. 932). On his view, social isolation and deprivation are general environmental factors that could account for the origin, development, and maintenance of psychotic experiences. In the case of delusions, for example, Rosenfarb (2013) follows Maher (1974/2005; see Chapter 7, section 7.1.) in understanding them as bizarre -yet adaptive- explanatory responses to

aberrant experiences; specifically, these would owe their bizarreness to the individuals' "lack [of] corrective social feedback to normalize their unusual experiences" (p. 934), ultimately due to their heightened social isolation. In this sense, contrary to cognitivist theories, which emphasize hypothetical individual, internal factors in the origin and maintenance of psychotic experiences, applied behavior analysis yields a deeply environmentalist perspective, which emphasizes the role of environmental sources of psychological distress. (This doesn't mean that biological or genetical factors are dismissed; what is rejected is not their influence, but rather their hypothetical role as necessarily "immediate" or "primary" causes of psychopathological behavior. In this sense, behavior analysts highlight the "loopy" character of the interplay between an organism's biological makeup and the history of interactions it establishes with the environment; see Rosenfarb, 2013).

However, although this general theory may provide us with clues for intervening at a social scale, most traditional behavior analytic research on psychotic experiences has focused on the analysis and modification of the specific environmental contingencies maintaining each individual's target behaviors. Early behavior analytic approaches to delusions and other psychotic phenomena go back to the origins of applied behavior analysis itself in the 1950's and the 1960's (see Cooper et al., 2019, p. 29; Rutherford, 2003) -which contrasts vividly with the current relative under-representation of behavior analytic interventions outside the sphere of developmental problems (Hayes, 2016). Lindsley and Skinner (see Lindsley 1956, 1959, 1963, 1964; see also Rutherford, 2003) conducted some of the first investigations along these lines. As early as 1956, Lindsley applied free-operant measurement methods to analyze psychotic behavior and the environmental variables maintaining them. Although he initially concluded that operant techniques were relatively ineffective in the modification of delusional and hallucinatory speech and other psychotic behaviors (Lindsley, 1959), subsequent research yielded more positive results. For example, Ayllon & Michael (1959) successfully trained the nursing workforce of a psychiatric hospital to record and measure the behavior of several patients, as well as to implement diverse behavioral techniques in order to reduce different target responses. In the case of a person with non-normative verbal behavior, the recording made by the nurses revealed that they were probably maintained by attention. Drawing from such observation, the authors trained the nurses to apply a systematic extinction procedure (i.e., attention diverting) which resulted in a pronounced reduction in the relative frequency of psychotic talk. Following a similar procedure, Ayllon & Haughton (1964) produced long-lasting changes in the frequency of another individual's delusional verbalizations following a pretreatment assessment which revealed that they were also probably maintained by attention. By contrast, other early functional analytic

interventions focused on the reinstatement of verbal behavior rather than its reduction (see Isaac et al., 1960; see also Sherman, 1965).

It must nonetheless be noted that, with the exception of these initial examples, most early functional analytic interventions with people with psychotic behaviors were largely based on the group application of token economies and other systematized treatment procedures to reduce inappropriate responses or promote the acquisition of various desired social skills (self-care behaviors, appropriate interpersonal behaviors, etc.) (see Mace et al., 1991; Hanley et al., 2003; Wilder et al., 2020); in other words, they didn't draw from a pre-treatment functional assessment of the particular contingencies maintaining each individual's problem behaviors, but rather employed "arbitrarily selected consequences (e.g., token reinforcement exchangeable for food or privileges and timeout from reinforcement) to override existing environmental contingencies" (Wilder et al., 2020, p. 318). By contrast, more contemporary FBA-based interventions with people with psychotic experiences have drawn from a pre-treatment assessment of the environmental contingencies that may be maintaining target behaviors to implement therapeutic procedures aimed at modifying such existing contingencies (e.g., Arntzen et al., 2006; Carr & Britton, 1999; DeLeon et al., 2003; Dixon et al., 2001; Horner et al., 1989; Lancaster et al., 2004; Mace et al., 1988; Mace & Lalli, 1991; Rehfeldt & Chambers, 2003; Travis & Sturmey, 2010; Wilder et al., 2001, 2003; see also Wong, 1996, 2006, 2014; Wilder et al., 2020).

As with other problem behaviors, FBAs of non-normative verbal behaviors have been conducted by indirect, descriptive (e.g., observational) and experimental methods (the latter constituting the "functional analysis" proper). The above-mentioned interventions by Ayllon & Michael (1959) and Ayllon & Haughton (1964) constitute an early version of this way of proceeding, which employed non-experimental methods to assess the functions of target behaviors. However, it wasn't until the 1980's, following Iwata et al.'s (1982/1994) development of a formal methodology for conducting experimental FBAs or functional analyses, that the first contemporary FBA-based interventions on delusions and other psychotic phenomena were carried out. Following several theoretical functional analyses of delusions and hallucinations (see Burns et al., 1983; Layng & Andronis, 1984), Mace et al. (1988) were among the first to apply the recently developed formal methods for conducting pretreatment FBAs in the intervention with a person with non-normative verbal behavior. Their functional analysis of the case revealed that non-normative verbal behavior was primarily maintained by both social positive and negative reinforcement, specifically, by effecting "either temporary escape from task demands or experimenter attention" (p. 295). Consequently, they implemented a combination of escape prevention and extinction procedures to modify the

individual's behavior, leading to a sustained reduction in the rate of non-normative verbalizations. According to the authors, these results suggested that "(a) bizarre vocalizations are a function of specific positive and negative reinforcement contingencies rather than hypothetical mental processes [...], and (b) interventions can be developed that interrupt the contingencies that maintain bizarre statements" (Mace et al., 1988 p. 295). Subsequent interventions have mainly followed Mace et al. (1988) in the use of experimental methods of FBA (DeLeon et al., 2003; Dixon et al., 2001; Lancaster et al., 2004; Rehfeldt & Chambers, 2003; Travis & Sturmey, 2010; Wilder et al., 2001, 2003), although some have also employed indirect and descriptive methods (Jimenez et al., 1996; McDonough et al., 2017; Vandbakk et al., 2012) or a combination of non-experimental and experimental assessment methods (Arntzen et al., 2006; Carr & Britton, 1999; Horner et al., 1989; Mace & Lalli, 1991).

Many different environmental contingencies have been analyzed and manipulated to assess their effect on target behaviors, including non-normative verbal behaviors: social attention, tangible reinforcers, access to preferred activities, interchangeable tokens, self-stimulation, escape from tasks or demands, time-out, etc.) (Wilder et al., 2020; Wong, 1996, 2006). According to Wilder et al. (2020), the experimental conditions that have most often featured in functional analyses (i.e., experimental functional assessment) of non-normative verbal behavior include the following:

> Typical conditions employed as part of functional analyses with individuals with schizophrenia include attention conditions, in which attention is delivered contingent upon the target behavior (test for social positive reinforcement), demand conditions, in which a brief break is provided contingent upon the target behavior (test for social negative reinforcement), alone conditions, in which no programmed consequences are provided for the target behavior (test for automatic reinforcement), and control conditions, in which patients are provided with non-contingent attention and no demands are delivered. (Wilder et al., 2020, p. 322)

Regarding treatment procedures, the most commonly employed intervention techniques have been operant in nature. These have usually included: a) different kinds of differential reinforcement of pro-therapeutic behaviors -e.g., differential reinforcement of alternative (DRA), other (DRO), or incompatible responses (DRI), or low rates of response (DRL)-, often combined with the extinction of target behaviors (including escape prevention in the case of behaviors maintained by negative reinforcement) (Anderson & Alpert, 1974; Arntzen et al., 2006; Ayllon & Haughton, 1964; Ayllon & Michael, 1959; DeLeon et al., 2003; Dixon et al., 2001; Horner et al., 1989; Jimenez et al., 1996; Lancaster et al., 2004; Mace et al., 1988; McDonough et al., 2017; Rehfeldt & Chambers, 2003; Travis & Sturmey, 2010; Vandbakk

et al., 2012; Wilder et al., 2001, 2003); b) non-contingent reinforcement (Carr & Britton, 1999; Lancaster et al., 2004; Mace & Lalli, 1991); and c) mild negative punishment procedures, such as time-out techniques (Davis et al., 1976; Haynes & Geddy, 1973). On the other hand, systematic desensitization has been the most commonly employed classical conditioning technique when psychotic experiences were hypothesized to be maintained by Pavlovian conditioning (Alumbaugh, 1971; Nydegger, 1972; Slade, 1972) (see also Froján-Parga et al., 2019; Wilder et al., 2020; Wong, 2006)[62].

As mentioned above, these methods have been commonly employed in institutional, in-patient settings, where there's a maximal control over the environmental contingencies that may control target behaviors. However, traditional behavior analysts have also aimed to understand the typical procedures employed in psychotherapy, which usually takes place in outpatient settings. Of special interest has been the understanding of the cognitive techniques and procedures (e.g., Socratic dialog), which play major role in CBT practice in general and in CBTp interventions on delusions and other psychotic experiences in particular (see Chapter 7, section 7.1.). As we saw in Chapter 1 (see section 1.3.2.), these techniques were historically developed to overcome the alleged limitations of more traditional behavioral procedures, which didn't tackle maladaptive reasoning processes "directly" (Alford & Beck, 1994). Functional analytic practitioners, both "orthodox" and "heterodox", don't question the utility and effectiveness of these cognitive procedures; as we saw, they just challenge the cognitivist understanding of their functioning (e.g., Jacobson, 1996; see also Froján et al., 2017, 2018). In particular, from a traditional behavior analytic point of view, what cognitive therapists understand as the "cognitive restructuring of the person's system of beliefs" can be reconceptualized as a series of verbal changes following the (primarily verbal) implementation of the same classical and operant conditioning procedures that are employed in inpatient contexts; specifically, cognitive techniques have been understood as a mix of verbal operant procedures (i.e., extinction or verbal punishment of problem responses plus the differential reinforcing, shaping, and chaining of desired ones) and verbal Pavlovian procedures (i.e., pairing of verbal stimuli) (see Calero et al., 2013; Froján et al., 2011, 2017, 2018; Pereira et al., 2018; see also Sturmey, 2020; Wong, 1996).

Regarding the efficacy of traditional behavior analytic interventions with people with delusions, although there are several narrative reviews (see Mace, 1994; Mace et al., 1991; Travis & Sturmey, 2008; Wong, 1996, 2006, 2014), quantitative syntheses are rare. There are several reasons for this. On the one hand, as we mentioned above, many behavior analysts

[62] For more detailed descriptions of these procedures, see Cooper et al. (2019).

have traditionally rejected medical-model-based methods of assessment of the efficacy of psychological interventions, which are based on the use of randomized controlled trials and the subsequent review and meta-analysis of their estimated effect sizes. The rationale behind this decision may be understandable -as we've mentioned, a stronger emphasis on the idiosyncratic character of each particular case is needed to promote advances in the psychological intervention with people with delusions and other psychotic experiences. However, systematic reviews and meta-analyses are nonetheless useful tools for summarizing existing data, providing approximate estimates of the efficacy of psychological interventions, establishing potential sources of bias, or promoting inter-theoretical exchanges, among other things.

A related second limitation is related to the kind of study designs and outcome measures typically employed in traditional behavior analytic interventions. Consistent with their strong emphasis on the individualization of assessment and treatment, the maximization of experimental control, and the prioritization of direct behavioral measures over inferential constructs, these interventions typically employ single-case designs (e.g., AB^k designs, multiple baseline designs, alternative treatment designs, etc.) and gather direct free operant measures (e.g., frequency or rate of target vs. alternative behaviors, etc.). By contrast, most effect size estimators are designed to estimate the results from group comparisons typically measuring outcomes via standardized tests or questionnaires, and classical effect size estimators of single-case interventions face a number of problems (Pustejovsky, 2018).

However, following a renewed interest in individual-centered practice, new statistical tools have been developed that allow for the quantitative synthesis of results from single-case interventions -most notably the Log Response Ratio (see Pustejovsky, 2018), which allows for the estimation of the effect size of single-case interventions using free-operant measures. This allowed us to conduct a systematic review and meta-analysis of effect sizes of FBA-based interventions on non-normative verbal behaviors related to the experience of hallucinations, delusions, and disorganized speech (see Froján-Parga et al., 2019).

In our review, we found 23 studies (24 cases) conducting FBA-based interventions on non-normative verbal behavior associated with delusions, hallucinations, and disorganized speech. These included a total of 29 interventions on non-normative verbal behaviors and 19 interventions on alternative normative behaviors. Some interesting descriptive results were the following: a) delusions were the most commonly assessed behavioral topography -approximately 58.3% of reviewed studies included an assessment of delusions, with 41.7% assessing delusions plus other topographies, namely hallucinations (33.3%); b) most

interventions employed an AB^k design (91%), and the most used FBA method was the experimental one (i.e., the functional analysis), which was employed in 58.3% cases (41.7% alone, 16.7% in combination with indirect or descriptive methods); c) in the majority of cases, non-normative verbal behavior was seen or hypothesized to be maintained by social positive reinforcement (75%), namely attention by others (62.5% only attention, 12.5% attention plus escape)[63]; and d) the most frequent behavior modification technique was differential reinforcement of alternative or other behaviors (66.7%), often combined with extinction (50% of total interventions, 75% of interventions employing differential reinforcement procedures).

Among all the interventions found, we could carry out a quantitative synthesis of 19 interventions aimed at reducing non-normative verbal behaviors. (By contrast, we couldn't analyze the average effect size of complementary interventions aimed at increasing alternative normative behaviors due to the sheer number of interventions which met the methodological requirements for conducting a quantitative synthesis -i.e., 11.). We found that the average effect size corresponded approximately to a percentage decrease in non-normative verbal behavior of 72%, with a 95% confidence interval ranging from 62% to 79%. We interpreted this result as showing that FBA-based interventions with people with non-normative verbal behaviors were effective in analyzing the environmental contingencies potentially maintaining these behaviors, adapting the behavior modification procedures accordingly, and achieving relatively large therapeutic effects -at least if these are exclusively measured in terms of the ability to reduce non-normative verbal behaviors.

In addition, this overall average effect size was seen to be moderated by the publication year and the quality analysis index, although subsequent analyses suggested that the moderating effect of the former could be at least partially explained by the moderating effect of the later. We interpreted this result as suggesting that more recent interventions tended to score higher in the quality analysis index and that those interventions with higher quality

[63] From our perspective, this result shouldn't be understood as implying that interventions (nor daily social relations for that matter) should consist in ignoring or rejecting people with psychotic experiences whenever they talk about their non-normative experiences or views of reality. Very much to the contrary, we think that such thing could actually be pernicious to the person, potentially increasing their stigmatization and social isolation, as well as potentially promoting the kind of abuse which people with psychotic experiences are too often subjected to (see Chapter 6, section 6.3.2.). Thus, far from the derogatory practices used by others (including of course behavior analysts), we think that this result must be strictly interpreted along the following lines: contingently paying greater attention to non-normative rather than normative verbal behaviors may be an important factor maintaining their frequency, duration, or intensity; thus, if a reduction in such parameters was something valued *by the person whose behavior is functionally assessed*, then therapeutic techniques consisting in the re-allocation of social reinforcement to alternative, competing, or other verbal behaviors (or simply to low rates of non-normative verbal behaviors) could be a useful intervention technique. The same goes for other environmental variables that could potentially have a maintaining role in a person's non-normative verbal behavior.

analysis index were more likely to report higher percentage decreases in problem behavior. By contrast, neither the type of functional assessment, intervention technique, behavior topography, nor diagnosis were found to be significant moderators. We interpreted these results as pointing to the general utility of conducting even indirect pre-treatment functional behavioral assessments to inform subsequent interventions, regardless of what specific behavior modification procedure is used, and regardless of the kind of non-normative verbal behavior analyzed or the received diagnostic label.

Despite these seemingly promising results, our study faced a number of limitations. Regarding the analysis of the overall efficacy of traditional behavior analytic interventions, a strong limitation of our study is that we only reviewed the efficacy of FBA-based interventions, leaving out many early behavior analytic studies that didn't count as FBA-based (see above). Furthermore, our study sample mainly consisted of interventions comparing the effects of the FBA-intervention to no treatment or treatment as usual conditions (e.g., reversal or withdraw conditions in AB^k designs). Thus, our results only constitute an estimate of the absolute efficacy of FBA-based interventions, not their relative efficacy when compared to other potential treatment conditions.

Nonetheless, we think there're reasons to be at least initially optimistic regarding the potential therapeutic impact of conducting pre-treatment FBAs as a means of tailoring intervention to each individual case and enhancing intervention results. On the one hand, although our synthesis included interventions on all three kinds of psychotic experiences (i.e., delusions, hallucinations and disorganized speech), delusional verbalizations were the most common assessed behavioral topography (58.3%). Since behavior topography was not a significant moderator of the average effect size, we can provisionally assume that the effect size of FBA-based interventions with people with psychotic experiences was not diminished in the case of delusions, as it's been observed in the case of cognitive behavioral therapy for psychosis (CBTp).

Along these lines, previous research has found that FBA-based interventions yield better results than non-FBA-based interventions. For example, Hurl et al. (2016) compared the efficacy of FBA-based with non-FBA-based interventions on a variety of problem behaviors. They found that, while the former had a large effect on the reduction of problem behavior, the latter had no effect when compared to no intervention. Future studies should draw similar comparisons in the case of interventions with people with delusions and other psychotic experiences. This way, we may be able to assess in more detail the potential therapeutic gains of adopting a causal-interventionist approach based on the FBA of target behaviors in each particular case.

Finally, another interesting result found by Hurl et al. (2016) was that the effect of FBA-based interventions on competing (e.g., incompatible, alternative, other, low rate, etc.) behaviors was four times greater than the effect found in non-FBA-based interventions. Although our study sample didn't contain enough interventions on alternative desired behavior, a great deal of functional analytic literature has focused on the promotion of self-care activities, recreational skills, and other personally and socially-valued behaviors through the management of arbitrarily chosen contingencies (e.g., via token economies) (see Hanley et al., 2003; Wilder et al., 2020; Wong, 1996). Therefore, the employment of pretreatment FBAs to tailor intervention to the specific characteristics of each case seems to be a promising tool not only for reducing non-normative behaviors (which not always is the desired outcome by users, as we'll see in the next section), but also for further enhancing the therapeutic power of functional analytic interventions aimed at increasing desired behavioral outcomes.

So far, we've seen how delusional and other psychotic experiences have been conceptualized, assessed and treated from a more traditional behavior analytic point of view. This straightforward non-cognitivist clinical tradition aims to eliminate all talk of cognitive, inner processes to describe delusional experiences via their operationalization in strictly behavioral terms. In the next section, we'll see how the more "heterodox" post-Skinnerian strand represented by Acceptance and Commitment Therapy (ACT) has approached the intervention with people with delusions. As we saw in Chapter 2 (section, 2.3.5.), the hallmark of this more heterodox functional analytic approach lies in that, instead of endorsing a straightforward eliminativist or ontologically radical approach, it endorses a more flexible, revisionary attitude, which is comfortable making use of -functionally definable- mentalistic talk to explain psychotic experiences and other mental health problems.

8.3. ACT interventions with people with delusions

As we saw in Chapter 1 (see section 1.5.2.2.), ACT is a "post-Skinnerian" functional analytic approach to mental health that primarily emerged within the field of clinical behavior analysis (hence CBA). In a nutshell, CBA encompasses a number of contemporary functional analytic approaches to psychotherapy that aim to provide an explanation of therapeutic change in out-patient, ambulatory settings with verbally competent users (Dougher, 2004; Dougher & Hayes, 2004; Guinther & Dougher, 2013; Follette et al., 1996; Kohlenberg et al., 1993, 2002; Madden et al., 2016). In its origins, the main goal of CBA was to provide an answer to the "talk therapy question" (i.e., why therapeutic changes achieved inside the clinic following a number of relatively brief, weekly or even monthly sessions of "talk therapy" can generalize and transfer to extra-clinical settings; see Kohlenberg et al., 1993, p. 271). In their attempt to

provide a satisfactory answer to this and related questions, which Skinnerian analyses of language apparently failed to provide (Hayes, 2004), some authors eventually developed an alternative, post-Skinnerian account of language and cognition: Relational Frame Theory (RFT) (Barnes-Holmes et al., 2001; Hayes et al., 2001), on which ACT relies. Before moving on to the analysis of ACT interventions with people with delusions, let's firs recap the main tenets of RFT.

As we saw, RFT proponents retrieve from cognitive models the idea of the causal primacy of cognition in the explanation of complex forms of behavior (e.g., concept formation, logical reasoning, etc.). However, they don't conceptualize cognition in internal, information processing terms, nor endorse strong commitments regarding the ontological status of cognitive variables; rather, RFT reconceptualizes some of these supposedly mediational variables (those that prove to be explanatory useful) in functional analytic terms. In particular, RFT draws from the core notion of *arbitrarily applicable relational responding* (i.e., an individual's responding to one event in terms of other events with which it bears no "natural" or "physical" resemblance) to explain how equivalence and non-equivalence relations are formed, which RFT views as the functional analytic equivalent to symbolic and inferential processes. Eventually, RFT's analyses of language and cognition had far-reaching philosophical and methodological consequences, which represent the main differences between the functional contextualist framework of RFT and ACT and that of traditional behavior analytic approaches. For instance, while the latter rejects the use of hypothetical constructs to explain behavior, the former is happy to make use of "topographically mentalistic terms" (Hayes, 2021, p. 239) as long as they're useful for predicting and influencing behavior. In addition, while traditional behavior analysts typically endorse the stricter methodological policy of assessing intervention efficacy through single-case experimental studies, functional contextualist researchers have widely endorsed the use of group comparisons and meta-analyses to assess the evidence base of treatment procedures (see below).

ACT builds onto this conceptual and methodological framework to explain psychopathology. Like CBT, it emphasizes the role of the person's interpretations of the world (i.e., the person's patterns of arbitrarily applicable relational responding) in the origin and maintenance of psychological problems. Specifically, these approaches draw from the assumption that verbal rules[64] may sustain both problem and alternative behaviors even in the

[64] Recall that rules, in the sense of the term being employed here, are bits of linguistic behavior that describe contingency relations (i.e., relations between a given response and certain environmental events), which have a causal role in the production and maintenance of other kinds of behavior. Thus, this sense of "rule" is equivalent to the understanding of rules as "regulative propositions" (Ryle, 1949/2009), i.e., to the nomological use of the

face of other competing contingencies. The relative insensibility of rule-governed behavior to actual contingencies is not only presumed to explain why therapeutic gains transfer from clinical to extra-clinical settings, but also why psychopathological behaviors and experiences are originated and maintained (Hayes et al., 1999). Thus, one core therapeutic strategy consists in modifying the control exerted by such verbal rules and inviting people to "make contact" with the actual contingencies operating in their environment.

The case of delusions is not different. For instance, Monestès et al. (2014) have recently argued that this rule-based insensitivity to changing environmental contingencies may precisely be what explains the maintenance of delusional ideas; in fact, in their pilot study, they found preliminary evidence in this direction. Along these lines, Stewart et al. (2016) have proposed to understand persecutory delusions as follows:

> We propose that (persecutory) delusional beliefs can be defined functionally as behaviors, and in particular, as [arbitrarily applicable relational responding]. Derived stimulus relating may explain why [persecutory delusion] sufferers respond with fear, anxiety, and worry, or even attempt to escape or avoid particular stimuli and events, despite having not encountering them previously. Increasingly complex instances of [arbitrarily applicable relational responding] (rules) may also be central in the development and maintenance of [persecutory delusions], further restricting the individual's behavioral repertoire, while efforts to respond in ways that are coherent with those rules may increase their influence. (Stewart et al., 2016, p. 240)

In this sense, ACT shifts somewhat away from the strict case-by-case policy adopted by many traditional behavior analytic practitioners and posits a common factor to account for many psychopathological behaviors: *experiential avoidance*, or the tendency to engage in avoidance behaviors to escape from noxious covert experiences (e.g., negative thoughts, aversive feelings, etc.), which was later incorporated into ACT's model of *psychological flexibility* (see Vilardaga et al., 2009). Specifically, psychological problems emerge from the adoption of inflexible behavioral rules that establish or reflect arbitrary (i.e., social-conventional) relations between events. Once established, these equivalence and non-equivalence relations are primarily maintained because of their escaping or avoidance function: they allow individuals to avoid noxious experiences (Hayes et al., 1999). This way, ACT attempts to explain the typical "backfiring" effect that common coping strategies like thought suppression or reason-giving have, i.e., that they actually increase the very same noxious experiences

term. "Rules" in this sense are different from "norms", as we've used the term in previous chapters (i.e., as normative standards that determine the correctness or incorrectness of different courses of action).

they are set up to eliminate. Drawing from this assumption, ACT therapy starts by teaching individuals a different way to relate to their own experiences; instead of trying to avoid or control them through typical thought suppression or reason-giving strategies, ACT therapists encourage people to engage with their inner experiences in a non-judgmental, accepting manner (i.e., one whereby the person does not struggle to eliminate or reason away such experiences, but instead accepts their presence), and to refocus on overt behavior and the achievement of valued goals (Bach & Hayes, 2002; Hayes, 2004; Hayes et al., 1999).

ACT for psychosis follows the same reasoning. In the case of auditory verbal hallucinations, ACT conceptualizes them as covert stimuli that typically prompt suppression or reason-giving strategies aimed at eliminating them (Bach & Hayes, 2002; Bach et al., 2006; García-Montes & Pérez-Álvarez, 2005; McLeod, 2009; Veiga-Martínez et al., 2008). In the case of delusions, ACT assumes that they might constitute themselves a form of verbal avoidance strategy; specifically, delusional thought would serve as a non-normative rationalizing strategy that prevents exposure to aversive thoughts and feelings of self-worthlessness (Bach & Hayes, 2002; Bach et al., 2006; García-Montes & Pérez-Álvarez, 2005; McLeod, 2009; Pankey & Hayes, 2003). As McLeod (2009) puts it, this conceptualization of delusions seems to be "a variant of the 'delusions as defense' hypothesis which proposes that delusions protect the individual from experiencing noxious affect" (p. 272). Several studies have provided preliminary evidence for this claim (Goldstone et al., 2010; Udachina et al., 2009, 2014). For instance, in their mediational analysis, Goldstone et al. (2010) found that experiential avoidance was a relevant mediating factor of the influence of life hassles on the frequency of delusional ideas and the distress associated to them in both clinical and non-clinical populations, and concluded that "coping with life hassles by attempting to suppress or avoid unwanted thoughts, may increase proneness to delusional ideation in people from the wider community, as well as facilitating the ongoing maintenance of delusions in people with a diagnosis" (p. 263). In a similar vein, two studies found that experiential avoidance also mediated tendency to have paranoid thoughts in both a clinical (Udachina et al., 2014) and a non-clinical student sample (Udachina et al., 2009). An alternative hypothesis could be that the main function of delusions is to reduce the distressing character of aberrant perceptions (such as the one's described by people with Capgras syndrome) by proving an explanation for such disturbing experiences, which would be in line with one factor theories of delusions (see Chapter 7, section 7.1.; see Bach et al., 2006; García-Montes & Pérez-Álvarez, 2005; McLeod, 2009; see also Gaudiano, 2015).

Drawing from this conceptualization, ACT intervention on delusions departs from both CBTp and traditional behavior analytic interventions in a number of relevant respects.

As we've seen, these two approaches typically target delusional contents and deploy different strategies to reduce their frequency -even if doing so in a non-confronting, collaborative manner (Alford & Beck, 1994). However, this common therapeutic goal does not always match what people with delusions themselves seek; as several service user led investigations have pointed out, many people with psychotic experiences don't view recovery as necessarily entailing a reduction in such experiences (Kilbride et al., 2013; Pitt et al., 2007; Wood et al., 2013).

By contrast, the intervention rationale of ACT is completely different to that of CBTp and traditional behavior analysis. ACT doesn't target delusional verbalizations or thoughts per se nor sets their reduction as its primary goal. In fact, the first ACT interventions with people with psychotic experiences emerged as a critique of the usual methods employed by CBTp (Bach & Hayes, 2002; see also Pankey & Hayes, 2003). From the ACT perspective, the more dialectical and reason-giving-based methods of CBTp approaches could in fact be promoting unhelpful coping strategies like thought suppression or avoidance (Bach et al., 2006; Bach & Hayes, 2002; Gaudiano et al., 2010; McLeod, 2009; Pankey & Hayes, 2003; see Gaudiano, 2015). This could explain the relative ambiguity in the evidence supporting the efficacy of CBTp interventions with people with delusions (see Chapter 7, section 7.2.). Instead, ACT interventions target the way in which many people with delusions relate to their own experiences, and aims to provide them with a different way of relating to such experiences, no matter whether this eventually leads to a reduction in their frequency or not (Bach et al., 2006; Bach & Hayes, 2002; García-Montes & Pérez-Álvarez, 2005; Gaudiano et al., 2010; McLeod, 2009; Pankey & Hayes, 2003; see Gaudiano, 2015). In consonance with its general view of psychopathology, ACT for delusions focus primarily on reducing the *believability* of disturbing delusional experiences (i.e., the "cognitive fusion" with the delusional content, or their steadfast, automatic interpretation in literal terms) and promoting the acceptance of both delusional experiences as well as those other potential feelings of depression or anxiety that delusional thoughts may be helping to avoid. In addition, this strategy is tightly connected with the focus of this type of intervention on personal values and on promoting alternative values-related courses of action in therapy: instead of presuming that the reduction of delusional thoughts per se is necessary for having a life worth living, ACT assumes that the specification of each individual's valued life horizons is a case-by-case task (Pérez-Álvarez, 2012). Pankey & Hayes (2003, p. 317) sum up the general ACT approach to interventions with people with psychotic experiences as follows:

> ACT focuses on the client's original aim of controlling their private experiences and situates willingness and defusion as the vehicles by which individuals learn that acceptance of aversive private emotion or bodily states is a process, not an outcome. The individual learns that through awareness, vulnerability, flexibility, and willingness one can begin to let go of old control agendas (e.g., "buying into" the veracity of delusional ideation) and learn that what needs to change is *the stance one has* in regard to negative private emotions or bodily states, not the emotion or bodily state itself. ACT shifts the focus from modifying the private experience to modifying one's reaction to the private experience. The goal is to assist the client in embracing more difficult psychological context while simultaneously focusing on valued overt behavior change. A key component to acceptance of private experience is teaching the client to defuse from tangled cognitions. Here, the client learns that the literal truth or falsity of the cognition need not be a target. Instead, the patient is directed toward their goals and behaviors. (Pankey & Hayes, 2003, p. 317)

Regarding the efficacy of ACT interventions with people with psychotic experiences, this change in focus needs to be taken into account. When assessed in terms of "symptomatology reduction", the results of ACT don't look very promising. For example, in the first randomized control trial assessing the efficacy of ACT for the treatment of psychotic experiences, Bach & Hayes (2002) compared the effects of a brief ACT intervention comprising just four 45-50-min individual sessions to treatment as usual (TAU): the authors found no statistically significant difference between the ACT and TAU groups in the frequency of psychotic experiences -among those participants who reported them- nor in the distress associated with them at follow-up; in fact, they found that those assigned to the ACT group were significantly more likely to report psychotic experiences at follow up. A replication study by Gaudiano & Herbert (2006) also found no significant effect of ACT on the frequency and severity of psychotic experiences. Subsequent studies have yielded mixed results: while some of them have continued to show similarly negative results on positive symptoms reduction (Shawyer et al., 2012), other have yielded partially positive results (see Gaudiano et al., 2013, 2015, 2020; Shawyer et al. 2017). However, the results of recent metanalyses rather point in the direction of a lack of effect on positive symptom frequency; although initial meta-analyses showed short-term small to moderate effect sizes on positive symptoms (Khoury et al., 2013; Cramer et al., 2016), others have found no significant effects when compared to control groups (Brown et al., 2021; Jansen et al., 2020; Louise et al., 2018; Tonarelli et al., 2016; see also Öst, 2014)[65].

[65] It should be noted that many of these meta-analyses merged ACT interventions with other third-wave behavior therapy approaches such as Mindfulness-Based Stress Reduction (MBSR; see Jansen et al., 2020) under the

These results are nonetheless in consonance with the acceptance-based focus of ACT. In this sense, the success of ACT interventions with people with psychotic experiences cannot be measured in the same terms as the efficacy of CBTp and traditional behavior analytic interventions; from this perspective, the main intervention goal is not to reduce the frequency of psychotic experiences, but to reduce their believability and risk of rehospitalization, and increase a number of mindfulness and acceptance measures, as well as social functioning (see Bach et al., 2012, 2013; Bach & Hayes, 2002; Gaudiano et al., 2010, 2013; Gaudiano & Herbert, 2006; Pankey & Hayes, 2003; White et al., 2011).

When measured in its own terms, the results of ACT interventions with people with psychotic experiences are somewhat more favorable. For example, Bach & Hayes (2002) reported a 50% reduction in the rate of rehospitalization over a 4-month follow-up period. They also reported a significantly lower believability of psychotic experiences in the ACT group; in fact, they found that none of the participants who admitted psychotic experiences at follow up *and* showed lower believability in them were hospitalized at follow-up. This was interpreted by the authors as pointing to the efficacy of ACT in promoting a more accepting attitude towards the person's psychotic experiences and their associated distress. Subsequent studies have shown relatively similar positive results (see Bach et al., 2012, 2013; Gaudiano & Herbert, 2006; Gaudiano et al., 2010). Bach et al. (2012) found that the results from their previous study maintained at one year follow-up. In their replication of the Bach & Hayes (2002) study, Gaudiano & Herbert (2006) found similar differences in rehospitalization rates between ACT and TAU, although they weren't statistically significant probably due to the smaller sample analyzed, and they also found increased social functioning. Bach et al. (2013) pooled the samples from Bach & Hayes (2002) and Gaudiano & Herbert (2006) together and found that experience believability mediated rehospitalization at 4 months follow-up.

Subsequent studies by Gaudiano et al., (2013, 2015, 2020) with participants with both depressive and psychotic experiences have shown positive therapeutic changes in hypothesized change mechanisms (e.g., experiential avoidance, psychological flexibility, mindfulness, values consistent living, etc.) associated with affect, functioning, and psychotic experience improvements. However, other recent studies have yielded mixed results regarding the efficacy of ACT on these alternative outcomes when compared to control groups (Shawyer et al., 2012, 2017; White et al., 2011; for a systematic review, see Wakefield, 2018). Evidence from metanalytic studies also yields a more ambiguous picture. On the one hand, Khoury et al. (2013) found that mindfulness, acceptance, and compassion measures were significantly

common term of "acceptance- and mindfulness-based interventions" with people with psychotic experiences, with the exception of Brown et al. (2021) and Tonarelli et al. (2016).

increased by acceptance- and mindfulness-based interventions compared to control groups, and that these measures moderated clinical effect size. Tonarelli et al. (2016) also found a significant difference in the efficacy of ACT on negative symptoms and rehospitalization rate when compared to TAU. Likewise, Jansen et al. (2020) found moderate to large effect sizes in hospitalization and acceptance, and small to moderate effects on negative symptoms, mindfulness measures, and social functioning. However, other meta-analyses have only found significant differences for mindfulness measures, but not for acceptance measures, nor negative symptoms, distress, or functioning (Cramer et al., 2016; Louise et al., 2018).

Finally, adding to these ambiguous results, there's also evidence that, when positive effects of ACT are found, these may be primarily explained by an enhanced management of hallucinations, but not delusional experiences. In their initial randomized control trial, Bach & Hayes (2002) found that the effect of ACT on believability was primarily found in the case of people with hallucinations; by contrast, reduced experience believability was not found for one third of participants with delusions in the ACT group who continued to deny symptoms. Pankey & Hayes (2003) attributed this reduced effect to the brief character of the intervention, and pointed out that "if delusions themselves are verbal avoidance strategies, it is not so much the delusional process that needs to be accepted but rather the feelings of failure, depression, anxiety, and so on that the delusions may help regulate" (p. 322). Gaudiano & Herbert (2006) found that changes in experience-related distress were mediated by changes in believability in the case of hallucinations, but they couldn't test such effect in the case of delusions due to the reduced number of participants that reported delusional thoughts (see also Gaudiano et al., 2010). By contrast, Shawyer et al. (2017) did report greater improvement in delusion-related distress in the control group than in the ACT group. On the other hand, no meta-analyses to date seem to have established comparisons between the effect of ACT interventions in enhancing the coping strategies to deal with stressful hallucinatory vs. delusional experiences. This might be due to an increased tendency in randomized control trials to report undifferentiated outcome measures (e.g., to assess "positive symptomatology" altogether).

So far, we've seen two different non-cognitivist, functional analytic approaches to the conceptualization, assessment and treatment of delusions. While traditional behavior analytic practitioners tend to adopt a rather ontologically radical approach towards the operationalization of delusional and other psychotic phenomena, ACT researchers and practitioners tend to adopt a rather revisionary attitude, which is happy to make use of middle-level terms (e.g., cognitive fusion) as long as these are explanatorily useful. According to the former, the debate about the doxastic status of delusions that we saw in Chapters 5 and 6 makes

no sense altogether, since beliefs *qua* mental entities "don't exist"; instead, delusions are operationalized as patterns of non-normative verbal behavior. The latter, by contrast, take it that the cognitivist understanding of delusions can in fact give clues about the actual processes involved in the origin and maintenance of delusions, but just fail to provide an adequate specification and explanation of these processes; instead, ACT conceptualizes the cognitive "deficits" that allegedly explain delusions (i.e., negative self-concept, jumping-to-conclusions, etc.) as fairly common and prevalent instances of arbitrarily applicable relational responding that yields inflexible behavioral rules, which are later maintained by experiential avoidance. In the next section, we'll see what are, from our point of view, the strengths and limitations of these two functional analytic approaches to delusions. As we'll see, some of the problems of these approaches may be due to the residual endorsement of an intellectualist view of the mind (Chapter 2, section 2.1.2.). Instead, we'll recommend the adoption of the kind of non-descriptivist approach to the mind (Chapter 4) as a way out of these and other related problems.

8.4. Non-descriptivism and the functional analytic approach to delusions

In this and the previous chapter, we've seen several issues regarding the currently available evidence on the efficacy of psychological interventions with people with delusions. Cognitive behavioral therapy for psychosis (CBTp), a traditional cognitivist account, has been the most studied so far, yet its efficacy in the case of delusions is somewhat ambiguous. Alternatively, the functional analytic approaches that we've focused on here present a number of conceptual benefits over traditional cognitivist accounts, and they afford a different way of understanding, assessing, and treating delusional experiences which may enhance the utility of psychological treatments. However, the evidence base available so far is relatively underdeveloped. Firstly, traditional behavior analytic interventions (and specially FBA-based interventions) offer an individually tailored, formulation-based, and causal-interventionist approach to mental health that has yielded overall promising results, at least if we understand therapeutic efficacy in terms of the reduction of non-normative verbal behavior. However, quantitative syntheses of the results of these interventions are rare and need further development. Secondly, drawing from a functional contextualist approach, ACT has emphasized the need to shift the focus from the reduction of problem behaviors to the modification of the person's way of relating to the world and their psychotic experiences. However, the evidence supporting this kind of approach is also relatively ambiguous -at least in the case of delusions.

Adding to the current relative underdevelopment of their evidence base, we think that functional analytic approaches to delusions may also be subject to certain conceptual issues, which may in turn limit their efficacy and clinical significance[66]. These conceptual issues, as we'll try to show now, have to do with a certain "residual" endorsement of the Cartesian view of the mind (in particular, its intellectualist assumptions) which is to a lesser or greater extent shared by many functional analytic approaches. This problematic commitment, as we'll see, is particularly clear in the case of ACT and other approaches within clinical behavior analysis, but traditional behavior analytic researchers also express some adherence to it. As we view it, the problem lies in the kind of reconceptualization strategies that many functional analytic researchers and practitioners adopt when trying to account for folk-psychological and cognitivist notions; particularly, those of "belief" or "mental representation". We'll first review these reconceptualization strategies, pointing out how they are related to certain objections traditionally raised by cognitivism. We'll then see how these responses retain a certain commitment to intellectualism, and how this yields several problems for traditional behavior analytical and ACT approaches to delusions, respectively related to their operationalization and explanation of delusional phenomena. Finally, we'll see how our non-descriptivist approach to mental-state ascriptions may provide a fit philosophical complement for functional analytic approaches to delusions. We'll argue that it may contribute to deepen the split away from the logical mold of Cartesianism, with potential therapeutic benefits.

8.4.1. What functional analytic researchers deny and don't deny

Functional analytic approaches have long faced certain systematic -yet unfounded- accusations by traditional cognitivists and other "schools of thought". Chief among them is the accusation of "denying people's mental lives", i.e., of denying the existence of people's feelings, thoughts, imaginings, and so on (see Baars, 2003). Certainly, many functional analytic researchers deny some mental states and processes (super-egos, deep cognitive schemas, etc.), or at least their explanatory utility -hence the insistence in overcoming mentalistic "explanatory fictions" and focusing on the environmental sources of behavioral control (e.g., Skinner, 1945, 1950, 1953, 1957, 1963, 1971, 1977, 1984, 1990). However, most of them also assume the existence of certain "mental" events (namely, occurrent episodes like imaginings, inner speech, and so on), although they propose to understand them in behavioral terms. The

[66] The difference between these two comes down to the difference between the statistical significance of some intervention effect (e.g., as measured in randomized controlled trials) vs. the degree of clinically relevant change that a person achieves following such intervention in actual cases.

following qualification by Skinner (1974) of the radical behaviorist approach to the mind is often invoked in response to this criticism:

> The statement that behaviorists deny the existence of feelings, sensations, ideas, and other features of mental life needs a good deal of clarification. Methodological behaviorism and some versions of logical positivism ruled private events out of bounds because there could be no public agreement about their validity. [...] Radical behaviorism, however, takes a different line. It does not deny the possibility of self-observation or self-knowledge or its possible usefulness, but it questions the nature of what is felt or observed and hence known. [...] Mentalism kept attention away from the external antecedent events which might have explained behavior, by seeming to supply an alternative explanation. Methodological behaviorism did just the reverse: by dealing exclusively with external antecedent events it turned attention away from self-observation and self-knowledge. Radical behaviorism restores some kind of balance. It does not insist upon truth by agreement and can therefore consider events taking place in the private world within the skin. It does not call these events unobservable, and it does not dismiss them as subjective. It simply questions the nature of the object observed and the reliability of the observations. (Skinner, 1974, p. 18)

This and similar remarks have been commonly advanced in response to the above-mentioned objection: behavior analysts (and contextual behavioral scientists) don't deny the existence of many cognitive or emotional states and processes; rather, they just deny their conceptualization as independent realms of fact, as separated from behavior (see Chapter 1, section 1.3.1.). Many cognitive (and affective) variables are not essentially separated kinds of events, which may always figure in causal chains as the immediate antecedent of overt behaviors; they *are* behavior, part of what needs explanation.

Of special interest to our present discussion is the functional analytic reconceptualization of thoughts or beliefs, understood as "mental representations" of reality. Here we need once again to distinguish between dispositional and occurrent understandings of these notions. On the one hand, mental representations might be conceptualized as dispositions, i.e., as being stable or not having "genuine duration" (see Chapter 4, section 4.2.2.). Thus understood, mental representations would correspond to what cognitivists understand as cognitive structures or cognitive schemas (see Chapter 1, section 1.3.2.). However, mental representations have also been understood as discrete phenomenal units of "happenings" -what cognitivists have often referred to as "automatic thoughts". This distinction is well-reflected in the classical functional analytic treatment of mental concepts as either pointing to larger patterns of interactions between an organism and the environment (e.g., Baum, 2011;

Skinner, 1953, 1957, 1963), or to instances of covert behavior (e.g., Moore, 2009; Skinner, 1957, 1974). The following two excerpts from *Verbal Behavior* (Skinner, 1957) capture well this dual treatment:

> The simplest and most satisfactory view is that thought is simply *behavior*—verbal or nonverbal, covert or overt. It is not some mysterious process responsible for behavior but the very behavior itself in all the complexity of its controlling relations, with respect to both man the behaver and the environment in which he lives. The concepts and methods which have emerged from the analysis of behavior, verbal or otherwise, are most appropriate to the study of what has traditionally been called the human mind. (Skinner, 1957, p. 449)

> [...] as a living organism a man is behaving in some sense while "doing nothing," even though his behavior may not be easily observed by others or possibly even by himself. We do not discuss these activities effectively because they are almost always accessible only to the "thinker" and useful verbal responses to them cannot easily be developed. [...]
> In a sense verbal behavior which cannot be observed by others is not properly part of our field. It is tempting to avoid the problems it raises by confining ourselves to observable events, letting anyone extend the analysis to his own covert behavior who wishes to do so. But there would then be certain embarrassing gaps in our account. In intraverbal chaining, for example, necessary links are sometimes missing from the observable data. When someone solves a problem in "mental arithmetic," the initial statement of the problem and the final overt answer can often be related only by inferring covert events. We also have to account for verbal behavior which is under the control of covert speech—which reports it [...] or qualifies it with autoclitics [...]. Covert behavior has also had to be considered in discussing grammar [...], sentence composition [...], editing [...], and other topics [...]. (Skinner, 1957, p. 434)

Thus, from a functional analytic point of view, a person's mental representations may be reconceptualized as either: a) large patterns of behavior, which may or may not involve instances of covert behavior, if understood in dispositional terms; and b) instances of covert verbal or non-verbal behavior, if understood in occurrentist terms (see Skinner, 1945, 1953, 1957, 1963, 1977, 1984). A key observation here is that neither the first "dispositional response" nor the second "occurrentist response" readily assume that belief-talk primarily refers to *verbal* behavior (in the sense of linguistic or vocal); in fact, Skinner (1957) pointed out that despite the tendency of many thinkers to equate "thought" with "verbal behavior", the "important and distinctive functions of verbal behavior [...] are nevertheless not relevant to a definition of thinking" (p. 448).

However, from our perspective, the overall tendency in the functional analytic treatment of delusions (and other clinically relevant irrational beliefs) has been to overemphasize the role of verbal behavior in the operational definition of delusional experiences. As we view it, this hints at a relatively residual endorsement of the intellectualist understanding of the mind, which might be primarily due to the fact that verbal behavior preserves "some of the magic we expect to find in a thought process" (Skinner, 1957, p. 447). In particular, this "magic" which Skinner refers to is the intentionality or representational capacity of thought. Beliefs, as mental representations, bear an intentional relation to the world: they're said to "represent" it, to "be about" it (Jacob, 2019). However, once we get rid of belief-talk and other supposed traces of mentalistic vocabulary, how to account for intentionality? The natural response is to turn our eyes to another, more tractable thing which is said to have such representational capacities: language. In particular, Skinner's notion of "tact" was developed in an attempt to provide an empirically plausible reconstruction of this representational capacity of language in behavioral terms. Here, representation is understood in non-mystical, causal (functional) terms: a tact is a verbal response which is reinforced by an audience in the presence of a given stimulus (i.e., the one to which the verbal response is said to "refer to").

Regardless of the success of Skinner's particular approach, what is important here is that the apparent pressure to reconceptualize intentionality and symbolism in behavioral terms is what may be behind the tendency to identify beliefs (as mental representations) with verbal behavior. We think that it's this identification of beliefs with verbal behavior which underlies the *overemphasis* of traditional behavior analytic and ACT approaches on the operationalization of delusional ideation in terms of non-normative verbal behaviors or inflexible rules, respectively. As we view it, this may hinder progress in promoting a sound non-cognitivist, functional analytic approach to delusions and other psychotic experiences.

8.4.2. Traditional behavior analysis and the superficiality objection

Let's first consider the case of traditional behavior analytic interventions with people with delusions. In section 8.2., we've seen that traditional approaches have typically conceptualized delusional ideation in terms of non-normative verbal behaviors. This relatively restricted focus on the person's verbalizations historically gave rise to a number of criticisms by CBTp theorists and practitioners; namely, that behavior analysis and behavior therapy just offered a "superficial treatment" of delusions, one which just dealt with the "symptoms" or overt manifestations of the delusional ideation, but which left its more "profound" roots (i.e., the person's cognitive structures) untouched. The following excerpt captures well the spirit of this criticism:

> Though there is much overlap between behavior therapy and cognitive therapy [...] traditional behavioral treatment of delusions is readily distinguished from the approach of cognitive therapy. The central difference is that between modification of *verbalizations,* or 'verbal behavior' [...] and belief modification [...]. As noted by Stahl and Leitenberg (1976) [...] "it has been clearly demonstrated [by behaviorists] that delusional speech can be controlled through operant techniques. An unresolved question is whether delusional 'thought' is modified by the same methods" (p. 234).
>
> [...] delusional thinking and beliefs are *not* necessarily modified by such therapies [...]. Furthermore, delusions are *by definition* conceptualized as cognitive phenomena, not 'behavioral' phenomena. To argue that they are merely 'verbal behavior'-as some behaviorists suggest [...]-is to greatly oversimplify the nature of delusional ideation. (Alford & Beck, 1994, p. 370)

This criticism -which we may refer to as the *superficiality objection*-, wasn't just raised as a specific worry with the behavioral understanding of delusions, but rather echoed a general concern regarding the apparent "shallowness" or simplicity of behavior analysis or behavioral approaches to psychology more broadly. This concern is still widely shared today -for instance, its echo can be heard in the above-mentioned strawman depiction of behaviorism as denying the existence of thoughts and feelings (see Baars, 2003)- and it help laid the foundations for the cognitivist approach to psychotherapy in general and psychosis in particular (see Alford, 1986; Alford & Beck, 1994).

We're skeptical -to say the least- about the ontological status or alleged causal-explanatory roles of any "profound" cognitive structures which traditional cognitivists may resort to. We don't think that the kind of selectionist, contextualist, and functionalist explanations of behavior provided by behavior analysts have the kind of "explanatory gaps" that shall be filled by invoking deeper cognitive states and processes. That said, we also think that, properly understood, there is some basis to the superficiality objection. In fact, we want to argue that this basis comes from the residual commitment to the very intellectualist framework that gave rise to traditional cognitivist approaches to psychotherapy in the first place. The overemphasis of traditional behavior analytic practitioners on the operationalization of delusional experiences in terms of non-normative verbal behaviors hints at this residual intellectualist commitment. It seems to imply an underlying conceptualization of delusions as beliefs -that is, an underlying acceptance of the doxasticist account of delusions (see Chapter 5)-, and an underlying conceptualization of beliefs as primarily verbal behavior. As a result, delusions are primarily operationalized as non-normative verbal behaviors, either overt or covert. Paired with the general -yet clearly disputable- tendency to understand therapeutic

success in terms of "symptom" or problem reduction, this explains why traditional behavior analytic interventions have typically set as their primary goal the reduction of delusional speech.

Now consider the case of a person who no longer claimed (either overtly or covertly) to be followed or spied by others -maybe after undergoing a functional analytic intervention-, but still lived in fear, anxiously checking their surroundings and avoiding large agglomerations. In this case, we would probably still ascribe them the belief that they are being spied or followed. We think that, in this scenario, most people -including many traditional behavior analysts- would feel strongly inclined to say that the intervention might have been *efficacious* (i.e., it might have achieved statistically significant or otherwise demonstrable effects in whatever outcome it set out to change) but not so much *effective* (i.e., its results weren't as *clinically significant* as it would have been desired). The determination of what may count as clinically significant is no doubt an evaluative question, which involves a necessary reference to the personal and social values of different stakeholders (e.g., the person undergoing therapy, their community, the therapist, other social agents, etc.), and hence it won't have an easy or straightforward answer. However, we think that the following rough proviso will be shared by many: in at least many cases, what will set the bar as to the clinical significance of therapy will be its ability to produce changes in the person's *state of mind*, not in their behaviors *per se*.

Surely, what cognitivists like Alford & Beck (1994) would take this to mean is something like the following: that changing one's verbal behavior is one thing, but changing the root cognitive *causes* of such verbal behavior is another. However, we think that our non-descriptivist approach offers a more parsimonious way of understanding this criticism; one that once again shifts focus from mental states to mental-state ascription practices and their role in therapy. This would lead us to rephrase our proviso as follows: in order to assess a behavioral change as clinically -and not just statistically- significant, this change should be intersubjectively evaluated as reflecting a broader change in the person's mental states.

As we view it, once we've scratched beneath its cognitivist surface, this is the actual criticism lying at the core of the superficiality objection: that a change in a person's covert or overt behavior does not necessarily reflect a change in their attitudes, i.e., in their beliefs, desires, intentions, expectations, etc., *as assessed from a particular evaluative standpoint*. And hence the obvious question arises: "behavior modification procedures may reduce what the person says they believe… but, do they *really* change what the person believes?". In other words: do these procedures have real, clinically significant effects?

So far, we've seen how the intellectualist operationalization of delusions as patterns of primarily verbal behavior may dampen the perceived utility of traditional behavior analytic interventions with people with delusions, giving way to the superficiality objection. Let's now consider how ACT interventions may also be hampered by this residual commitment to intellectualism.

8.4.3. From mental to verbal representations: the specter of intellectualism

As we've seen, according to cognitive therapists, the operationalization of beliefs (and thus delusions) as verbal behaviors is left lacking because it doesn't address how the person's "deep cognitive structures" (beliefs or cognitive schemas) of the world might change. In the face of this objection, many traditional behavior analysts adopt a straightforward eliminativist response: beliefs, as overall views of reality which stand in logical or justificatory relations with other such views, wouldn't exist. After all, as we saw in Chapters 3 and 4 (sections 3.2.2. and 4.2.2.), no genuinely "justificatory" or "logical" -i.e., normative- relations obtain in a purely descriptive worldview; all we may sensibly talk about is the person's overall patterns of (verbal) behavior.

By contrast, other functional analytic scholars (e.g., ACT proponents, but also other clinical behavior analysts) see this response as unsatisfactory. These approaches share with cognitive models the assumption that there's something lacking in the kind of account provided by traditional behavior analysts; in particular, what seems to be missing is an account of the hypothetical *causal link* that apparently needs to be established between the environmental contingencies and the person's behavior when the former doesn't seem to effectively control the latter (Hayes et al., 1999). The "causal glue" or "mediational variables" invoked by cognitivists were the person's beliefs or mental representations of the world. For many clinical behavior analysts, something like these mental representations needs to be posited in order to explain why the modification of primarily verbal behavior in clinical contexts can lead to further behavior changes in extra-clinical contexts, where the environmental contingencies that gave rise to psychological problems in the first place may still be operating (i.e., the "talk therapy problem"). As we saw in Chapter 1, Hayes (2004) takes this to be the main reason for the success of cognitivism: that it provided an account of how the *arbitrary relations* that a person establishes among different events may causally impact subsequent behaviors, often in spite of competing environmental contingencies.

In this sense, many clinical behavior analysts take the superficiality objection seriously: no functional analytic approach to behavior will be complete until an account of these "beliefs" or "mental representations" is given. Their response to this challenge thus involves providing some kind of behavioral redefinition of "beliefs" or "mental representations" that

preserves their causal powers. On this view, beliefs, as well as the justificatory or logical relations that may stand among them, do exist, but these are reinterpreted in terms of behavioral rules, i.e., verbal descriptions of the functional relations that may hold between certain behaviors and certain environmental events. ACT's Relational Frame Theory (RFT) precisely aims to provide such reinterpretation. According to RFT, beliefs and the logical relations among them can be understood in terms of arbitrarily applicable relational responding, whereby individuals come to establish increasingly complex arbitrary relations among events, eventually yielding relatively inflexible rules (see section 8.3.; see also Chapter 1, section 1.5.2.2.). Alternative accounts, some of them more akin to the tenets of traditional behavior analysis, reject the concept of "relational responding", providing instead a reverse account of these arbitrary relations: it's not relational responding which yields the formation of behavioral rules, but rather the learning of behavioral rules which, through Pavlovian conditioning processes, yields the establishment of associative relations among different verbal and non-verbal stimuli (e.g., Froján-Parga et al., 2017). Be that as it may, these approaches share the reconceptualization of "beliefs" and other "deep cognitive structures" in terms of verbal rules, whose covert or overt emission explains the "insensibility to contingencies" that is often observed both in problem responses as well as in the maintenance of alternative ones in the (often-unsupportive) extra-clinical environment.

As we view it, this reconceptualization of beliefs and mental representations in terms of verbal rules is fundamentally flawed; namely, because it amounts to a mere secularization of the intellectualist legend of Cartesianism. The idea of "believing as entertaining mental representations" is not radically questioned; it's just replaced by the more secular notion of "believing as entertaining *verbal* representations". To be sure, we completely agree that understanding how shaping verbal behavior in the clinical context translates into changes in many other behaviors across many different settings is one of the most important questions to be addressed. We're convinced that research on rule-governed and complex verbal behavior is absolutely necessary, and that a sound functional analytic approach to clinical practice will surely need to incorporate this. However, the problem comes when this explanatory project is confounded with the old reductivist project of translating belief-talk to the language of behavioral science.

This yields several problems. In their attempt to retrieve the explanatory power of traditional cognitivist approaches, ACT supporters and other clinical behavior analysts buy into the conceptualization of "beliefs" and "mental representations" as primarily *causal* devices. On this view, to believe that certain state of affairs is the case ultimately comes down to having established certain relations among events which are reflected in -or effected

through- verbal rules, i.e., verbal representations of the world. Hence, "acting on the grounds of one's beliefs" comes down, once again, to behaving in certain ways as a *causal effect* of some anterior or ongoing, overt or covert operation of "planning what to do" (Ryle, 1949/2005, p. 20)[67].

Note how little this conceptualization of beliefs departs from traditional cognitivist thinking. Clutton's own cognitive-phenomenological approach (see Chapter 5, section 5.2.2.) is not very different: according to Clutton (2018), believing that *p* is just a matter of judging that *p*, i.e., entertaining certain propositions or mental representations (e.g., verbal rules) "before one's mind" (i.e., covertly) and b) taking them to be true -which, for RFT researchers, would be a relational behavior of its own. The reconceptualization of beliefs as verbal rules just introduces two amendments to this picture: a) the demystification of mental representations, whereby these are seen as primarily verbal responses, not essentially different from other kinds of behavior; and b) their de-internalization, whereby covert and overt utterances of these verbal rules are seen as equivalent. From our point of view, these amendments are insufficient. They're just not *radical* enough; they may introduce an even more secular account of what it is to have a certain belief -one that rejects dualism-, but which nonetheless leaves the descriptivist, factualist, and intellectualist roots of the Cartesian picture of the mind relatively untouched.

Other than the conceptual problems associated to these commitments (see Chapters 2, 3, and 4), they may negatively impact the efficacy of functional analytic interventions with people with delusions. In particular, the ACT approach is especially vulnerable to the sort of limitations faced by CBTp interventions (see Chapter 7, section 7.3.). One core positive feature of early and contemporary behavior analysis, as we've been stressing here, is its individually tailored, formulation-based, and causal-interventionist approach to assessment and treatment. From this perspective, no prior assumptions about the factors maintaining a given behavioral pattern are pre-empirically warranted; they must be empirically tested on a case-by-case basis via the functional assessment -and preferably functional analysis- of behavior. By contrast, a corollary of the reconceptualization of beliefs as verbal rules is that any instance of what we may, from a folk-psychological point of view, describe as "acting in accordance with certain beliefs" will *necessarily* be understood as an instance of rule-governed behavior, i.e., of behavior that is controlled by verbal rules. Drawing from this

[67]We have defended a similar view in a discussion of the possible behavioral processes involved in cognitive techniques (Froján-Parga et al., 2017, 2018), although relying on Pavlovian pairing processes rather than relational frames as mediational explanatory devices. Insofar as these were presumed to explain one and every possible instance of "acting on the grounds of one's beliefs", our account was also faultily committed to intellectualism.

theoretical framework, it seems like ACT practitioners should already know all there is to know about the controlling variables that maintain any instance of behavior that is commonly understood in terms of beliefs. Why conduct an FBA then?

In the ACT approach to delusions, this shift away from the functional analytic tradition is obvious. Delusions, whose conceptualization as beliefs doesn't seem to be questioned by ACT either, are readily understood as inflexible verbal rules (see section 8.3.). These are supposed to exert control over the person's other behaviors, as well as to be primarily maintained by experiential avoidance, i.e., by allowing the individual to avoid or escape certain noxious experiences (e.g., feelings of self-worthlessness). The possibility of cases where no verbal source of control is to be found, let alone the potential role of other kind of reinforcers other than "experiential avoidance" (e.g., access to various social or non-social reinforcers) is completely bypassed by the theoretical conviction that, since delusions are beliefs, delusional behavior must be readily explainable in terms of rule-governed behavior. Once again, the empirical aspect of clinical practice seems to be straitjacketed by pre-empirical assumptions; this, in turn, could be one of the reasons why ACT interventions with people with delusions have shown mixed results, even when measured by its own standards.

8.4.4. A non-descriptivist response to the superficiality objection

Now, what could be a better response to the superficiality objection? From our perspective, cognitivists raise an obvious point: that modifying what someone says -either overtly or covertly- does not necessarily imply that someone's beliefs have changed. However, this doesn't mean that the functional analytic conception of delusions is lacking an obliged reference to hypothetical underlying cognitive structures -nor behavioral rules or relating operations, for that matter. Instead, we might understand this claim along the lines of the pragmatist kind of non-descriptivist approach to the mind that we exposed in Chapter 4.

As we view it, what traditional behavior analytic accounts of delusions lack are not more complex descriptions of facts (i.e., of the hypothetical deep cognitive causes of delusional behavior), but more complex accounts of the plurality of language games and social practices that may be at play in therapy -something which, from our perspective, other cognitivist and non-cognitivist approaches also lack. In particular, what is lacking here is a recognition of the actual role of mental-state ascriptions in therapy. From our non-descriptivist approach, these do not subserve a descriptive function -at least not primarily-, but an evaluative one; their place is not to be found in nomological accounts of human affairs, aimed at the goals of prediction and control; rather, they feature in normative, meaning-making practices, aimed at the goals of comprehension, rationalization, and responsibilizing. In therapy, mental-state ascriptions don't point to any hidden objects nor hypothetical *primum*

movens whose existence we must affirm or deny; rather, they just set the bar for what we will evaluate as a *clinically significant* intervention: one which is able to produce the kind of changes that, within a certain community, will be assessed as "actual changes in the person's mental state" -that is, in the person's norm-following (not rule-governed) behavior.

This directly affects both the operationalization of the intervention goals as well as the determination of the causal variables at play in each particular case. To begin with, our non-descriptivist approach puts the recognition of value plurality at the center of the operationalization of therapeutic goals. In this sense, we would be better off if we took the results from user-led research seriously and abandoned the misplaced assumption that the overall purpose of therapy must always be "symptomatology reduction" (see Wood et al., 2013) -a deeply entrenched inheritance from the medical model, from our point of view. In this sense, ACT's recognition of the plurality of models of recovery, as well as its focus shift from problem reduction to improving one's ability to cope more adequately with one's experiences, are two of its key virtues, from our point of view. The determination of the therapy goals must always involve an analysis of personal and social values at play in each particular case.

But even if we were to proceed under a "problem reduction" model of recovery, what may count as believing or ceasing to believe in a delusional content in each particular case will still be an evaluative task. And, in some occasions, there may be disagreement over whether someone has or has not ceased to believe in a certain content. These disagreements may not always dissolve by "appealing to facts", since at least part of them may arise due to conflicting evaluative standards: what one party evaluates as an instance of "believing that p", another party may not (Curry, 2020; Pérez-Navarro et al., 2019). Thus, not only determining the goals of therapy, but also their actual fulfilment (i.e., the clinical significance of the intervention) is an irreducibly evaluative task which may sometimes or even typically require the balancing of different perspectives. Sometimes it will suffice with reducing the rate of non-normative verbal behaviors; sometimes, it won't. In many cases, perhaps the majority of them, the clinician will need to assess and intervene on a much broader range of responses other than non-normative verbal behavior. A larger register of how the person behaves, verbally and non-verbally, overtly and covertly, across time and across contexts (e.g., in the clinic, with the family, with friends, etc.), will be needed to determine whether they can be assessed as "not believing that a vengeful deity is after them" or as believing other less damaging contents. In other words: most likely, the response to the superficiality objection doesn't lie in providing a behavioral reconstruction of the hypothetical cognitive structures posited by cognitive therapists, but in a) assessing a broader range of behaviors other than non-normative verbal behavior; b) taking the need for generalization and follow-up analyses

more seriously; and, above all, c) taking the *evaluative* -rather than descriptive- complexity of mental health practice at face value.

In addition, our non-descriptivist view of belief ascriptions also has some consequences for the issue of explanation. In particular, it helps re-emphasize the need for grounding interventions in pre-treatment FBAs of the potential contingencies maintaining behavior, regardless of whether this can be understood as "belief-behavior" or not. It thus helps dispel the specter of intellectualism. It is one thing to attempt to explain or induce behavior changes by analyzing the controlling role of verbal behavior, and quite another to try to reinterpret every instance of what may be assessed as "belief behavior" in terms of rule-governed behavior. As we view it, a sound functional analytic approach to clinical practice surely needs the former, but not the latter -in fact, it can well be counterproductive. At least in some functional analytic approaches, these two explanatory projects seem to be confounded. What is at stake here is a confusion between the notion of "norms" (evaluative criteria) and the notion of "behavioral rules" (i.e., regulative propositions), and hence between "norm-following behavior" (i.e., acting in accordance to certain norms, which may or may not be subjectable to explicit formulation) and "rule-governed behavior" (i.e., instances of behavior that are totally or partially controlled by behavioral rules).

Instead, our non-descriptivist approach to the mental draws an important distinction between these two notions. Acting on the grounds of one's beliefs (or desires, intentions, etc.) is an instance of norm-following behavior, i.e., of behavior that is *logically* or "grammatically" linked to certain norms and further courses of action. By contrast, not every action which we might describe in folk-psychological terms necessarily is an instance of rule-governed behavior, i.e., of behavior that is *causally* controlled by certain overt or covert rules. Neither believing mad contents is necessarily a matter of entertaining mad regulative propositions in one's theatre of consciousness or having knotty neurocognitive makeups, nor it is a matter of engaging in mad verbal behavior -either covertly or overtly. In other words: whether someone's overall patterns of actions and reactions can be properly assessed as an instance of "believing that *p*" is *orthogonal* to whether their behavior is totally or partially maintained by verbal sources of control. The former, as we've seen, is an evaluative issue, one which no determinate fact may settle down for every possible case; by contrast, the latter is something that only an FBA can help determine.

To be fair, many of the FBAs conducted in FBA-based interventions, at least in the case of people with delusions, analyze a shockingly narrow set of environmental contingencies -typically reduced to attention, escape from demands, self-reinforcement, and control conditions (see Froján-Parga et al., 2019; Wilder et al., 2020; Wong, 1996). As we view it, these

functional assessments would be much richer and would considerably gain in explanatory accuracy if they also considered additional potential sources of verbal control, such as the ones that have been emphasized by some clinical behavior analysts. In any case, what more traditional behavior analysts are right to point out is that there's no shortcut to analyzing the variables controlling the target behaviors in each particular case. No matter whether these are understood as instances of "belief behavior" or not, an FBA must always be carried out if we're to properly explain the problem at hand and determine the most appropriate treatment methods (Froxán-Parga, 2020). Our non-descriptivist approach to belief ascription practices contributes to re-emphasize this point.

Once the Cartesian mist of intellectualism has been dispelled, we're in a better position to assess the main virtues of the two functional analytic approaches to delusions that we've reviewed here. On the one hand, the greatest appeal of the more traditional functional analytic branch of FBA-based interventions is its strict emphasis on the need for conducting individualized pre-treatment FBAs of the target behaviors at stake, as well as to provide a careful analysis of the evolution of the intervention across treatment phases. This provides a non-cognitivist variety of the individually tailored, formulation-based, and causal-interventionist approach to psychological interventions that many are starting to advocate for. In doing so, it contributes to a more solid, idiosyncratic, and potentially more effective approach to the intervention with people with delusions and other psychotic experiences; one whose explanatory and intervention potential does not depend on conceptual debates about the doxastic or non-doxastic status of delusional phenomena. On the other hand, the main virtues of ACT and other approaches to clinical behavior analysis are namely two: a) that they draw attention to the possible verbal sources of behavioral control, which may need to be tackled in order to foster the generalization and maintenance of therapeutic gains over time; and b) that it encourages alternative ways of thinking about recovery, following user-led research in the promotion of a shift from "problem-reduction" to the development of more appropriate copying strategies and the realignment of the person's actions with their own values as core therapeutic goals.

As we view it, a "both-ways" functional analytic approach to mental health -one that combines the analysis of verbal behavior and its potential controlling functions, while remaining strictly committed to the FBA methodology- is worth exploring. Such an approach could promote the advancement of psychological interventions with people with delusions and other psychotic experiences -a field which yet remains a fiefdom of pharmacotherapy, with its many downsides (see Chapter 1, section 1.1.). In addition, it would provide a more complete functional analytic approach to process research, i.e., the analysis of the processes

underlying therapeutic efficacy. This, precisely, has been the overarching goal of some functional analytic researchers at the Autonomous University of Madrid, whom during the last 15 years have been investigating the basic behavioral processes involved in the verbal interaction in therapy (e.g., Alonso-Vega et al., 2019; Froján-Parga, 2011; Froján-Parga et al., 2006, 2008, 2010a, 2016, 2017; Montaño-Fidalgo et al., 2013; Ruiz-Sancho et al., 2015; Pascual-Verdú et al., 2019)

This work is part of a larger effort to advance in that direction. As we view it, our pragmatist and non-descriptivist approach to the mind provides a good complementary philosophical position in various respects. In a nutshell, its main contributions are the following: a) it encourages a return to the fundamentals of functional analytic interventions, shifting explanatory efforts from an individualistic focus on putative cognitive mediators to the relational analysis of the particular environmental contingencies maintaining each target behavior; and b) it frees contemporary functional analytic approaches from the intellectualist assumptions that have traditionally conditioned cognitive models, which may undermine therapeutic efficacy and effectiveness. It does so by helping to disentangle several questions which are often confounded, and which have consistently appeared throughout the present work. In particular, it helps us distinguish ontological puzzles about the nature and causal role of mental events from the analysis of the truth-conditions of mental-state ascriptions (see Chapters 2, 3, and 4). This allows us to dismiss the former while retaining the latter, resisting the urge to engage in flawed and potentially anti-therapeutic reconceptualizations of folk-psychological talk. In other words: it helps functional analytic oriented researchers to deepen the detachment from the logical mold of Cartesianism, with all its concomitant problems, while avoiding falling into narrow and confounding reductivisms or eliminativisms of any sort. Instead, it fosters a more complete and coherent view of the different language games that continuously crisscross in the field of mental health. In doing so, it pays due attention to the complaints of many cognitivist thinkers without compromising the overtly non-cognitivist approach of functional analysts to the nomological aspects of the science of behavior. From our point of view, this partnership could contribute to the development of new synergies among research areas; not only between strands of functional analytic approaches to psychotherapy, but also among non-cognitivist approaches in general (e.g., the enactivist approach to mental health, see de Haan, 2020a, 2020c, 2021).

8.5. Conclusion

In this chapter, we've presented a non-cognitivist, functional analytic approach to the intervention with people with delusions and other psychotic experiences. We've differentially

reviewed the main conceptual tenets and evidence status of two main strands within the functional analytic approach to delusions: one drawing from more traditional forms of behavior analysis, and another one drawing from the functional contextualist framework of Acceptance and Commitment Therapy.

As we've seen, functional analytic approaches share some common central features, among which the following may be highlighted: a) the focus on behavior as the primary unit of analysis, broadly understood as the relation between an organism and the environment; b) the assumption that this formulation encompasses both overt and covert forms of behavior; and c) the explanation of behavior in functional and selectionist terms. However, there also are a number of important differences. On the one hand, traditional behavior analysts endorse a more "orthodox" reading of Skinner radical behaviorism. This is characterized by a stricter preference for single-case experimental designs to assess therapy outcomes, a marked tendency towards the direct observation of behavioral changes vs. the postulation or modelling of hypothetical mediators, and a strict adherence to Functional Behavioral Assessment (FBA) methods to carry out the case formulation and guide subsequent interventions. Traditional behavior analytic interventions have thus typically taken place in in-patient settings, where there's a maximum control over the environmental contingencies potentially maintaining target behaviors.

On the other hand, ACT and other clinical behavior analysts are primarily interested in explaining and promoting clinical changes in ambulatory, outpatient contexts. Thus, they place a greater emphasis on the analysis of verbal behavior and its potential mediational role in the explanation of the transfer, generalization, and maintenance of behavioral changes across time and settings. In particular, ACT draws from a more "heterodox" functional analytic approach, lately identified with Contextual Behavioral Science the philosophy of functional contextualism. This approach is characterized by relatively looser methodological and conceptual commitments, with a more permissible attitude towards the modelling of hypothetical mediators and the use of group comparisons and meta-analytic methods to assess intervention efficacy.

These two functional analytic strands also differ regarding the operationalization, explanation, and treatment of delusions. Within traditional behavior analysis, delusions and other psychotic experiences have been commonly operationalized in terms of non-normative verbal behaviors, which may be maintained by different environmental contingencies. The most commonly analyzed ones have been social positive reinforcers like attention, social negative reinforcers like escape from demand, and self-reinforcement. Typical behavior modification procedures have included individual techniques like extinction plus differential

reinforcement of alternative, incompatible, other, or low-rate responses, and group procedures like token economies. From the 1980's onwards, traditional behavior analytic interventions have been progressively characterized by the employment of a pre-treatment FBA (either by indirect, descriptive, or experimental methods) to assess the possible environmental functions of target behaviors and design interventions accordingly.

On the other hand, ACT researchers and practitioners emphasize the verbal and relational roots of delusions. In line with its general approach to psychopathology, ACT conceptualizes delusions as instances of arbitrarily applicable relational responding that yield inflexible rules, which control behavior in spite of competing environmental contingencies. In turn, these are hypothesized to be negatively reinforced by experiential avoidance, in particular by the avoidance of either thoughts and feelings of self-worthlessness (i.e., the "delusions as defense" hypothesis) or noxious or ambiguous experiences (i.e., the behavioral analogue of the one factor theory of delusions; see Chapter 7, section 7.1.). Regarding therapy, ACT interventions depart from common "problem reduction" models of recovery and focus instead on changing the person's relation to their own experiences. Thus, ACT interventions have commonly targeted experience believability, rather than experience frequency, as well as other indirect indicators of well-being, such as the rate of rehospitalization, social functioning, etc.

Overall, the evidence supporting these two models is promising, yet we've identified a number of limitations. In the case of traditional behavior analytic approaches to delusions, we mainly focused on the discussion of the results of FBA-based interventions. In our recent meta-analysis of FBA interventions with people with non-normative verbal behaviors, we found an overall effect size of 72% in target behavior reduction, with a 95% confidence interval ranging from 62% to 79%. Although promising, a proper assessment of the efficacy of FBA-based interventions with people with delusions is limited by the sheer number of quantitative syntheses conducted so far.

In the case of ACT interventions with people with delusions and other psychotic experiences, these interventions seem to have mixed results when intervention efficacy is measured in terms of problem reduction. However, this outcome is to be expected once we take into account that the primary goal of these interventions is not to reduce problem behavior, but to promote different ways of relating to psychotic experiences. When assessed by its own measure, these interventions seem to yield somewhat better results, although their observed efficacy may be primarily due to their effect on hallucinations, not delusions; a result which parallels the results found in the case of CBTp (see Chapter 7, section 7.2.).

In the last section, we've discussed how our non-descriptivist approach to mental-state ascriptions may offer some possible ways to improve the prospects of functional analytic approaches to delusions. In particular, we've claimed that the perceived utility and efficacy of these approaches may be limited by a residual commitment to the Cartesian view of the mind and its intellectualist construal of mental states. This residual commitment can be hinted at in the typical responses that functional analytic approaches have yielded against a common objection made by supporters of cognitivism and other therapeutic models: namely, that functional analytic approaches deny the existence of mental states. In the case of delusions, this more general objection is expressed in what we've called the superficiality objection. It entails the idea that functional analytic approaches to delusions are inherently superficial: they may be efficacious in reducing non-normative verbal behaviors, but they nonetheless fail to address the supposedly deeper causal roots of these behaviors, i.e., the person's cognitive schemata.

Ultimately, the functional analytic response to this objection depends on the reconceptualization strategy that is followed to account for the notion of "belief", understood in terms of "mental representations". Functional analytic oriented researchers have typically endorsed two possible strategies: a) to claim that "beliefs" just amount to a person's overall patterns of verbal and non-verbal, covert and overt behavior, across time and different contexts; or b) to assume that belief-talk refers to covert episodes which may figure as intermediary steps in some explanations of certain responses (e.g., calculating an equation). We've argued that, despite none of these strategies emphasize verbal over non-verbal behavior, this tendency to equate thinking and language may be observed in clinical practice, at least in the case of the functional analytic approach to delusions. It's this overemphasis on verbal behavior which, from our perspective, hints at a residual commitment to the intellectualist view of the mind, and which may be limiting the perceived utility and efficacy of functional analytic interventions with people with delusions.

On the one hand, traditional behavior analytic interventions operationalize delusions as patterns of non-normative verbal behavior. What seems to underlie this operationalization is the identification of delusions with beliefs (thus implicitly endorsing the doxasticist characterization) and beliefs with patterns of primarily verbal behavior. This is what gave rise in the first place to the superficiality objection: it is one thing to change what one says one believes and another one to change what one *really* believes. Despite being skeptical of the usual cognitivist framing of this objection, we've nonetheless claimed that there's a sense in which the objection is accurate. In particular, we've claimed that it is mental-state ascriptions, not mere descriptions of behavior, what set the bar for what we typically value as

"effective" interventions (i.e., those with clinically significant results). Thus, the emphasis on the identification of delusions with patterns of primarily verbal behavior may dampen the perceived utility of traditional behavior analytic interventions with people with delusions.

On the other hand, clinical behavior analytic approaches like ACT take the superficiality objection seriously: no functional analytic account of psychopathology will be complete until a proper behavioral reformulation of beliefs -understood as mental representations- and their causal roles is given. These approaches reconceptualize beliefs as verbal rules that either reflect or establish certain relations among events, and which may in turn override other sources of control, making behavior insensible to actual contingencies. We've argued that this reconceptualization poses two main problems. Firstly, in its attempt to retrieve the mediational causal role of beliefs, it sticks its feet further into the Cartesian quicksand. On this account, "acting in accordance with one's beliefs" is once again conceptualized as a dual occurrence, involving the action itself and an anterior (or concomitant) operation of uttering -either overtly or covertly- certain verbal rules. In Chapter 7 (section 7.3.) we claimed that this intellectualist assumption could dampen the efficacy of CBTp interventions by unduly constricting the range of possible causal factors analyzed. Despite their behavioral reformulation of the notion of mental representation, the efficacy of ACT interventions may be limited by the same token.

Given these problems, we've discussed how the pragmatist kind of non-descriptivism endorsed in this dissertation may afford an alternative response to the superficiality objection. Our response, once again, is to understand the belief ascriptions that figure in the conceptualization, assessment, and treatment of delusions as primarily evaluative, not descriptive devices. These belief ascriptions don't point to underlying hypothetical cognitive or neural mediators, but to overall patterns of behavior (both overt and covert, verbal and non-verbal) which are evaluated as instances of norm-following within a particular evaluative framework, grounded on shared social practices.

We've then outlined several consequences of this non-descriptivist approach for the operationalization and explanation of delusions within a functional analytic approach. Firstly, we've stressed that if belief ascriptions set the bar as to what we may count as clinically significant interventions, then the operationalization of therapeutic goals will be evaluative through and through, even if we endorse a "problem reduction" model of recovery. We've claimed that a proper response to the superficiality objection will probably lie along the lines of broadening the range of behaviors to be assessed, securing the generalization of behavioral changes across contexts and their maintenance over time, and, above all, taking the evaluative dimension of mental health seriously. Secondly, we've emphasized the need

to shift away from intellectualist approaches to the causal explanation of belief-like (e.g., delusional) behavior. We've claimed that, whereas a proper functional analytic account of verbal sources of control is needed to account for certain generalization and maintenance issues (e.g., the "talk therapy" problem), this explanatory project must not be confounded with the flawed attempt to reformulate belief ascriptions in behavioral terms. In this sense, our non-descriptivist approach makes a clear distinction between the notions of "norm-following behavior" (e.g., acting in accordance with a certain belief) and "rule-following behavior" (behaving as a causal effect of engaging in certain verbal behaviors): not every instance of the former in an instance of the latter. Hence, there's no shortcut from a case-by-case analysis of the environmental contingencies potentially controlling behavior. In other words: intervention procedures will always need to be grounded on a pre-treatment FBA of target behaviors, whether these are evaluated as instances of "believing a certain content" or not.

Finally, we've laid out what we see as a proper non-cognitivist, functional analytic approach to the intervention with people with delusions: one that pays due attention to the analysis of verbal sources of behavioral control without renouncing to the individually tailored, formulation-based, and causal-interventionist approach to mental health afforded by the FBA methodology. In addition, we've highlighted different ways in which our pragmatist, non-descriptivist analysis of mental-state ascriptions may foster the advance of this non-cognitivist approach to delusions. By providing it with a more complex and accurate account of the different language games and social practices that crisscross in the field of mental health, it may a) promote a further shift away from both the Cartesian theory of mind and its reductivist and eliminativist offshoots; b) free intervention from straitjacketing assumptions about the nature and causes of delusions and other mental health problems; and c) promote inter-theoretical exchanges with other non-cognitivist approaches, such as the recent enactivist approach to mental health.

Chapter 9
Conclusion. Toward a philosophy of mental health without mirrors

The year is 2022. A war has begun. The pandemic still ravages the world. A new economic crisis is on the way. And a mental health one most probably too. Hopelessness, angst, panic, derealization, paranoia, all thrive amidst this climate of crisis and shrink many people's hearts. Some believe there must be a grim hand behind it all pulling the strings of these macabre scenes. Some believe it all comes down to capitalism's intrinsic expansive and self-destructive dynamics. Some hope things will get better -or just won't get worse.

There's been a lot going on while writing this dissertation, and these introductory statements just describe -in a broad sense of the term- part of it. One of our main goals has been precisely to introduce some distinctions between the different sorts of claims at stake here -specifically, between those which are used to describe some temporally and spatially localizable state of affairs (e.g., "A war has begun") and those which say something about people's mental states (e.g., "Some hope things will get better"). The reason, we've said, is that there's an intimate link between mind and normativity, i.e., between our folk-psychological interpretation practices, the "mindgames" that are ubiquitous in our common interactions with each other, and the "ought" dimension of action, which entails the possibility of error and success, merit and demerit, correctness and incorrectness, etc. Hence the mental is not reducible to nor replaceable by mere descriptions of material states of affairs. Yet, at the same time, we've stressed that this doesn't mean that mental vocabulary points to mythical creatures, nor that mental-state ascriptions lack truth value; otherwise, our naturalist or non-naturalist inclinations would be ultimately indistinguishable from mere Jabberwocky verses. We've seen how a pragmatist and non-descriptivist conception of mind, based on Wittgenstein's and Ryle's work, affords a different view of our folk-psychological practices

(one which highlights their evaluative and regulative dimension) and allows us to reconcile two seemingly opposing ideas: a) that mental-state ascriptions don't describe some given fact about an agent; and b) that that doesn't mean that they lack truth value. After all, albeit perhaps in different ways, all the introductory claims above are *true*.

Our main argument has been that this kind of approach provides a sounder framework for mental health research and practice; a "philosophy of mental health without mirrors" that affords a more adequate response to the problems of mind and normativity. In doing so, it's able to resist the diverse reductivist and eliminativist tendencies that pervade many of the debates among competing therapeutic models, as well as the skeptical, individualistic, and, to some extent, non-naturalist inclinations of certain critical approaches (e.g., Szasz's). We've also seen some of its implications for a longstanding debate regarding the standard conceptualization of delusions as beliefs and its consequences for the intervention with people with delusions. In a nutshell, it allows us to retain the best of the doxasticist understanding of delusional experiences and behaviors (i.e., its ethico-political utility), while at the same time dispelling the specter of Cartesianism that straitjackets their clinical and scientific understanding.

In this chapter, we'll summarize the main conclusions of this dissertation and point out some interesting topics to be addressed in future work. In section 9.1., we'll provide a summary of this dissertation, highlighting its main contributions. In section 9.2., we'll sketch out several possible lines for the future development of our non-descriptivist approach to mental health. To do so, we'll turn back to the four major themes of the philosophy of mental health that we saw at the Introduction (i.e., those related to the analogy, boundary, priority, and integration problems), pointing out how our non-descriptivist approach can provide new insights into these matters.

9.1. Summary and main contributions of the dissertation

In the first part of the dissertation, our main goal was to explore the conceptual underpinnings of recurring debates about the nature of mental health problems, and to argue that our non-descriptivist approach to the mind offers a better starting point to address them. We began in Chapter 1 with an overview of the conceptual history of mental health and the main classical and contemporary therapeutic models of mental health problems. We saw the so-called "medical model", which many take to be the prevailing therapeutic approach, actually admits various interpretations. From a minimal interpretation, it just amounts to saying that mental health problems can be usefully understood in medical terms, for a number of purposes (e.g., administrative, epidemiological, etc.). According to a stronger interpretation,

mental disorders are natural kinds; our taxonomies should aspire to "carve nature at its joints" and help us discover the actual neurobiological nature of mental health problems. Both interpretations, especially the latter, have been contested. Critical approaches to mental health pointed out that two cornerstones of psychiatry's legitimation as a medical discipline (i.e., the analogy between mental and somatic disorders and the distinction between "mad" and "bad") amounted to no more than mere myths. Psychological models, on the other hand, criticized the medical model for its internalist or biologicist assumptions, prioritizing instead the analysis of psychological (e.g., environmental, cognitive, etc.) factors involved in the origin and maintenance of mental health problems. The biopsychosocial model then emerged as a synthetic attempt to provide a conciliatory approach for the health sciences in general, including both biological and psychosocial factors as relevant to explain and address mental health problems. Finally, several contemporary approaches (e.g., precision psychiatry, contemporary functional analytic approaches, and the enactivist approach to mental health) have recently emerged, partly in response to the reliability and validity crises of the medical model, partly in response to the integration problems of the biopsychosocial model.

At the end of Chapter 1, we spelled out the two major conceptual issues underlying these debates: a) the problem of mind, comprising a series of issues related to the ontological and epistemic relations between mind and nature; and b) the problem of normativity, related to the place of norms and values in a naturalist worldview and the clash between the manifest and the scientific images of the world and human beings. In Chapter 2, our main goal was to provide a plausible account of the origin of these problems and to expose the different philosophies of mind underlying the diverse therapeutic models. We began by tracing these two problems back to the Cartesian theory of mind, highlighting its ontological and epistemological commitments (dualism, factualism, causalism, intellectualism, and representationalism) and pointing out its semantic cornerstone: descriptivism, or the idea that mental-state ascriptions describe or represent some given state of affairs. We then saw several contemporary attempts to deal with the ontological puzzles of Cartesianism (i.e., the mind-body problem). Drawing from a common commitment to ontological naturalism (defined by the ideas of monism, materialism, and the principle of causal closure), naturalist approaches implement three different strategies to account for the mental: a) to accommodate mental objects within a naturalist ontology (ontologically conservative approaches like straightforward reductivism, functionalism, and emergentism); b) to retain them only in so far as they prove to have some explanatory value, eliminating them otherwise (ontologically revisionary approaches like discourse eliminativism); or c) to dismiss mental states and processes as mythical creatures or explanatory fictions (ontologically radical approaches like

straightforward eliminativism). After reviewing how these different approaches underly the different therapeutic models, we concluded that they all lack a proper account of the normative dimension of the mental; this, in turn, makes them unable to provide a proper conceptual framework for mental health research and practice. We claimed that, to escape this situation, we must shift our focus from the ontological puzzles of mind to the analysis of the meaning and function of mental-state ascriptions.

In Chapter 3, we delved into the issue of descriptivism, analyzing how it has restricted the range of possible answers to the mind-body problem and how it yields untenable forms of naturalism and normativism. We began by situating descriptivism at the core of mind-dreading conceptions of folk psychology, shared by all competing approaches to the mind-body problem, and according to which folk psychology subserves a primarily descriptive, causal explanatory role. We then distinguished two main versions of descriptivism at the semantic level: a) the shallow, affirmative version, according to which the meaning or content of declarative sentences (e.g., mental-state ascriptions) lies in a description or representation of some possible state of the world; and b) the deep, conditional version, which amounts to the claim that only those declarative sentences that successfully describe some state of affairs are truth-evaluable. We then saw how ontological naturalism can be seen as imposing certain restrictions on what may count as a possible state of the world, and hence on what may count as a "successful" description. This leads to the translatability assumption, whereby the compatibility between mind and nature hinges on the possibility of reducing or translating mental-state ascriptions to descriptions of material states of affairs. This leaves naturalists with just two ways to account for the mental: reductive compatibilism or non-reductive incompatibilism. Both, as we saw, lead to a flawed, self-defeating kind of naturalism; one which, in its inability to account for the truth-evaluability or normative force of mental-state ascriptions (which only compatibilists and non-reductivists can retain, respectively), is itself eliminated or reduced to mere descriptions of the neural states or verbal behavior of naturalists. Non-naturalist approaches fare no better: apart from their antiscientific character, they fall prey of Wittgenstein's argument against private rule-following, eventually leading to a self-defeating kind of normativism. Finally, we claimed that, to avoid this dilemma -the puzzle of translatability, we called it- we need to find some kind of non-reductive, yet compatibilist account of the relation between mind and nature; to do so, we must reject descriptivism.

In Chapter 4 we set forth our non-descriptivist approach to the mental. Contrary to other non-descriptivist approaches, ours rejects descriptivism at both the pragmatic and semantic levels, as well as in both the shallow and deep versions of the latter. Instead, it

assumes functional pluralism (the idea that language might be used for many purposes other than stating how the world is) and pluralism about truth (i.e., the idea that the truth value of different declarative sentences might be determined in various ways for different kinds of claims). Drawing from a pragmatist reading of Wittgenstein's and Ryle's work, it assumes that the meaning (and truth-conditions) of a given expression lies in its possible uses in different norm-governed communicative practices (i.e., language games). These possible uses are determined by the inferential or justifiability relations that an expression keeps with other expressions and courses of action (i.e., their logical geography); in turn, these inferential connections are themselves grounded on the various social practices in which we're trained by our community. Knowing what our utterances mean -and what their truth value is, when relevant- is thus viewed as a matter of "knowledge how", rather than "knowledge that". We've then given three arguments (non-durability, truth-conditional dependence, and normative force) which support the idea that mental-state ascriptions have an evaluative and regulative, rather than descriptive function; their main goal is not to predict and control behavior, but to rationalize and justify it. Finally, we've also seen that this doesn't mean that mental-state ascriptions aren't truth-evaluable; rather, their truth or falsity depends on the many different norms that competent speakers follow in folk-psychological interpretative practices (first person authority, overall consistency, etc.). Against the idea that there's some golden rule of interpretation, our non-descriptivist approach endorses a pluralist account, whereby the relevant criteria to determine the truth or falsity of different mental-state ascriptions must be determined on a case-by-case basis. All in all, this approach affords a post-ontological account of the place of mind on nature, which avoids both the mind-body problem and the problem of normativity. To account for the truth-evaluability and normative force of mental-state ascriptions, we don't need to posit ontological weirdos; rather, we just need to recognize the plurality of language games that we play when we try to account for each other's behavior.

This pragmatist, non-descriptivist, evaluativist, and regulativist view of mind provides, from our perspective, a sounder conceptual architecture for mental health research and practice. In the second part of the dissertation, we've applied it to a particular debate in the philosophy of mental health: that concerning the conceptualization of delusions as beliefs and its implications for research and intervention.

In Chapter 5 we introduced this debate. We saw that the standard doxasticism that characterizes the prevailing approaches to delusions (e.g., DSM, cognitive neuropsychiatry, cognitive behavioral therapy, etc.) has been challenged for a number of reasons. Antidoxasticists, who mainly draw from interpretivist and functionalist theories of belief, point out

that delusions don't fit the stereotypical rational or causal profiles of belief. By contrast, several authors have defended doxasticism by introducing certain amendments to the interpretivist and functionalist frameworks (i.e., revisionist doxasticism) or rejecting them in favor of alternative theories of belief (i.e., non-revisionist doxasticism). Revisionist doxasticists (Bayne & Pacherie, 2005; Bortolotti, 2010, 2012) establish more relaxed criteria for something to count as a belief, and invoke an "all-things-being-equal" clause to argue that the inconsistencies displayed by some people with delusions can be properly excused by non-standard features of the case. Clutton (2018), by contrast, rejects functionalism and interpretivism on the grounds of their "anti-realist" tendencies and proposes instead to adopt a cognitive-phenomenological theory of belief, according to which to believe that *p* just amounts to having a disposition to "mentally assent" to *p* whenever the possibility of *p* is entertained in one's mind. Despite their differences, we've highlighted two common desiderata that motivate defenses of doxasticism: a) the scientific desideratum, according to which doxasticism leaves us in a better position to understand the causes of delusions; and b) the ethico-political desideratum, according to which doxasticism provides a way to understand delusions in terms of their intelligibility, and hence stands as a conceptual barrier against undue deagentializing practices and possible forms of unjust treatment.

In Chapter 6, we've argued that neither revisionist nor non-revisionist doxasticisms can live up to both desiderata. Revisionist doxasticisms, on the one hand, fail to meet the scientific desideratum. In particular, they end up being forced to assume some kind of ascriber-relativist view of the truth of belief ascriptions, whereby the truth of a belief ascription might be dependent on the ascriber's standards; thus, to characterize delusions as beliefs wouldn't be informative regarding their possible causes. On the other hand, Clutton's non-revisionist approach fails to meet the ethico-political desideratum. The reason is that his cognitive-phenomenological theory, no matter how it's construed, renders a normatively inert notion of belief; one which cannot rationalize behavior nor hence inform judgements about the person's agency. We've then argued that doxasticism can and should be defended -not on scientific, but on ethico-political grounds-, and that our non-descriptivist approach to folk psychology allows for a more robust defense of it. Firstly, it captures the plurality of norms at play in belief ascription practices -sometimes, we privilege the interpersonal norm of first-person authority over considerations about the overall consistency of someone's behavior (as it seems to be the case with delusions); other times, we don't. Secondly, we've claimed that this allows us to see not only how our belief ascription practices in fact work, but also why they should continue to work as such. The reason is that antidoxasticism, in its allegiance to an idealistic conceptualization of our belief ascription practices, might promote

unwarranted deagentializing practices towards people with delusions. By contrast, from a non-descriptivist perspective, doxasticism can be viewed as a sounder and more desirable conceptualization policy: by default, we should take the person's belief self-ascriptions at face value. Finally, we've seen how non-descriptivism, in distinguishing the main purposes of folk and scientific psychology, protects doxasticism from eliminativist tendencies; no matter how far scientific psychology goes, our folk-psychological, doxasticist understanding of delusions will make sense on its own terms.

In Chapter 7, we've focused on whether the kind of doxasticism endorsed by Clutton -scientific doxasticism, or the doxastic conception of delusions at play in traditional cognitivist models of delusions- actually provides a good scientific model of delusional experiences. We've first reviewed two main traditional cognitivist approaches to delusions: cognitive behavioral therapy for psychosis (CBTp) and cognitive neuropsychiatry. These two complementary approaches share a conceptualization of mind and cognition in information processing terms, and hence explain delusions as departures from "normal" information processing mechanics. In particular, the underlying cognitive models of delusions emphasize the role of specifically cognitive factors like Beck's cognitive triad (i.e., the person's "deep" schemas about the world, the future, and themselves, especially the latter), and various cognitive biases (i.e., jumping to conclusions, attributional biases, and Theory of Mind deficits). We've later conducted a narrative review of the existing evidence in support of CBTp's efficacy. We've seen that, although there's evidence that CBTp in general has small to moderate effects, its efficacy in the case of people with delusions is more ambiguous. Furthermore, the evidence doesn't support cognitivist explanations of delusions; only jumping to conclusions and negative self-schemas are significantly increased in people with delusions, and none of them have been found to mediate the efficacy of CBTp interventions with these people, not even when they were specifically directed at those factors. After that, we've pointed out that these results might be partially explained by the intellectualist understanding of the mind that characterizes traditional cognitivism, i.e., the idea that having a certain belief or acting in accordance with it is a matter of entertaining certain regulative propositions before the mind's eye and then acting in consequence. In particular, we've argued that, in the case of delusions, this leads to a neglect of the variable environmental factors that might be at play in different cases of delusions and of the various intervention techniques aimed at tackling them.

Taking all this into account, we've claimed that non-cognitivist, functional analytic approaches might provide a better model for the intervention with people with delusions. In Chapter 8 we've conducted a narrative review of these approaches. We've separately

considered two main functional analytic approaches: traditional (applied) behavior analysis and Acceptance and Commitment Therapy (ACT). The former conceptualizes delusions as non-normative verbal behaviors and emphasizes the need for conducting pre-treatment Functional Behavioral Assessments (FBA). The latter, by contrast, conceptualizes delusions as inflexible rules primarily maintained by their avoidance function, and emphasizes the need to change the person's relation to their own experiences. The evidence for FBA-based interventions with people with delusions seems promising, but further quantitative syntheses should be conducted. In the case of ACT, when measured in its own terms, the existing evidence suggests that its benefits are mostly due to its effect on hallucinations, rather than delusions. We've then argued that the efficacy and perceived utility of functional analytic approaches might be hindered by a residual commitment to intellectualism. On the one hand, traditional behavior analysis exhibits this residual commitment in its narrow operationalization of delusions as non-normative verbal behaviors. Hence it faces the superficiality objection raised by cognitivists, which, viewed from our non-descriptivist point of view, amounts to saying that mental-state ascriptions, and not mere descriptive reports of what a person says, set the bar to clinically significant changes. On the other hand, ACT researchers stick their feet further into the intellectualist quicksand in their attempts to reformulate beliefs in terms of causally effective verbal behavior (e.g., behavioral rules) and relational responding processes. This might lead them to neglect other potential sources of environmental control, hence explaining their ambiguous evidence status. We've finally discussed how our non-descriptivist approach to folk psychology might benefit functional analytic approaches. By pointing out the difference between norm-following behavior (e.g., "acting on the grounds of one's beliefs") and rule-governed behavior (i.e., behavior controlled by verbal rules), it frees functional analytic approaches from straitjacketing and preconceived assumptions about the causes of psychopathology and encourages a more encompassing, values-based approach to the determination of therapeutic goals in mental health practice.

As we warned in the Introduction, this dissertation has been a long journey -longer than Yellow's, indeed. We've come across multiple debates, addressed from diverse disciplines and subdisciplines, from different angles and in different languages. It's -we hope it merits the name- a piece of work in the philosophy of mental health. If any, the value of this dissertation mainly resides in the kind of conceptual bridges that we've tried to build, as well as in its vindication of the relevance of philosophical inquiry for mental health research and practice, often overshadowed by the specter of scientism.

Other than that, we think that its main contributions lie in the implementation of the pragmatist kind of non-descriptivism endorsed here to the philosophy of mental health. In

general terms, the attractiveness of this approach lies in its ability to accommodate both mind and normativity within a naturalist worldview, hence dissolving a tension which, as we've seen, pervades discussions in the field of mental health since at least the second half of the 20th century. In the particular case of delusions, this approach helps to: a) explain and defend, in a most robust way, the ethico-political benefits of doxasticism; and b) dispel straitjacketing assumptions from clinical practice. Specifically, by pointing out that beliefs are not material entities with causal powers, nor belief-guided behavior the result of entertaining representations of any sort -neither mental nor verbal-, our non-descriptivist approach encourages the case-by-case determination of therapeutic goals and the case-by-case analysis of how best to achieve them; at the same time, it allows us to generally characterize delusions as beliefs, which avoids undue deagentializing practices towards people with delusions.

However, as a piece of work in the philosophy of mental health, it hasn't said much -explicitly, at least- about the four major themes that we identified in the Introduction (i.e., the analogy, boundary, priority, and integration problems). We should like to finish now by coming back to these four themes. We'll sketch out some further reflections on these matters, so as to establish several possible paths towards the future development of a non-descriptivist approach to the philosophy of mental health.

9.2. Further notes on non-descriptivism and the philosophy of mental health

Our starting examples (Purple's mad madness and Yellow's Karnatahclaniard madness) helped us to illustrate the main topics of the philosophy of mental health. On the one hand, these examples prompted issues about the role of *norms and values* -and the social niches which institute them- in the determination of what counts as pathological and what doesn't. Purple's mad madness was mad, precisely, because there didn't seem to be anything wrong with them, despite their different neural make up. It was mad because, were someone to insist that Purple had a mental health problem, we would feel inclined to respond "Yes, ok, Purple's neural make up is *different*, but there's nothing intrinsically *wrong* about that. Look at them! They're clearly healthy and well!". By contrast, our interpretation of Yellow's case seemed to vary according to what Karnatahclaniards themselves thought about the case. If they considered Yellow's behavior and experience as something relatively normal -a "second adolescent phase", we said- we wouldn't be so inclined to say that they have "a problem"; we would just say that they are different from us (and a bit peculiar, to say the least). But, in our case, Karnatahclaniards evaluate Yellow's behavior, cognition, and experience as we would

evaluate Purple's if they in fact displayed the same behavioral, cognitive, and experiential patterns. In this case, we feel inclined to say that there's something *amiss* with Yellow, that their suffering constitutes a call for help and empathic understanding. If someone insisted that we cannot pronounce ourselves about the wrong character of Yellow's suffering, most of us would feel inclined to disagree.

On the other hand, our initial examples also invited us to think about the inter-play among different kinds of facts, addressed from different scales of analysis, in mental health. Where should we look for Purple's or Yellow's mental health problems, if they had them? In both cases, their neural (or bodily) states and processes could be relevant for determining possible interventions -if required-, but they didn't seem to be so relevant for determining whether there was or not a problem in the first place. Rather, it was their overall patterns of action and reaction, and of interaction with their environment, what seemed to matter for determining their mental health status. In addition, Yellow's example implied the role of environmental variables in the development of psychopathology (e.g., their four years of social isolation, the stress related to their PhD studies, etc.). In this sense, these examples prompt questions about the constitutive and causal role that different kinds of factors (e.g., biological, psychological, social, etc.), might play in the analysis of mental health problems and their origin and maintenance.

The four major themes of the philosophy of mental health appear in these cases. These were, as we saw, the analogy problem (i.e., the issue of the analogy between mental and somatic health problems), the boundary problem (i.e., the issue of the relation between psychopathology and social deviancy), the priority problem (i.e., whether some scale of analysis should enjoy causal or constitutive priority over the rest), and the integration problem (i.e., the issue of the integration among different scales of analysis). As we view it, the first two are more closely related to the role of norms and values in the conceptualization of mental health problems; the second two, in contrast, are more tightly related to factual questions regarding how best to intervene on them. Hence, in what follows, we'll discuss the possible contributions of our non-descriptivist approach to these matters separately.

9.2.1. Non-descriptivism and the analogy and boundary problems

As we saw in Chapter 1, concerns about the analogy between mental and somatic disorders and the possibility to draw a firm boundary between "mad" and "bad" lie at the core of the classic criticisms against the medical model. Before we address these problems themselves, however, let us say something about the convenience of the medical understanding of psychological problems. In this dissertation, we've clearly opposed the biomedical, or strong interpretation of the medical model; we think that understanding mental health problems

as primarily originated in internal machineries gone wrong just diverts attention from the clear contextual sources of psychological distress. However, we remain agnostic as to the convenience of framing psychological problems in medical terms and the use of diagnostic labels from a minimalist standpoint. To be sure, as long as it invites individualistic and internalist thinking about the origin of mental health problems, or as long as it pathologizes mere forms of social deviancy (or even common experiences within a lifetime), we think that the medical understanding of psychological problems must be rejected. But we might separate the two issues here. Framing psychological problems in medical terms and having certain diagnostic kinds with which to identify different forms of psychological suffering subserves a number of important practical purposes, which go far beyond the alleged "interprofessional communication" function that has been so often invoked. In administrative terms, it helps people whom are suffering from psychological problems access a number of social security resources (i.e., social benefits, sick leaves, etc.), which are themselves a necessary material condition for recovery in multiple cases. Diagnostic labels also play an important function as hermeneutical tools for people to understand what's going on with them, which might help them acquire a certain distanced and less self-blaming perspective on their suffering. Finally, pro-diversity movements in the realm of mental health (i.e., neurodiversity or psychodiversity movements) are increasingly reappropriating, politicizing, and depathologizing these diagnostic labels to advocate for the recognition and revaluing of certain forms of psychological diversity (e.g., see Chapman 2020; Singer, 1999).

As long as diagnostic labels play these functions, they might be worth keeping - better in the later kind of depathologizing and repoliticized manner, perhaps. But what interests us here are the following questions: regardless of whether talk of "mental health" problems or "mental disorders" is worth keeping or not, are mental health problems *analogous* to somatic health problems? And, how are we to spell out the difference between social deviancy and psychopathology?

Let's begin with the analogy problem. Commonly, answers to this problem recur to one or another notion of "illness" or "disorder" in general. As some authors have claimed (see Fulford & van Staden, 2013; Thornton, 2007), traditional defenses and criticisms of the analogy (e.g., Kendall, 1975; Szasz 1961/1974) have typically drawn from a narrow definition of "disorder"; namely, one which assumes that the notion of disorder can be spelled out in value-free terms (i.e., in terms of pure descriptions of material states of affairs). The disputes, then, concerned one or another descriptive notion of "pathology" -e.g., Szasz's Virchowian understanding of it vs. Boorse's (2014) redefinition of it as "a state of statistically species-subnormal biological part-functional ability, relative to sex and age" (p. 684)-, as

well as whether these criteria captured mental health problems or not. More recent approaches have drawn attention to the fact that neither mental nor somatic health problems can be spelled out in purely descriptive, value-free terms (e.g., Fulford & van Staden, 2013; Graham, 2010b; Thornton, 2007; Varga, 2015). These "having it both ways" approaches, as they've been called, argue instead that the consideration of *any* kind of condition as pathological necessarily involves value judgements. Hence, they conclude that there's no essential difference between mental and somatic health problems. The only difference between them would lie in the amount of *agreement* or *disagreement* among different stakeholders within a given community regarding the values and norms at stake in different cases, which would determine whether something can be *evaluated* as a pathology or not. It's assumed, then, that while in the case of somatic health problems there's a minimum degree of value variability, in the case of mental health problems there's a maximum degree of dissonance as to what and whose values should determine the categorization of some condition as a kind of pathology.

We won't go into detail here about these having it both ways approaches. We just want to point out that at least some of these approaches are attractive from a non-descriptivist point of view; in fact, some of them draw from a somewhat similar non-descriptivist approach to value judgments or rely on similar Wittgensteinian and Rylean arguments (e.g., Fulford & van Staden, 2013; Thornton, 2007). However, we think that these approaches fall short of their non-descriptivist analyses. In particular, they assume a non-descriptivist approach to the disorder aspect of the notion of "mental disorder", but fail to apply that very same analysis to the *mental* aspect of the notion. Hence, they might be right in pointing out that the diagnosis of both somatic and mental health problems relies on a "bedrock" of values, but that doesn't mean that both mental and somatic health problems are analogous.

As we view it, these authors reason as if in both mental and somatic health problems, diagnoses had a dual function: to describe some material state of affairs, and to evaluate it as pathological. Our non-descriptivist analysis goes further: while this might capture how the diagnosis of somatic health problems works, it doesn't fit mental health assessment practices. The "somatic" in "somatic disorder" can be spelled out in purely descriptive terms; the "mental" in "mental health" cannot. Of course, in mental health we talk about a person's patterns of actions and reactions; however, what qualifies these as "signs" of *mental* (vs. somatic) health problems is that these patterns of behavior are assessed in primarily personal (hence evaluative), rather than subpersonal terms. Mental health problems primarily affect one's status as an *agent*, not as a mere organism. They are thus constituted by alterations in a person's ability to find meaning and value in their relation to themselves and their social

world; or, to put it in de Haan's (2020) terms, in their "existentialized sense-making" abilities (p. 11). That's why mental health problems "dissolve if one succeeds in changing one's way of interacting with the world" while "secondary effects of somatic disorders on sense-making (…) do not disappear by interacting with the world in a different way"; in other words: mental health problems, *qua* mental states, "are not of the brain, not even of the body, but of *persons*" (p. 11, emphasis added).

Again, this doesn't mean that understanding psychological problems in medical terms cannot have beneficial consequences, nor that mental health services and resources are a mere form of statal control of the individual's sacred individuality, nor that the attribution of the "sick role" to someone on account of their psychological suffering cannot make sense sometimes (e.g., to give them time and space for rest and recovery). In this sense, we don't think that the medical understanding of psychological problems is a necessarily perverse metaphor. But we do think that, as a matter of fact, it's a metaphor -or, at least, that somatic and mental health problems are not analogous to each other. The disanalogy between mental and somatic health problems is best revealed in the fact that it's utterly strange to speak (in a literal sense) of "sick wills", "ill beliefs", or "disordered desires", just like it's utterly strange to literally speak of an individual's "sick morality" or "pathological values". Minds, norms, and values might be involved in the determination of what counts as "ill" or "disordered"; however, they are not *themselves* the kind of thing of which it makes sense to say that are "ill" or "disordered", in literal terms; to think otherwise amounts to the kind of category mistake pointed out by Ryle (1949). Our non-descriptivist approach to folk psychology helps us understand why talk of "mental disorders" shouldn't be understood in literal terms (i.e., as analogous with talk of somatic disorders); at the same time, it resists the tendency to call these expressions "myths" or disregard mental-state ascriptions (including mental health assessments) as lacking truth value.

This has also consequences for the boundary problem, or the possibility to distinguish between "psychopathology" and "social deviancy". Once again, our discussion of the disanalogy between mental and somatic health problems reveals that mental health assessment is a normative or evaluative task through and through; from the determination of what counts as being in certain mental states, to the determination of which of these mental states are "wrong" or "amiss" in some sort of way. Social norms and values are deeply entrenched in both kinds of judgement; hence, telling "psychopathology" and "social deviancy" apart is not a straightforward task, to say the least. As we view it, what (and whose) norms and values should count in determining what falls on the side of psychopathology (and hence merits the attribution of the sick role, with its related benefits and disadvantages), and what falls on the

side of mere social deviancy (and hence must be respected if it harms no one, such as one's sexual orientation or gender identity) is something that will probably vary across types of mental health problems and across particular cases. In this sense, our non-descriptivist approach aligns with the kind of value particularism endorsed by Thornton (2007, 2013, 2014); no general maxims nor principles will allow us to ascertain, for all and every possible case, what should be intervened on.

However, if we were to speak in general terms, we think that a good rule of thumb is to be found in the consideration of how a person's behavior and experiences align or not with their *own* values. This goes in line with functional analytic approaches like ACT and recent enactive approaches (e.g., see de Haan, 2021; Nielsen, 2021), which emphasize the *self-defeating* nature of mental health problems, i.e., the fact that at least many of them can be primarily characterized in terms of harmful systematic deviations from one's own preferred ways of living. From this point of view, the "hallmark of psychopathology" (vs. mere social deviancy) is, or should be, the presence of psychological suffering due to a conflict between one's doings and the values that one endorses.

We think that our non-descriptivist approach yields interesting insights on this matter, which would be worth developing in future research. In a nutshell, its main potential lies in its ability to provide a non-reductivist and non-individualist view of what a person's values and value conflicts amount to. An example of this can be seen in contrast with "reifying" and subjectivist views of values that follow from their strict identification with what the person says their values are, or with the "verbal rules" that the person emits (e.g., ACT), as well as other kinds of individualistic approaches that seem to assume that one's values are somehow chosen at will (e.g., Szasz's approach). In short, from our perspective, the determination of what a person's values (i.e., their evaluative *beliefs*) are cannot be reduced to a mere description of what the person says their values are; rather, value clarification requires viewing a person's behavior in relation to the myriad *social* practices and institutions that configure the person's socio-cultural niche and ground the norms they follow. Value clarification thus is an irreducibly hermeneutical and contextualized task, which requires engaging in interactive evaluative and regulative "mindshaping" practices to differentiate what are the norms that we actually follow, those that we would like to follow, and those that others want us to follow. Our non-descriptivist approach gives meaning and truth-evaluability to these interpretative practices without reducing them to the formulation of empirical hypotheses about the verbal rules that a person emits (nor about hidden inner processes that mediate their relation with the world). In addition, it provides a non-individualist understanding of what "one's own values" amounts to, and hence of the "self-defeating criterion" to distinguish

psychopathology from social deviancy; what differentiates mental health problems from diverse forms of social deviancy is that the former, unlike the latter, involves a particular kind of socio-normative conflict: one which arises not just between an individual's actions and *any* social values or expectations, but between their actions and those social norms and values that they self-profess, or that they enact or *express* when they're not trapped in the common loops of mental distress. This provisional non-intellectualist definition could be, from our perspective, useful for further research on the boundary problem.

9.2.2. Non-descriptivism and the priority and integration problems

Now let's turn back to the priority and integration problems. Should we privilege any given scale of analysis (e.g., the biological, the psychological, the social, etc.) in the conceptualization of mental health problems or the intervention on their causes? And how should we think of the integration among different explanatory projects in mental health research and practice?

First of all, as we view it, it goes without saying that a proper general framework for mental health services should count with many different professionals working at different levels or scales of analysis. In this sense, we think that scientific investigation on the *causes* of mental health problems should be addressed within a multi-disciplinary attempt to spell out the different kinds of causal dynamics involved in their development and maintenance. Whether all these dynamics can be properly unified within an integrative *ontological* framework is something about which we're happy to remain agnostic; after all, it's likely that different scientific languages and methods prove to be incommensurable -something which simply restoring to "emergent" properties might not resolve. In any case, we think that, within a common naturalist framework, a healthy epistemological pluralism about the different possible causal-explanatory approaches to mental health is recommendable, and that integrative efforts, inasmuch as they yield new productive lines of research or maximize our intervention abilities on psychological suffering, should be welcome.

In the same vein, we think that research on the bodily processes that people with mental health problems undergo is obviously necessary for gaining a proper understanding on how psychological suffering affects the organism; among other reasons, because this helps uncover the profound effects that the stressful or unjust social dynamics to which we're sometimes subject can have on us. Our stance on psychopharmacology and the use of psychiatric drugs to alleviate mental health problems is similar; to the extent that these drugs are used and prescribed in a *responsible* manner, and to the extent that the people who take them are duly informed of both their possible benefits and secondary effects, we think that psychiatric medications can be helpful in many situations -for instance, by providing

the necessary conditions (e.g., rest, reduced agitation, etc.) for a person to start their recovery process.

That said, however, a different question is whether this or another scale of analysis should be given *conceptual* priority over the rest, and hence endorse it as our focal or basic scale of analysis, i.e., that where our primary "subject matter" is to be found. We think that it's clear from the present dissertation that our sympathies in this case primarily lie with psychological forms of intervention -especially, with those that take the interaction between a person and their environment as their starting point. In this sense, we agree with functional analytic approaches -as well as recent enactivist proposals- that the basic unit of analysis in mental health research and practice is, and should continue to be, the person's *behavior*, broadly construed, and not their brain circuits.

Our non-descriptivist framework pairs well with these assumptions. Regarding the tension between cognitivist and non-cognitivist psychological models, our non-descriptivist approach helps to point out that the psychological scale of analysis should be understood in irreducibly contextual terms. Mental states are not inner facts which mediate between perception and action; likewise, mental health problems are not to be found in hypothetical inner information processing mechanisms gone wrong. From our non-descriptivist point of view, the field of mental health is primarily involved with a person's interaction with their natural and social environments.

Likewise, this approach helps us to see what's exactly wrong about "neuro-reductivist" (or "neuro-eliminativist") tendencies in mental health that establish the "brain circuitry" level as the focal unit of analysis in mental health research and practice. As we've seen, the "mental" in "mental health" cannot be reduced to neither a mere description of an individual's brain states nor to a mere description of their behavioral patterns; it cannot be reduced to mere statements of facts whatsoever. However, it's also clear from our analysis that not all types of facts play an equal role in the determination of what counts as having certain mental states (e.g., having certain beliefs about oneself or the world). When we engage in folk-psychological interpretative practices, it's each other's *actions and reactions* (motor, symbolic, inferential, phenomenological, etc.) which we're putting under evaluation to decide whether they "merit" certain mental-state ascriptions or whether they are properly made intelligible by them. Our brains -our bodies, in general- allow us to act and react, but are not themselves what is under evaluation when we assess each other's mental states. Insofar as certain bodily processes are found to be altered in certain mental health problems, including neural ones, the "brain circuitry" level is obviously relevant for mental health research; but to elevate it as the focal unit of analysis in mental health is, once again, to fall

prey of the category mistake that characterizes the dogma of the Ghost in the Machine. In short, brain circuits have an *enabling* role in mental health problems; the person's behaviors and experiences, assessed from particular evaluative frameworks, *constitute* the subject matter itself of mental health research and practice. The futuristic vision of mental health services that we tried to depict in our introductory examples still attracts many; we think that our non-descriptivist approach provides sound reasons for realizing that it's essentially a wrong vision, one based on misplaced assumptions about the nature of mind and its adversities.

To conclude, we hope to have provided compelling arguments in support of our pragmatist, evaluativist, and regulativist kind of non-descriptivist framework, as well as for its underlying focus shift from the ontological puzzles of Cartesianism to the analysis of our linguistic practices in the field of mental health. It's time to overcome the "you are more dualist than I am" games (Pinedo-García, 2020) that have for so long pervaded discussions about the status of the "mental" in "mental health", and we think that this dissertation is a step in that direction.

Capítulo 9
Conclusión. Hacia una filosofía de la salud mental sin espejos

Es el año 2022. Ha comenzado una guerra. La pandemia sigue campando a sus anchas. Una nueva crisis económica está en camino. Y muy probablemente también una de salud mental. La desesperanza, la angustia, el pánico, la desrealización, la paranoia, todas ellas prosperan en medio de este clima de crisis. Algunos creen que debe haber una mano negra moviendo los hilos detrás de estas macabras escenas. Algunos creen que todo se reduce a la dinámica intrínseca expansiva y autodestructiva del capitalismo. Algunos tienen la esperanza de que las cosas mejoren, o simplemente de que no empeoren.

Mucho ha ocurrido durante la redacción de esta tesis, y estos enunciados introductorios sólo describen -en un sentido amplio del término- parte de todo ello. Uno de nuestros principales objetivos ha sido precisamente introducir algunas distinciones entre los diferentes tipos de afirmaciones que están en juego aquí -específicamente, entre las que se utilizan para describir algún estado de cosas temporal y espacialmente localizable (por ejemplo, "Ha comenzado una guerra") y las que dicen algo sobre los estados mentales de alguien (por ejemplo, "Algunos esperan que las cosas mejoren"). La razón, hemos dicho, es que hay un vínculo íntimo entre mente y normatividad; es decir, entre nuestras prácticas interpretativas basadas en la psicología del sentido común -los "juegos mentales" que caracterizan buena parte de nuestras interacciones con los demás, y la dimensión normativa de la acción, que implica la posibilidad de error y éxito, mérito y demérito, corrección e incorrección, etc. Por tanto, lo mental no es reducible a ni sustituible por meras descripciones de estados de cosas materiales. Pero, al mismo tiempo, hemos subrayado que esto no significa que el vocabulario mental apunte a criaturas míticas, ni que las descripciones de estados mentales carezcan de valor de verdad; de lo contrario, nuestras inclinaciones naturalistas o antinaturalistas serían, en última instancia, indistinguibles de meros versos del *Jabberwocky* de Lewis Carroll.

Hemos visto cómo una concepción pragmatista y antidescriptivista de la mente, basada en la obra de Wittgenstein y Ryle, ofrece una visión diferente de nuestras prácticas interpretativas (que resalta su dimensión evaluativa y regulativa) y nos permite reconciliar dos ideas aparentemente opuestas: a) que las atribuciones de estados mentales no describen ningún hecho en particular sobre un agente; y b) que eso no significa que carezcan de valor de verdad. Al fin y al cabo, aunque quizás de forma diferente, todas las afirmaciones de la introducción de este capítulo son *verdaderas*.

Nuestro principal argumento ha sido que este tipo de enfoque proporciona un marco conceptual más sólido para la investigación en salud mental y la práctica clínica; "una filosofía de la salud mental sin espejos", capaz de proveer una respuesta más adecuada al problema de lo mental y al problema de la normatividad. De este modo, nuestra propuesta es capaz de resistir las diversas tendencias reduccionistas y eliminativistas que permean gran parte de los debates entre modelos terapéuticos, así como las inclinaciones escépticas, individualistas y, en cierto sentido, antinaturalistas de ciertos enfoques críticos (por ejemplo, el de Szasz). Hemos destacado también las implicaciones de nuestro marco antidescriptivista respecto al debate sobre la conceptualización estándar de los delirios en términos de creencias, y sus consecuencias para la intervención con personas con delirios. En pocas palabras, esta aproximación nos permite conservar lo mejor de la comprensión doxasticista de los delirios (es decir, su utilidad ético-política), al tiempo que disipa el espectro del cartesianismo que encorseta su comprensión clínica y científica.

En este capítulo, resumiremos las principales conclusiones de esta tesis doctoral y señalaremos algunas líneas de investigación interesantes para abordar en futuros trabajos. En la sección 9.1., haremos un resumen de esta disertación, destacando sus principales aportaciones. En la sección 9.2., haremos un esbozo de posibles líneas de investigación futuras a desarrollar en el marco de nuestra aproximación antidescriptivista al campo de la salud mental. Para ello, volveremos sobre los cuatro grandes temas de la filosofía de la salud mental que vimos en la introducción (es decir, los relacionados con los problemas de la analogía entre problemas de salud mental y física, del límite entre psicopatología y desviación social, de la priorización de niveles de análisis en salud mental y de la integración entre diversos niveles de análisis), señalando cómo nuestro enfoque anti-descriptivista puede aportar nuevas ideas sobre estas cuestiones.

9.1. Resumen y principales aportaciones de la tesis doctoral

En la primera parte de la disertación, nuestro principal objetivo fue explorar los fundamentos conceptuales de ciertos debates recurrentes, relativos a la naturaleza de los problemas

de salud mental, así como mostrar cómo nuestra concepción anti-descriptivista de la mente ofrece un mejor punto de partida para abordarlos. Comenzamos en el capítulo 1 con una visión general de la historia conceptual de la salud mental y de los principales modelos terapéuticos clásicos y contemporáneos. Vimos que el llamado "modelo médico", que muchos toman como el enfoque terapéutico predominante, admite en realidad varias interpretaciones. De acuerdo con una interpretación mínima, el modelo médico simplemente equivale a la comprensión de los problemas de salud mental en términos médicos, lo cual puede resultar útil para diversos fines (toma de decisiones en cuestiones administrativas, elaboración de estudios epidemiológicos, etc.). De acuerdo con una interpretación más fuerte, los trastornos mentales son clases naturales; nuestras taxonomías psiquiátricas, por tanto, aspiran o deberían aspirar a "cortar la naturaleza por sus junturas" y ayudarnos así a establecer la verdadera naturaleza neurobiológica de los problemas de salud mental. Ambas interpretaciones, especialmente esta última, han sido puestas en cuestión. Los enfoques críticos en salud mental atacaron las dos piedras angulares de la legitimación de la psiquiatría como disciplina médica (es decir, la analogía entre los trastornos mentales y somáticos y la distinción entre "psicopatología" y "desviación social"), entendiéndolas como meros mitos. Los modelos psicológicos, por su parte, criticaron el modelo médico por sus supuestos internalistas o biologicistas, priorizando en cambio el análisis de los factores psicológicos (ambientales, cognitivos, etc.) implicados en el origen y mantenimiento de los problemas de salud mental. A modo de síntesis, el modelo biopsicosocial surgió como un intento de proporcionar un enfoque conciliador para las ciencias de la salud en general, asumiendo que tanto los factores biológicos como los psicosociales son igualmente relevantes para explicar y abordar los problemas de salud mental. Por último, diversos modelos contemporáneos (la "psiquiatría de precisión", los enfoques analítico-funcionales contemporáneos y la aproximación enactivista a la salud mental) han intentado articular distintas soluciones tanto a las crisis de fiabilidad y validez del modelo médico como a los problemas de integración del modelo biopsicosocial.

Al final del capítulo 1, expusimos las dos principales cuestiones conceptuales que subyacen a estos debates: a) el problema de lo mental, que comprende una serie de cuestiones relativas a la relación ontológica y epistémica entre mente y naturaleza; y b) el problema de la normatividad, relacionado con el lugar de las normas y valores en una cosmovisión naturalista, así como con el choque entre las imágenes manifiesta y científica del mundo y los seres humanos. En el capítulo 2, nuestro principal objetivo fue ofrecer una explicación plausible del origen de estos problemas y exponer las diferentes filosofías de la mente que subyacen a los diversos modelos terapéuticos. Comenzamos remontándonos a la teoría

cartesiana de la mente, destacando sus compromisos ontológicos y epistemológicos (dualismo, factualismo, causalismo, intelectualismo y representacionalismo). Señalamos también la concepción semántica de las atribuciones de estados mentales subyacente a dichos fundamentos: el descriptivismo, o la idea de que las atribuciones de estados mentales describen o representan algún estado de cosas determinado. A continuación, vimos varios intentos contemporáneos de abordar el principal rompecabezas ontológico del cartesianismo, el problema mente-cuerpo. Partiendo de un compromiso común con el naturalismo ontológico (definido por las ideas del monismo, el materialismo y el principio del cierre causal), los enfoques naturalistas implementan tres estrategias diferentes para dar cuenta de lo mental: a) acomodar los objetos mentales dentro de una ontología naturalista (enfoques ontológicamente conservadores como el reduccionismo directo, el funcionalismo y el emergentismo); b) mantener los conceptos mentales sólo en la medida en que demuestren tener algún valor explicativo, eliminándolos en caso contrario (enfoques ontológicamente revisionistas como el eliminativismo discursivo); o c) descartar los estados y procesos mentales como criaturas míticas o ficciones explicativas (enfoques ontológicamente radicales como el eliminativismo directo). Después de mostrar cómo estas diversas aproximaciones naturalistas a lo mental subyacen a los distintos modelos terapéuticos, concluimos que ningún es capaz de ofrecer una explicación adecuada de la dimensión normativa de lo mental; esto, a su vez, los hace incapaces de proporcionar un marco conceptual adecuado para la investigación e intervención en el ámbito de la salud mental. Para salir de esta situación, insistimos en la necesidad de desplazar el foco de análisis, pasando de los rompecabezas ontológicos de lo mental al análisis del significado y la función de las atribuciones de estados mentales.

En el capítulo 3, profundizamos en la cuestión del descriptivismo, analizando cómo ha restringido la gama de posibles respuestas al problema mente-cuerpo, y cómo este nos lleva a formas insostenibles de naturalismo y normativismo. Comenzamos situando el descriptivismo en el núcleo de la imagen estándar de la psicología del sentido común, compartida por todas las aproximaciones al problema mente-cuerpo discutidas en el capítulo anterior. De acuerdo con dicha imagen, la psicología del sentido común desempeña un papel explicativo principalmente descriptivo y causal. A continuación, distinguimos las dos principales versiones del descriptivismo en el nivel semántico: a) la versión superficial y afirmativa, según la cual el significado o el contenido de las oraciones declarativas (entre ellas, las atribuciones de estados mentales) reside en una descripción o representación de algún posible estado del mundo; y b) la versión profunda y condicional, que equivale a la afirmación de que sólo las oraciones declarativas que describen con éxito algún estado de cosas son veritativo-evaluables. A continuación, vimos cómo el naturalismo ontológico puede ser

entendido como la imposición de ciertas restricciones sobre lo que puede ser un posible estado del mundo, y por lo tanto sobre lo que puede contar como una descripción "exitosa". Esto lleva al supuesto de la traducibilidad, según la cual la compatibilidad entre mente y naturaleza depende de la posibilidad de reducir o traducir las descripciones de estados mentales a descripciones de estados materiales. Esto deja a los naturalistas con sólo dos formas de explicar lo mental: el compatibilismo reductivo o el incompatibilismo no reductivo. Ambos, como vimos, conducen a un tipo de naturalismo defectuoso y autodestructivo; uno que, en su incapacidad para dar cuenta del carácter veritativo-evaluable y la fuerza normativa de las atribuciones de estados mentales -que sólo el compatibilismo y el antirreduccionismo pueden retener, respectivamente-, queda él mismo eliminado o reducido a meras descripciones de los estados neurales o el comportamiento verbal de quienes defienden el naturalismo. Las aproximaciones antinaturalistas, por otra parte, no ofrecen una mejor alternativa: aparte de su carácter anticientífico, son presa del argumento Wittgensteiniano contra el seguimiento privado de reglas, lo que finalmente conduce a un tipo de normativismo autodestructivo. Por último, dijimos que, para evitar este dilema, al que nos referimos como el "rompecabezas de la traducibilidad", necesitamos encontrar algún tipo de explicación no reduccionista, pero compatibilista, de la relación entre mente y naturaleza; para ello, debemos rechazar el descriptivismo.

En el capítulo 4 expusimos nuestra aproximación antidescriptivista a lo mental. A diferencia de otros enfoques antidescriptivistas, el nuestro rechaza el descriptivismo tanto en el nivel pragmático como en el semántico, así como en las versiones superficial y profunda de este último. En su lugar, asume el pluralismo funcional (la idea de que el lenguaje puede usarse para muchos propósitos además de hacer afirmaciones sobre cómo es o deja de ser el mundo) y el pluralismo sobre la verdad (es decir, la idea de que el valor de verdad de diferentes oraciones declarativas puede determinarse de distintas maneras en distintos tipos de afirmaciones). Partiendo de una lectura pragmática de la obra de Wittgenstein y Ryle, nuestra aproximación asume que el significado (y las condiciones de verdad) de una determinada expresión reside en sus posibles usos en diferentes prácticas comunicativas regladas (es decir, en distintos juegos del lenguaje). De acuerdo con esta perspectiva, los posibles usos de una expresión están determinados por las relaciones inferenciales o de justificación que una expresión mantiene con otras expresiones y cursos de acción (es decir, su geografía lógica); a su vez, estas conexiones inferenciales se basan en las distintas prácticas sociales en las que somos entrenados por nuestra comunidad de referencia. Saber lo que significan nuestras proferencias -y cuál es su valor de verdad, cuando ello es pertinente- se considera, pues, una cuestión de "saber cómo", más que de "saber que". Partiendo de esta base, hemos

dado tres argumentos (no durabilidad, dependencia de las condiciones de verdad y fuerza normativa) que apoyan la idea de que las atribuciones de estados mentales tienen una función evaluativa y regulativa, más que descriptiva; su objetivo principal no es predecir y controlar la conducta, sino racionalizarla y justificarla. Por último, también hemos visto que esto no significa que las atribuciones de estados mentales no sean veritativo-evaluables; más bien, su verdad o falsedad depende de las muy diversas normas que las hablantes competentes de un lenguaje siguen en sus prácticas interpretativas (autoridad de primera persona, coherencia general, etc.). En contra de la idea de que existe una regla de oro que regule dichas prácticas, nuestro enfoque antidescriptivista ofrece una visión pluralista de los criterios que determinan la verdad o la falsedad de distintas atribuciones de estados mentales, enfatizando la necesidad de analizar la relevancia de uno u otro caso por caso. En definitiva, este enfoque ofrece una explicación post-ontológica de la relación entre mente y naturaleza, que evita tanto el problema mente-cuerpo como el problema de la normatividad. Para dar cuenta del carácter veritativo-evaluable y la fuerza normativa de las atribuciones de estados mentales, no necesitamos postular la existencia de extrañas criaturas ontológicas, sino simplemente reconocer la pluralidad de juegos del lenguaje a los que jugamos cuando intentamos dar cuenta de nuestro comportamiento y el de otras personas.

Esta visión pragmatista, antidescriptivista, evaluativista y regulativista de la mente proporciona, desde nuestra perspectiva, una arquitectura conceptual más sólida para la investigación y práctica en el ámbito de la salud mental. En la segunda parte de la tesis doctoral, la hemos aplicado a un debate particular en la filosofía de la salud mental, relativo a la conceptualización de los delirios como creencias y sus implicaciones para la investigación y la intervención con personas con delirios.

En el capítulo 5 introdujimos este debate. En primer lugar, vimos que el doxasticismo estándar que caracteriza las aproximaciones predominantes a la conceptualización de los delirios (el DSM, la neuropsiquiatría cognitiva, la terapia cognitivo-conductual, etc.) ha sido cuestionado por varias razones. El antidoxasticismo, basado fundamentalmente en teorías interpretacionistas y funcionalistas de la creencia, señala que los delirios no se ajustan a los perfiles racionales o causales estereotípicos de las creencias. Por el contrario, diversos autores y autoras han defendido el doxasticismo por medio de dos vías: la introducción de ciertas revisiones de los marcos interpretacionista y funcionalista (lo que caracteriza al doxasticismo revisionista), o el rechazo de dichas teorías en favor de conceptualizaciones alternativas de la creencia (lo que caracteriza al doxasticismo no revisionista). Quienes defienden el doxasticismo revisionista (Bayne y Pacherie, 2005; Bortolotti, 2010, 2012) establecen criterios más laxos para la determinación de aquello que puede contar como una creencia, y

asumen que las incoherencias mostradas por algunas personas con delirios pueden ser adecuadamente excusadas apelando a características especiales o "no estándar" de los casos analizados. Clutton (2018), por el contrario, rechaza el funcionalismo y el interpretacionismo por sus tendencias "antirrealistas". Propone, en cambio, adoptar una teoría cognitivo-fenomenológica de la creencia, según la cual creer que *p* equivale simplemente a tener la disposición a "asentir mentalmente" a *p* siempre que la persona valore la posibilidad de *p* en su mente. A pesar de sus diferencias, hemos destacado dos desiderátum comunes que motivan las defensas del doxasticismo: a) el desiderátum científico, según el cual el doxasticismo nos deja en una mejor posición para entender las causas de los delirios; y b) el desiderátum ético-político, según el cual el doxasticismo proporciona una manera de entender los delirios en términos de su inteligibilidad, erigiéndose así como como una barrera conceptual frente a prácticas indebidas de desagencialización y posibles formas de trato injusto que puedan derivarse de las mismas.

En el capítulo 6, hemos mostrado que ninguna de las defensas del doxasticismo presentadas son capaces de satisfacer ambos desiderátum. Los doxasticismos revisionistas, por un lado, no cumplen con el desiderátum científico. Su problema, en concreto, es que acaban viéndose forzados a comprometerse con alguna forma de relativismo respecto a la verdad de las atribuciones de creencia; de acuerdo con dicha postura, la verdad de una atribución de creencia podría depender de los estándares de quien la atribuye. Si así fuera, entonces la caracterización de los delirios en términos doxásticos no sería en sentido alguno informativo respecto a sus posibles causas. Por otra parte, el enfoque no revisionista de Clutton no satisface el desiderátum ético-político. La razón es que su teoría cognitivo-fenomenológica, independientemente de cómo se interprete, da lugar a una noción "normativamente inerte" de la creencia; una que no puede racionalizar el comportamiento ni, por tanto, informar juicio alguno sobre la agencia o autonomía de una persona. Después, hemos argumentado que el doxasticismo puede y debe ser defendido -no sobre la base de su atractivo científico, sino más bien de su atractivo ético-político-, y que nuestra aproximación antidescriptivista a la psicología del sentido común permite una defensa más robusta del mismo. En primer lugar, porque es capaz de captar la pluralidad de normas que entran en juego en las prácticas de atribución de creencias. En ocasiones, privilegiamos la autoridad de primera persona, considerada como una norma habitual en la comunicación interpersonal, por encima de consideraciones sobre la coherencia general del comportamiento de una persona (este, precisamente, parece ser el caso de los delirios). Otras veces, en cambio, razonamos a la inversa. En segundo lugar, hemos dicho que esto nos permite no sólo ver cómo funcionan de hecho nuestras prácticas de atribución de creencias, sino también por qué deberían seguir

funcionando de esta manera. La razón es que el antidoxasticismo, basado en una conceptualización idealista de nuestras prácticas de atribución de creencias, podría promover prácticas injustificadas de desagencialización de las personas con delirios. Por el contrario, desde una perspectiva antidescriptivista, el doxasticismo puede verse como una norma o política de conceptualización más sólida y deseable, a saber: por defecto, deberíamos asumir como ciertas las autoatribuciones de creencia de la persona. Por último, hemos visto cómo el antidescriptivismo, al distinguir los usos y objetivos principales de la psicología popular y la científica, protege al doxasticismo de tendencias eliminativistas. Desde este punto de vista, por muy lejos llegue la psicología científica, la conceptualización de los delirios como creencias seguirá teniendo sentido por derecho propio.

En el capítulo 7, nos hemos centrado en determinar si el tipo de doxasticismo que Clutton defiende -el doxasticismo científico, o la concepción doxástica de los delirios al uso en modelos cognitivistas tradicionales de los delirios- proporciona realmente un buen modelo científico de las experiencias delirantes. En primer lugar, hemos reseñado los dos principales enfoques cognitivistas tradicionales en la investigación sobre delirios: la terapia cognitivo-conductual para la psicosis (TCCp) y la neuropsiquiatría cognitiva. Partiendo de una conceptualización común de la mente y la cognición, estos dos enfoques complementarios explican los delirios como desviaciones de las mecánicas normales de procesamiento de la información. En concreto, los modelos cognitivos subyacentes enfatizan el papel de factores específicamente cognitivos, entre ellos la tríada cognitiva de Beck (es decir, las creencias basales de la persona sobre el mundo, el futuro y ella misma, con énfasis especial en estas últimas), y varios sesgos cognitivos, a destacar el salto a las conclusiones, ciertos sesgos atribucionales y déficits en la Teoría de la Mente. Posteriormente, hemos realizado una revisión narrativa de la evidencia disponible respecto a la eficacia de la TCCp. Hemos visto que, aunque hay evidencia de que la TCCp en general produce efectos entre pequeños y moderados, su eficacia en el caso de las personas con delirios es más ambigua. Es más, la evidencia disponible no avala las explicaciones cognitivistas de los delirios; sólo el salto a las conclusiones y los autoesquemas negativos aparecen significativamente en mayor medida en las personas con delirios, y no se ha encontrado que ninguno de los supuestos factores específicos medie la eficacia de las intervenciones cognitivo-conductuales con estas personas, ni siquiera cuando estas estaban específicamente dirigidas a alterar dichos factores. Más adelante, hemos señalado que estos resultados podrían explicarse parcialmente por la comprensión intelectualista de la mente que caracteriza al cognitivismo tradicional, es decir, la idea de que "tener una determinada creencia" o actuar de acuerdo con ella se reduce a contemplar ciertas proposiciones regulativas "frente a los ojos de la mente" y luego actuar en consecuencia.

En particular, hemos argumentado que, en el caso de los delirios, esto lleva a descuidar el efecto de los factores ambientales que puedan estar en juego en diferentes casos de delirios, así como las diversas técnicas de intervención destinadas a abordarlos.

Teniendo en cuenta todo esto, hemos defendido que los enfoques no cognitivistas, en concreto las aproximaciones analítico-funcionales, proporcionan un mejor modelo para la intervención con personas con delirios. En el capítulo 8, hemos realizado una revisión narrativa de estas aproximaciones. Hemos considerado por separado los dos principales enfoques analítico-funcionales: el análisis (aplicado) de la conducta "tradicional" y la terapia de aceptación y compromiso (ACT). El primero conceptualiza los delirios como conductas verbales no normativas y enfatiza la necesidad de realizar Evaluaciones Funcionales de la Conducta (EFC) antes del tratamiento. ACT, por el contrario, conceptualiza los delirios como reglas inflexibles mantenidas principalmente por su función de evitación, y enfatiza la necesidad de cambiar la relación de la persona con sus propias experiencias. La evidencia disponible respecto a las intervenciones con personas con delirios basadas en una EFC parece prometedora, pero hacen falta más síntesis cuantitativas. En el caso de ACT, cuando se mide su eficacia en sus propios términos, la evidencia disponible sugiere que sus beneficios se deben principalmente a su efecto sobre las alucinaciones, más que los delirios. A continuación, hemos visto que tanto la eficacia como la utilidad percibida de los enfoques analítico-funcionales podrían verse comprometidas debido a ciertos supuestos intelectualistas residuales. Por un lado, el análisis de la conducta tradicional exhibe este intelectualismo residual en su restrictiva operativización de los delirios como conductas verbales no normativas. Debido a ello, es objeto de la acusación de superficialidad planteada por el cognitivismo. Vista desde una perspectiva antidescriptivista, dicha acusación equivale a decir que son las atribuciones de estados mentales, y no los meros informes descriptivos de lo que dice una persona, los que fijan el listón de los cambios clínicamente significativos. Por otro lado, quienes trabajan bajo el marco de la ACT, profundizan su compromiso con el intelectualismo al tratar de reformular las creencias en términos de reglas verbales y procesos de comportamiento relacional. Esto puede derivar en el descuido de otras posibles fuentes de control ambiental, lo que explicaría en parte la ambigüedad de la evidencia disponible sobre la eficacia de las intervenciones ACT con personas con delirios. Finalmente, hemos discutido cómo nuestra aproximación antidescriptivista a la psicología del sentido común podría beneficiar a los enfoques analítico-funcionales. Al señalar la diferencia entre el comportamiento describible en términos de seguimiento de normas (por ejemplo, "actuar de acuerdo con nuestras propias creencias") y el comportamiento gobernado por reglas (es decir, el comportamiento controlado o mantenido por reglas verbales), el antidescriptivismo libera a los enfoques

analítico-funcionales de preconcepciones sobre las causas de la psicopatología que encorsetan la investigación, fomentando así un enfoque más amplio, basado en valores, para la determinación de los objetivos terapéuticos.

Como ya advertimos en la Introducción, esta tesis doctoral ha sido un largo viaje -más largo que el de Amarillo, a decir verdad. Nos hemos encontrado con múltiples debates, abordados desde diversas disciplinas y subdisciplinas, desde distintos ángulos y en diferentes lenguajes técnicos. Constituye así, si merece dicho nombre, una aportación a la filosofía de la salud mental. Si acaso, el valor de esta tesis doctoral reside, principalmente, en el tipo de puentes conceptuales que hemos tratado de establecer entre distintas disciplinas, así como en su reivindicación de la relevancia de la investigación filosófica para la investigación y la práctica clínica en el ámbito de la salud mental, cuya importancia se ve a menudo ensombrecida por el espectro del cientificismo.

Aparte de eso, pensamos que sus principales aportaciones residen en la aplicación del antidescriptivismo pragmatista que aquí hemos defendido al ámbito de la filosofía de la salud mental. En términos generales, el atractivo de este enfoque reside en su capacidad de acomodar lo mental y la normatividad en el seno de una cosmovisión naturalista. De este modo, contribuye a disolver una tensión que, como hemos visto, atraviesa diversos debates fundamentales en el ámbito de la salud mental desde al menos la segunda mitad del siglo XX. En el caso particular de los delirios, este enfoque ayuda a: a) explicar y defender de manera más robusta los beneficios ético-políticos del doxasticismo; y b) disipar compromisos conceptuales infundados que encorsetan la práctica clínica. En concreto, al señalar que las creencias no son entidades materiales con poderes causales, ni actuar conforme a una creencia el resultado de contemplar representaciones de ningún tipo -ni mentales ni verbales-, nuestra aproximación antidescriptivista fomenta la determinación caso por caso de los objetivos terapéuticos y el análisis caso por caso de la mejor manera de alcanzarlos. Al mismo tiempo, nos permite caracterizar de forma general los delirios como creencias, lo que evita prácticas de desagencialización indebida de las personas con delirios.

Sin embargo, como investigación en el ámbito de la filosofía de la salud mental, lo cierto es que no hemos dicho demasiado -explícitamente, al menos- acerca de los cuatro ejes principales de dicha disciplina, que identificamos en la Introducción (es decir, los problemas de la analogía, de los límites, de la prioridad y de la integración). Nos gustaría concluir este trabajo volviendo a estos cuatro temas. El objetivo será esbozar algunas reflexiones adicionales sobre estas cuestiones, de manera que puedan servir de base para el desarrollo futuro de una aproximación antidescriptivista a la filosofía de la salud mental.

9.2. Últimas notas sobre antidescriptivismo y filosofía de la salud mental

Nuestros ejemplos de partida (la locura loca de Púrpura y la locura karnatahclaniense de Amarillo) nos ayudaron a ilustrar los principales temas de la filosofía de la salud mental. Por un lado, estos ejemplos suscitan cuestiones sobre el papel de las *normas y los valores* -y los nichos sociales que los instituyen- en la determinación de lo que cuenta como patológico y lo que no. La locura loca de Púrpura es loca, precisamente, porque no parece haber nada malo o incorrecto sobre su estado de salud, a pesar de sus atípicos estados cerebrales. Era una locura loca porque, si alguien insistiese en que Púrpura tiene un problema de salud mental, nos sentiríamos inclinados a responder "Sí, vale, los estados cerebrales de Púrpura son *diferentes,* pero no hay nada intrínsecamente *malo* en ello. ¡Mírale! Está claro que se encuentra perfectamente de salud". En cambio, nuestra interpretación del caso de Amarillo parece variar en función de lo que el propio pueblo karnatahclaniense piense sobre el caso. Si consideraran el comportamiento y la experiencia de Amarillo como algo relativamente normal -una "segunda fase de la adolescencia", dijimos- no estaríamos tan inclinados a decir que tiene "un problema"; como mucho, diríamos que sus reacciones son diferentes a las nuestras (y un poco peculiares, a decir verdad). Pero, en nuestro caso, el pueblo karnatahclaniense evaluaba el comportamiento, la cognición y la experiencia de Amarillo como evaluaríamos en la Tierra los de Púrpura si actuase y reaccionase de igual manera. Siendo así, nos sentimos inclinados a decir que algo *anda mal* en el caso de Amarillo, que su sufrimiento constituye una llamada de ayuda y despierta nuestra empatía. Si alguien insistiera en que no podemos pronunciarnos sobre el carácter erróneo del sufrimiento de Amarillo, la mayoría nos sentiríamos inclinados a discrepar.

Por otra parte, nuestros ejemplos iniciales también nos invitan a pensar en la interrelación entre diferentes tipos de hechos, abordados desde distintas escalas de análisis, en el ámbito de la salud mental. ¿Dónde localizaríamos los problemas de salud mental de Púrpura o Amarillo, si es que tuviesen alguno? En ambos casos, sus estados y procesos cerebrales (o corporales) pueden ser relevantes para determinar posibles intervenciones -si se requiriese-, pero no parecen serlo a la hora de establecer si hay o no un problema en primer lugar. Más bien, son sus patrones generales de acción y reacción, de interacción con el entorno, lo que parece importar más para determinar su estado de salud mental. Por otro lado, el ejemplo de Amarillo pone de manifiesto el importante rol de las variables ambientales en el desarrollo de la psicopatología (por ejemplo, sus cuatro años de aislamiento social, el estrés relacionado con sus estudios de doctorado, etc.). En este sentido, estos ejemplos

suscitan preguntas sobre el papel constitutivo y causal que los diferentes tipos de factores (biológicos, psicológicos, sociales, etc.), pueden desempeñar en el análisis de los problemas de salud mental y su origen y mantenimiento.

Los cuatro grandes temas de la filosofía de la salud mental aparecen en estos casos. Se trata, como vimos, del problema de la analogía entre problemas de salud mental y física, el problema de los límites entre psicopatología y desviación social, el problema de la prioridad (es decir, si alguna escala de análisis debe gozar de prioridad causal o constitutiva sobre el resto) y el problema de la integración entre diversas escalas o niveles de análisis. Desde nuestro punto de vista, las dos primeras cuestiones están más estrechamente relacionadas con el papel de las normas y los valores en la conceptualización de los problemas de salud mental; las dos segundas, en cambio, están más estrechamente relacionadas con cuestiones de hecho relativas a la mejor manera de intervenir sobre los mismos. En lo que sigue, discutiremos por separado las posibles aportaciones de nuestro enfoque antidescriptivista a ambos tipos de cuestiones.

9.2.1. El antidescriptivismo y los problemas de la analogía y de los límites

En el capítulo 1, vimos que los problemas relativos a la analogía entre trastornos mentales y físicos, y a los límites entre psicopatología y desviación social se encuentran en el centro de las críticas clásicas contra el modelo médico. Sin embargo, antes de abordar estos problemas, nos gustaría decir algo sobre la conveniencia de la comprensión médica de los problemas psicológicos. En esta tesis doctoral, nos hemos opuesto claramente a la interpretación biomédica, o fuerte, del modelo médico; pensamos que entender los problemas de salud mental como originados principalmente en maquinarias internas defectuosas sólo desvía la atención de las obvias raíces contextuales del malestar psicológico. Sin embargo, somos agnósticos respecto a la conveniencia de enmarcar los problemas psicológicos en términos médicos, así como respecto al uso de etiquetas diagnósticas desde una perspectiva minimalista. Sin duda, en la medida en que ello invite a un pensamiento individualista e internalista sobre el origen de los problemas de salud mental, o en la medida en que patologice meras formas de desviación social (o incluso meras experiencias comunes en el curso de una vida), la comprensión médica de los problemas psicológicos debe ser rechazada. Sin embargo, creemos que es posible aquí separar las dos cuestiones. Enmarcar los problemas psicológicos en términos médicos y disponer de ciertas etiquetas diagnósticas con las que identificar las diferentes formas de sufrimiento psicológico sirve a importantes propósitos prácticos, que van mucho más allá de la supuesta facilitación de la "comunicación interprofesional", tan a menudo invocada en defensa de las taxonomías tradicionales. Desde el punto de vista administrativo, ayuda a las personas que sufren problemas psicológicos a acceder a recursos

de la seguridad social (prestaciones por discapacidad, bajas laborales, etc.), que en muchos casos proveen las condiciones materiales necesarias para poder iniciar un proceso de recuperación. Las etiquetas diagnósticas también desempeñan una importante función como herramientas hermenéuticas, sirviendo para que las personas puedan dar nombre y articular en cierto marco de inteligibilidad lo que les ocurre; ello, a su vez, puede ayudarles a adquirir una cierta perspectiva distanciada y menos autoculpabilizadora respecto a su sufrimiento. Por último, los movimientos a favor de la diversidad en el ámbito de la salud mental (es decir, los movimientos de la neurodiversidad o la psicodiversidad) están progresivamente reapropiándose, repolitizando y despatologizando cada vez más estas etiquetas diagnósticas en un esfuerzo por promover el reconocimiento y la revalorización de ciertas formas de diversidad psicológica (por ejemplo, véase Chapman 2020; Singer, 1999).

Mientras las etiquetas diagnósticas desempeñen estas funciones, puede que merezca la pena mantenerlas -mejor, probablemente, en el marco del proceso de despatologización y repolitización que acabamos de mencionar. Sin embargo, lo que nos interesa aquí son las siguientes preguntas: independientemente de si vale la pena mantener los términos "problemas de salud mental" o "trastornos mentales", ¿son los problemas de salud mental *análogos* a los problemas de salud física? Y, ¿cómo hemos de precisar la diferencia entre desviación social y psicopatología?

Empecemos por el problema de la analogía. Comúnmente, las respuestas a este problema recurren a una u otra noción de "enfermedad" o "trastorno" en general. Como han afirmado algunos autores (véase Fulford & van Staden, 2013; Thornton, 2007), las defensas y críticas tradicionales de la analogía (por ejemplo, Kendall, 1975; Szasz 1961/1974) se han basado típicamente en una definición estrecha de "trastorno"; a saber, una que asume que la noción de trastorno puede ser explicada en términos puramente descriptivos o no evaluativos. Estas disputas, por tanto, se dan en torno a la conveniencia de una u otra noción descriptiva de "patología"; por ejemplo, entre la comprensión virchowiana de Szasz y la redefinición de Boorse (2014) de la misma en términos de "un estado de disfunción biológica parcial estadísticamente por debajo de lo normal para la especie, en relación con la cohorte correspondiente de sexo y edad" (p. 684; traducción del autor). De igual modo, las disputas giran en torno a si estos criterios permiten incluir o no los problemas de salud mental. Enfoques más recientes han llamado la atención sobre el hecho de que ni los problemas de salud mental ni los somáticos pueden explicarse en términos puramente descriptivos y libres de valores (por ejemplo, Fulford y van Staden, 2013; Graham, 2010b; Thornton, 2007; Varga, 2015). Por el contrario, estas aproximaciones intermedias sostienen que la consideración de *cualquier* tipo de condición como patológica implica necesariamente juicios de valor. Por tanto,

concluyen que no hay ninguna diferencia esencial entre los problemas de salud mental y los físicos; la única diferencia entre ellos radicaría en el grado de *acuerdo* o *desacuerdo* entre diferentes partes interesadas dentro de una determinada comunidad respecto a los valores y normas en juego en diferentes casos, lo que determinaría si tiene sentido o no *evaluar* cierta condición como patológica. Se asume entonces que, mientras que en el caso de los problemas de salud física hay un grado mínimo de variabilidad respecto a los valores de las distintas partes, en el caso de los problemas de salud mental hay un grado máximo de disonancia en cuanto a qué valores (y los de quién) deben determinar la categorización de alguna condición como un tipo de patología.

No vamos a entrar en detalles aquí sobre estas aproximaciones intermedias. Sólo queremos señalar que estos enfoques, o al menos algunos de ellos, son atractivos desde un punto de vista anti-descriptivista; de hecho, algunos de ellos se basan en un enfoque anti-descriptivista similar respecto a los juicios de valor o se apoyan en argumentos wittgensteinianos y ryleanos similares (por ejemplo, Fulford y van Staden, 2013; Thornton, 2007). Sin embargo, creemos que estos enfoques se quedan cortos en sus análisis antidescriptivistas. En concreto, aunque asumen un enfoque antidescriptivista respecto a la noción de "trastorno" incluida en la noción de "trastorno mental", estos enfoques olvidan aplicar dicho análisis al aspecto de lo *mental*. Desde nuestro punto de vista, por tanto, estos enfoques están en lo cierto al señalar que el diagnóstico de los problemas de salud, tanto físicos como mentales, descansa necesariamente sobre un lecho de valores; sin embargo, eso no significa necesariamente que unos y otros sean análogos.

Desde nuestra perspectiva, estos enfoques asumen que, en ambos casos, los diagnósticos tienen una doble función: describir algún estado material y evaluarlo como patológico. Nuestro análisis antidescriptivista va más allá: aunque esto podría captar cómo funciona el diagnóstico de los problemas de salud físicos, no termina de ajustarse a las prácticas de evaluación en salud mental. Lo "físico" en "trastorno físico" puede explicarse en términos puramente descriptivos; lo "mental" en "salud mental", no. Por supuesto, en salud mental hablamos de los patrones de acción y reacción de una persona; sin embargo, lo que los califica como "signos" de problemas de salud *mental* (frente a problemas de salud física) es que estos patrones de comportamiento se evalúan en términos principalmente personales (por tanto, evaluativos), en lugar de subpersonales. Los problemas de salud mental afectan principalmente a la condición de la persona como *agente*, no como mero organismo. Por lo tanto, están constituidos por alteraciones en la capacidad de la persona para encontrar significado y valor en su relación consigo misma y con su mundo social; o, por decirlo en términos de de Haan (2020), en sus capacidades de "creación de significado existencial" (p. 11, traducción

del autor). Es por ello que los problemas de salud mental "se disuelven si una persona logra cambiar su forma de interactuar con el mundo", mientras que "los efectos secundarios de los trastornos físicos en la creación de significado, por el contrario, no desaparecen al interactuar con el mundo de una manera diferente"; en otras palabras: los problemas de salud mental, *en tanto que* estados mentales, "no son del cerebro, ni siquiera del cuerpo, sino de *las personas*" (p. 11, traducción del autor, énfasis añadido).

De nuevo, esto no significa que entender los problemas psicológicos en términos médicos no pueda tener consecuencias beneficiosas, ni que los servicios y recursos en salud mental sean una mera forma de "control estatal" de la sacrosanta individualidad del individuo, ni que la atribución del "rol de enfermo" a alguien a causa de su sufrimiento psicológico no pueda tener sentido en ocasiones (por ejemplo, para darle tiempo y espacio para descansar y poder iniciar un proceso de recuperación). En este sentido, no creemos que la comprensión médica de los problemas psicológicos sea una metáfora necesariamente perversa. Pero sí creemos que, de hecho, es una metáfora -o, al menos, que los problemas de salud física y mental no son análogos entre sí. La disanalogía entre los problemas de salud mental y física queda patente en el hecho de que hablar (en un sentido literal) de "libre albedríos patológicos", "creencias enfermas" o "afecciones de los deseos" es de todo punto extravagante, al igual que lo es hablar literalmente de la "moral enferma" o los "valores patológicos" de un individuo. Las mentes, las normas y los valores pueden estar implicados en la determinación de lo que cuenta como "enfermo" o "patológico"; sin embargo, no son *en sí mismos* el tipo de cosas de las que tiene sentido decir, en términos literales, que están "enfermas" o que "sufren una afección"; pensar lo contrario equivale al tipo de error categorial señalado por Ryle (1949). Nuestra aproximación antidescriptivista a la psicología del sentido común nos ayuda a entender por qué hablar de "trastornos mentales" no debería entenderse en términos literales (es decir, en estricta analogía con los trastornos físicos); al mismo tiempo, se resiste a la tendencia a calificar estas expresiones de "mitos" o a despreciar las atribuciones de estados mentales (incluidas las evaluaciones de salud mental) como carentes de valor de verdad.

Esto también tiene consecuencias para el problema de los límites, o la posibilidad de distinguir entre "psicopatología" y "desviación social". Una vez más, nuestra discusión sobre la disanalogía entre los problemas de salud mental y salud física revela que la evaluación de la salud mental es una tarea normativa o evaluativa de principio a fin; desde la determinación de lo que cuenta como estar en ciertos estados mentales, hasta la determinación de cuáles de ellos son "erróneos" o "andan mal" de alguna manera. Las normas y los valores sociales están profundamente arraigados en ambos tipos de juicios; por lo tanto, distinguir la

"psicopatología" de la "desviación social" no es, como mínimo, una tarea sencilla. Desde nuestro punto de vista, qué normas y valores (y de quién) deben contar para determinar qué cae del lado de la psicopatología (y, por lo tanto, merece la atribución del rol de enfermo, con sus beneficios y desventajas correspondientes), y qué cae del lado de la mera desviación social (y, por lo tanto, debe respetarse si no perjudica a nadie, como la orientación sexual o la identidad de género) es algo que probablemente variará entre tipos de problemas de salud mental y en función del caso particular. En este sentido, nuestro enfoque antidescriptivista se alinea con el tipo de particularismo normativo que defiende Thornton (2007); no hay máximas ni principios generales que nos permitan determinar, para todos y cada uno de los casos posibles, sobre qué se debe intervenir.

Sin embargo, si tuviéramos que hablar en términos generales, creemos que una buena regla general viene dada por la consideración de cómo el comportamiento y las experiencias de una persona se alinean o no con *sus propios* valores. Esto va en la línea de los enfoques analítico-funcionales como ACT y las recientes aproximaciones enactivistas (por ejemplo, véase de Haan, 2021; Nielsen, 2021), que hacen hincapié en la naturaleza *autoboicoteadora* de los problemas de salud mental; esto es, el hecho de que al menos muchos de ellos pueden caracterizarse principalmente en términos de desviaciones sistemáticas de las propias formas de vida que una persona aspira a tener y que provocan malestar emocional. Desde este punto de vista, el "sello distintivo de la psicopatología" (frente a la mera desviación social) es, o debería ser, la presencia de sufrimiento psicológico debido a un conflicto entre las acciones de una persona y los valores que la misma hace suyos.

Creemos que nuestro enfoque antidescriptivista arroja ideas interesantes sobre esta cuestión, que merecerían ser desarrolladas en futuras investigaciones. En pocas palabras, su principal potencial reside en su capacidad para proporcionar una visión antirreduccionista y antiindividualista de los valores y los conflictos de valores de una persona. Un ejemplo de esto puede verse en contraste con las aproximaciones "cosificadoras" y subjetivistas de los valores que se desprenden de su estricta identificación con lo que la persona dice que son sus valores, o con las "reglas verbales" que la persona emite (por ejemplo, ACT), así como otros tipos de enfoques individualistas que parecen asumir que los valores de una persona son algo elegible a capricho por la misma (por ejemplo, el enfoque de Szasz). En resumen, desde nuestra perspectiva, la determinación de cuáles son los valores de una persona (es decir, sus *creencias* evaluativas) no puede reducirse a una mera descripción de lo que la persona dice sobre los mismos; más bien, la clarificación de valores requiere ver el comportamiento de una persona en relación con las múltiples prácticas e instituciones *sociales* que configuran el nicho sociocultural de la persona y fundamentan las normas que sigue. La

clarificación de valores es, por tanto, una tarea irreductiblemente hermenéutica y contextualizada, que requiere participar en prácticas interactivas de evaluación y regulación para diferenciar las normas que realmente seguimos, las que nos gustaría seguir y las que otras personas quieren que sigamos. Nuestro enfoque antidescriptivista da sentido y veracidad a estas prácticas interpretativas sin reducirlas a la formulación de hipótesis empíricas sobre las normas verbales que emite una persona (ni sobre procesos internos ocultos que supuestamente median su relación con el mundo). Además, proporciona una comprensión antiindividualista de lo que suponen los "valores propios" y, por tanto, del "criterio del autoboicot" que nos permitiría, como hemos dicho, distinguir psicopatología y mera desviación social. Desde este punto de vista, lo que diferencia los problemas de salud mental de las diversas formas de desviación social es que los primeros, a diferencia de las segundas, implican un tipo particular de conflicto socio-normativo: el que surge no entre las acciones de una persona y *cualquier* valor o expectativa social, sino entre sus acciones y las normas y valores sociales que ella misma defiende explícitamente, o que *expresa* implícitamente en su acción cuando se encuentra atrapada en los bucles característicos del malestar psicológico. Esta definición provisional antiintelectualista podría ser, desde nuestra perspectiva, útil para seguir investigando el problema de los límites en filosofía de la salud mental.

9.2.2. El antidescriptivismo y los problemas de la prioridad y la integración

Volvamos ahora a los problemas de la prioridad y de la integración. ¿Debemos privilegiar alguna escala de análisis (por ejemplo, la biológica, la psicológica, la social, etc.) en la conceptualización de los problemas de salud mental o en la intervención sobre sus causas? Por otro lado, ¿cómo debemos abordar la integración de los diversos proyectos explicativos en la investigación en salud mental y en la práctica clínica?

En primer lugar, a nuestro entender, huelga decir que un marco general adecuado para la provisión de servicios en el ámbito de la salud mental debería contar con profesionales de todo tipo trabajando desde distintos niveles o escalas de análisis. En este sentido, pensamos que la investigación científica sobre las *causas* de los problemas de salud mental debe enfocarse en el marco de un esfuerzo multidisciplinar dirigido a establecer los diferentes tipos de dinámicas causales implicadas en su desarrollo y mantenimiento. Nos mantenemos agnósticos, por otro lado, respecto a la posibilidad de analizar adecuadamente todas estas dinámicas dentro de un marco *ontológico* integrador; después de todo, es posible que diferentes lenguajes y métodos científicos resulten ser inconmensurables, y eso es algo que la apelación a propiedades "emergentes" probablemente no pueda resolver. En cualquier caso, pensamos que es recomendable adoptar un pluralismo epistemológico sano -siempre desde un marco naturalista- respecto a los diferentes enfoques causal-explicativos posibles

en el ámbito de la salud mental, y que los esfuerzos integradores, en la medida en que arrojen nuevas líneas productivas de investigación o maximicen nuestras capacidades de intervención sobre el sufrimiento psicológico, deben ser bienvenidos.

En la misma línea, la investigación sobre los procesos corporales involucrados en los problemas de salud mental (entre ellos, los cerebrales) es evidentemente necesaria para comprender cómo el sufrimiento psicológico afecta al organismo; entre otras razones, porque esto nos permite mostrar los profundos efectos que tienen sobre las personas las dinámicas sociales estresantes o injustas a las que a menudo se encuentran expuestas. Nuestra postura sobre la psicofarmacología y el uso de psicofármacos para aliviar los problemas de salud mental es similar; en la medida en que estos se utilicen y prescriban de forma *responsable,* y en la medida en que las personas que los toman estén debidamente informadas tanto de sus posibles beneficios como de sus posibles efectos secundarios, los psicofármacos pueden ser útiles en muchas situaciones -por ejemplo, proporcionando las condiciones necesarias (por ejemplo, descanso, reducción de la agitación, etc.) para que una persona pueda iniciar un proceso de recuperación.

Dicho esto, sin embargo, otra cuestión es si debemos dar prioridad *conceptual* a una u otra escala de análisis sobre el resto y, por tanto, si debemos adoptarla como nuestra escala de análisis focal; esto es, aquella que establece nuestro principal objeto de estudio. En este punto, la postura defendida en esta tesis doctoral es claramente afín a los distintos modelos psicológicos -especialmente, a aquellos que toman como punto de partida la interacción entre la persona y su entorno. Coincidimos, por tanto, con los enfoques analítico-funcionales -así como con las recientes propuestas enactivistas- en que la unidad básica de análisis en la investigación y la práctica de la salud mental es, y debe seguir siendo, la *conducta de la persona* (entendida en sentido amplio) y no sus circuitos cerebrales.

Nuestro marco antidescriptivista encaja bien con estos supuestos. Respecto a la tensión entre los modelos psicológicos cognitivistas y los no cognitivistas, la aproximación antidescriptivista nos permite señalar que lo psicológico debe entenderse en términos irreduciblemente contextuales. Los estados mentales no son hechos internos que median entre la percepción y la acción; del mismo modo, los problemas de salud mental no se encuentran en supuestos déficits o alteraciones en hipotéticos mecanismos internos de procesamiento de la información. Desde nuestra perspectiva antidescriptivista, el ámbito de la salud mental tiene que ver principalmente con la interacción entre una persona y su entorno natural y social.

Del mismo modo, este enfoque nos ayuda a ver qué es exactamente lo que está mal en las tendencias "neuro-reduccionistas" (o "neuro-eliminativistas") en salud mental, que

establecen los "circuitos cerebrales" de una persona como la unidad de análisis focal en la investigación en salud mental y en la práctica clínica. Como hemos visto, lo "mental" en "salud mental" no puede reducirse a una mera descripción de los estados cerebrales de una persona ni a una mera descripción de sus patrones de comportamiento; de hecho, no puede reducirse a meras descripciones de hechos de ningún tipo. Sin embargo, de nuestro análisis también se desprende que no todos los tipos de hechos desempeñan el mismo papel en la determinación de lo que cuenta como tener ciertos estados mentales (por ejemplo, las creencias que una persona tiene sobre sí misma o el mundo). Cuando ejercemos el tipo de práctica interpretativa característica de la psicología del sentido común, son las *acciones y reacciones* propias y de otras personas (motoras, simbólicas, inferenciales, fenomenológicas, etc.) las que estamos sometiendo a evaluación para decidir si "merecen" ciertas atribuciones de estados mentales, o si son inteligibles cuando son vistas a la luz de las mismas. Nuestros cerebros -nuestros cuerpos, en general- nos permiten actuar y reaccionar, pero no son en sí mismos el objeto de evaluación cuando valoramos los estados mentales de cada cual. En la medida en que ciertos procesos corporales se encuentran alterados en ciertos problemas de salud mental, incluidos los cerebrales, el nivel de los "circuitos cerebrales" es obviamente relevante para la investigación en salud mental; sin embargo, elevarlo a la categoría de "unidad focal de análisis" constituye, una vez más, una muestra del error categorial que caracteriza el dogma del Fantasma en la Máquina. En resumen, los circuitos cerebrales tienen un rol *posibilitador* en los problemas de salud mental; los comportamientos y experiencias de la persona, evaluados desde marcos evaluativos particulares, *constituyen* el objeto mismo de la investigación en salud mental y en la práctica clínica. La visión futurista de los servicios de salud mental que describimos al inicio de esta tesis doctoral sigue atrayendo a muchos y muchas; por el contrario, creemos que nuestra aproximación antidescriptivista proporciona razones sólidas para entender por qué esta es una visión fundamentalmente errónea, basada en suposiciones equivocadas sobre la naturaleza de la mente y sus diversas tribulaciones.

A modo de conclusión, esperamos haber dado argumentos convincentes en favor del tipo de antidescriptivismo pragmatista, evaluativista y regulativista que constituye nuestro enfoque, así como de la necesidad de desplazar el foco de análisis de los rompecabezas ontológicos del cartesianismo al análisis de nuestras prácticas lingüísticas en el ámbito de la salud mental. Es hora de superar la tendencia a abordar las discusiones sobre el estatus de lo "mental" en "salud mental" desde el marco del "a ver quién es más dualista que quién" (Pinedo-García, 2020, traducción del autor). Desde nuestro punto de vista, esta tesis doctoral es un paso en esa dirección.

References

Acero, J. J., (Ed.) (2019). *Guía Comares de Wittgenstein*. Comares.

Acero, J. J., & Villanueva, N. (2012). Wittgenstein y la intencionalidad de lo mental. *Análisis filosófico*, *32*(2), 117-154.

Aftab, A. (2021). Conceptual psychiatry: the ground beneath our feet. *International Review of Psychiatry, 33*(5), 443-445. https://doi.org/10.1080/09540261.2020.1864303

Aftab, A., & Nielsen, K. (2021). From Engel to Enactivism: Contextualizing the Biopsychosocial Model. *European Journal of Analytic Philosophy*, *17*(2), M2-22. https://doi.org/10.31820/ejap.17.2.3

Aftab, A., & Rashed, M. A. (2021). Mental disorder and social deviance. *International Review of Psychiatry*, *33*(5), 478-485. https://doi.org/10.1080/09540261.2020.1815666

Alexander, S. (1966). *Space, Time, and Deity: The Gifford Lectures at Glasgow 1916–1918*. Palgrave Macmillan. (Original work published 1920).

Alford, B. A. (1986). Behavioral treatment of schizophrenic delusions: A single-case experimental analysis. *Behavior Therapy*, *17*(5), 637-644. https://doi.org/10.1016/S0005-7894(86)80101-0

Alford, B. A., & Beck, A. T. (1994). Cognitive therapy of delusional beliefs. *Behaviour research and therapy*, *32*(3), 369–380. https://doi.org/10.1016/0005-7967(94)90134-1

Almagro-Holgado, M. (2020). Límites de la noción de 'affordance' y de la concepción de lo mental en el marco de la psicología ecológica. *Teorema: Revista internacional de filosofía*, *39*(1), 135-149.

Almagro-Holgado, M. (2021). Seeing hate from afar: the concept of affective polarization reassessed. (Doctoral dissertation, Universidad de Granada). http://hdl.handle.net/10481/70432

Almagro-Holgado, M. & Fernández-Castro, V. (2019). The social cover view: a non-epistemic approach to mindreading. *Philosophia*, 48, 483–505. https://doi.org/10.1007/s11406-019-00096-2

Almagro-Holgado, M., & Moreno-Zurita, A. (2022). Affective Polarization and Testimonial and Discursive Injustice. In D. Bordonaba, V. Fernández-Castro, & J. R. Torices (Eds.). (2022). *The Political Turn in Analytic Philosophy: Reflections on Social Injustice and Oppression* (pp. 259-278). Walter de Gruyter GmbH & Co KG.

Almagro-Holgado, M. A., Navarro-Laespada, L., & de Pinedo-García, M. (2021). Is testimonial injustice epistemic? Let me count the ways. *Hypatia*, *36*(4), 657-675. https://doi.org/10.1017/hyp.2021.56

Alonso-Vega, J. (2021). Análisis funcional de la interacción verbal entre terapeuta y cliente con diagnóstico de Trastorno Mental Grave. (Doctoral dissertation, Universidad Autónoma de Madrid). http://hdl.handle.net/10486/700789

Alonso-Vega, J., Ávila-Herrero, I., Núñez de Prado-Gordillo, M., & Pereira Xavier, G. (2020). Análisis de la conducta y prácticas culturales. In M. X. Froxán Parga (Coord.) *Análisis funcional de la conducta humana: Concepto, metodología y aplicaciones* (pp. 179-201). Pirámide.

Alonso-Vega, J., Núñez de Prado-Gordillo, M., Pereira, G. L., & Froxán-Parga, M. X. (2019). El tratamiento de Enfermedades Mentales Graves desde la investigación de procesos. *Conductual*, *7*, 44-65. https://www.conductual.com/articulos/El%20tratamiento%20de%20enfermedades%20mentales%20graves%20desde%20la%20investigacion%20de%20procesos.pdf

Alumbaugh, R. V. (1971). Use of behavior modification techniques toward reduction of hallucinatory behavior: A case study. *The Psychological Record, 21*(3), 415–417. https://doi.org/10.1007/BF03394034

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders (*DSM-II*)*. American Psychiatric Association.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (*DSM-III*)*. American Psychiatric Association.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders (*DSM-III-TR*)*. American Psychiatric Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders (*DSM-IV*)*. American Psychiatric Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (*DSM-IV-TR*)*. American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Anderson, L. T., & Alpert, M. (1974). Operant analysis of hallucination frequency in a hospitalized schizophrenic. *Journal of Behavior Therapy and Experimental Psychiatry, 5*(1), 13–18. https://doi.org/10.1016/0005-7916(74)90007-X

Andreasen, N. C. (1997). Linking mind and brain in the study of mental illnesses: a project for a scientific psychopathology. *Science*, *275*(5306), 1586-1593. https://doi.org/10.1126/science.275.5306.1586

Andreasen, N. C. (1997). Linking mind and brain in the study of mental illnesses: A project for a scientific psychopathology. *Science, 275*(5306), 1586–1593. https://doi.org/10.1126/science.275.5306.1586

Andreasen, N. C. (2001). *Brave New Brain: Conquering Mental Illness in the Era of the Genome*. Oxford University Press.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *The American Psychologist, 61*(4), 271-285.

APA Task Force on Promotion and Dissemination of Psychological Procedures. (1995). Training in and dissemination of empirically-validated psychological treatments. *The Clinical Psychologist, 48*(1). 3-23.

Arntzen, E., Tonnessen, I. R., & Brouwer, G. (2006). Reducing aberrant verbal behavior by building a repertoire of rational verbal behavior. *Behavioral Interventions*, *21*(3), 177–193. https://doi.org/10.1002/bin.220

Assaz, D. A., Roche, B., Kanter, J. W., & Oshiro, C. K. B. (2018). Cognitive defusion in acceptance and commitment therapy: What are the basic processes of change? *The Psychological Record, 68*(4), 405–418. https://doi.org/10.1007/s40732-017-0254-z

Austin, J. L. (1961). Other minds. In J. L. Austin, J. O. Urmson, & G. J. Warnock (Eds.), *Philosophical papers* (pp. 44-84). Oxford University Press. (Original work published 1946).

Austin, J. L. (1961). Performative Utterances. In J. L. Austin, J. O. Urmson, & G. J. Warnock (Eds.), *Philosophical papers* (pp. 220-240). Oxford University Press. (Original work published 1946).

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Avramides, A. (2020). Other minds. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/other-minds/

Ayer, A. J. (1936). *Language, truth, and logic*. Victor Gollancz.

Ayllon, T, & Haughton, E. (1964). Modification of symptomatic verbal behaviour of mental patients. *Behaviour Research and Therapy*, 87–97. https://doi.org/10.1016/0005-7967(64)90001-4

Ayllon, T., Haughton, E., & Hughes, H. B. (1965). Interpretation of symptoms: Fact or fiction? *Behaviour Research and Therapy*, *3*(1), 1-7. https://doi.org/10.1016/0005-7967(65)90037-9

Ayllon, T., & Michael, J. (1959). The psychiatric nurse as a behavioral engineer. *Journal of the Experimental Analysis of Behavior*, *2*(4), 323–334. https://doi.org/10.1901/jeab.1959.2-323

Baars, B. J. (2003). The double life of B.F. Skinner: Inner conflict, dissociation and the scientific taboo against consciousness. *Journal of Consciousness Studies, 10*(1), 5–25.

Bach, P., Gaudiano, B. A., Hayes, S. C., & Herbert, J. D. (2013). Acceptance and commitment therapy for psychosis: intent to treat, hospitalization outcome and mediation by believability. *Psychosis*, *5*(2), 166-174. https://doi.org/10.1080/17522439.2012.671349

Bach, P. A., Gaudiano, B., Pankey, J., Herbert, J. D., & Hayes, S. C. (2006). Acceptance, Mindfulness, Values, and Psychosis: Applying Acceptance and Commitment Therapy (ACT) to the Chronically Mentally Ill. In R. A. Baer (Ed.), *Mindfulness-based treatment approaches: Clinician's guide to evidence base and applications* (pp. 93–116). Elsevier Academic Press. https://doi.org/10.1016/B978-012088519-0/50006-X

Bach, P., & Hayes, S. C. (2002). The use of acceptance and commitment therapy to prevent the rehospitalization of psychotic patients: a randomized controlled trial. *Journal of consulting and clinical psychology*, *70*(5), 1129–1139. https://doi.org/10.1037//0022-006X.70.5.1129

Bach, P., Hayes, S. C., & Gallop, R. (2012). Long-term effects of brief acceptance and commitment therapy for psychosis. *Behavior modification*, *36*(2), 165–181. https://doi.org/10.1177/0145445511427193

Baum W. M. (2011). What is Radical Behaviorism? A Review of Jay Moore's *Conceptual Foundations of Radical Behaviorism*. *Journal of the Experimental Analysis of Behavior*, *95*(1), 119–126. https://doi.org/10.1901/jeab.2011.95-119

Bandura, A. (1969). *Principles of behavior modification.* Holt, Rinehart, & Winston.

Bandura, A. (1974). Behavior theory and the models of man. *American Psychologist, 29*(12), 859–869. https://doi.org/10.1037/h0037514

Banner, N. F., & Thornton, T. (2007). The new philosophy of psychiatry: its (recent) past, present and future: a review of the Oxford University Press series International Perspectives in Philosophy and Psychiatry. *Philosophy, Ethics, and Humanities in Medicine*, *2*(1), 1-14. https://doi.org/10.1186/1747-5341-2-9

Bar-On, D. (2015). Transparency, expression, and self-knowledge. *Philosophical Explorations*, *18*(2), 134-152.

Bar-On, D., & Sias, J. (2013). Varieties of expressivism. *Philosophy Compass*, *8*(8), 699-713. https://doi.org/10.1111/phc3.12051

Barnes-Holmes, D. (2000). Behavioral pragmatism: No place for reality and truth. *The Behavior Analyst*, *23*(2), 191-202. https://doi.org/10.1007/BF03392010

Barnes-Holmes, D., Barnes-Holmes, Y., Smeets, P. M., Cullinan, V., & Leader, G. (2004). Relational frame theory and stimulus equivalence: Conceptual and procedural issues. *International Journal of Psychology and Psychological Therapy, 4*, 181-214.

Barnes-Holmes, Y., Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). Relational frame theory: a post-Skinnerian account of human language and cognition. *Advances in child development and behavior*, *28*, 101–138. https://doi.org/10.1016/s0065-2407(02)80063-5

Barnes-Holmes, Y., Hussey, I., McEnteggart, C., Barnes-Holmes, D., & Foody, M. (2016). Scientific Ambition: The Relationship between Relational Frame Theory and Middle-Level Terms in Acceptance and Commitment Therapy. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley handbook of contextual behavioral science* (pp. 365-382). John Wiley & Sons.

Barrett L. (2015). Why Brains Are Not Computers, Why Behaviorism Is Not Satanism, and Why Dolphins Are Not Aquatic Apes. *The Behavior analyst*, *39*(1), 9–23. https://doi.org/10.1007/s40614-015-0047-0

Barrett, L. (2019). Enactivism, pragmatism... behaviorism? *Philosophical Studies*, *176*(3), 807-818. https://doi.org/10.1007/s11098-018-01231-7

Battie, W. (1758). *A Treatise on Madness*. J. Whiston & B. White.

Baum, W. M. (2011). What is radical behaviorism? A review of Jay Moore's Conceptual Foundations of Radical Behaviorism. *Journal of the experimental analysis of behavior*, *95*(1), 119-126. https://doi.org/10.1901/jeab.2011.95-119

Baum, W. M. (2017). Ontology for behavior analysis: Not realism, classes, or objects, but individuals and processes. *Behavior and Philosophy*, *45*, 63-80.

Bayne, T. (2010). Delusions as doxastic states: Contexts, compartments, and commitments. *Philosophy, Psychiatry, & Psychology*, *17*(4), 329-336.

Bayne, T., & Hattiangadi, A. (2013). Belief and its bedfellows. In N. Nottelman (Ed.) *New essays on belief* (pp. 124-144). Palgrave Macmillan.

Bayne, T., & Pacherie, E. (2004a). Bottom-Up or Top-Down? Campbell's Rationalist Account of Monothematic Delusions. *Philosophy, Psychiatry, & Psychology, 11*(1), 1–11. https://doi.org/10.1353/ppp.2004.0033

Bayne, T., & Pacherie, E. (2004b). Experience, belief, and the interpretive fold. *Philosophy, Psychiatry, & Psychology*, *11*(1), 81–86.

Bayne, T., & Pacherie, E. (2005). In defence of the doxastic conception of delusions. *Mind & Language*, *20*(2), 163–188. https://doi.org/10.1111/j.0268-1064.2005.00281.x

Beavers, G. A., Iwata, B. A., & Lerman, D. C. (2013). Thirty years of research on the functional analysis of problem behavior. *Journal of applied behavior analysis*, *46*(1), 1-21.

Beck, A. T. (1952). Successful outpatient psychotherapy of a chronic schizophrenic with a delusion based on borrowed guilt. *Psychiatry: Journal for the Study of Interpersonal Processes, 15,* 305–312. https://doi.org/10.1080/00332747.1952.11022883

Beck, A. T. (1961). A systematic investigation of depression. *Comprehensive Psychiatry, 2*(3), 163–170. https://doi.org/10.1016/S0010-440X(61)80020-5

Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, *9*(4), 324-333. https://doi:10.1001/archpsyc.1963.01720160014002

Beck, A. T. (1964). Thinking and depression: II. Theory and therapy. *Archives of general psychiatry*, *10*(6), 561-571. https://doi:10.1001/archpsyc.1964.01720240015003

Beck, A. T. (1970). Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy, 1*(2), 184–200. https://doi.org/10.1016/S0005-7894(70)80030-2

Beck, A. T. (Ed.). (1979). *Cognitive therapy of depression*. Guilford press.

Beck, A. T., & Alford, B. A. (2009). *Depression: Causes and treatment*. University of Pennsylvania Press. (Original work published 1967).

Bedau, M. A. (1997). Weak emergence. *Philosophical perspectives*, *11*, 375-399. https://doi.org/10.1111/0029-4624.31.s11.17

Bedau, M. A., & Humphreys, P. E. (2008). *Emergence: Contemporary readings in philosophy and science*. MIT press.

Bell, V., Halligan, P. W., & Ellis, H. D. (2006). Explaining delusions: A cognitive perspective. *Trends in Cognitive Sciences, 10*(5), 219–226. https://doi.org/10.1016/j.tics.2006.03.004

Bensusan, H., & Pinedo-García, M. (2007). Minimal empiricism without dogmas. *Philosophia*, *35*(2), 197-206.

Bentall, R. P. (2003). *Madness explained: Psychosis and human nature*. Penguin UK.

Bernstein, R. J. (2010). *The pragmatic turn*. Polity Press.

Berrios, G. E. (1991). Delusions as "wrong beliefs": A conceptual history. *The British Journal of Psychiatry, 159*(Suppl 14), 6–13.

Berrios, G. E. (1996). *The history of mental symptoms: Descriptive psychopathology since the nineteenth century*. Cambridge University Press.

Bickle, J. (1992). Revisionary physicalism. *Biology and Philosophy*, *7*(4), 411-430. https://doi.org/10.1007/BF00130060

Blackburn, S. (2006). Antirealist expressivism and quasi-realism. In D. Coop (Ed). *The Oxford Handbook of Ethical Theory* (pp. 146-162). Oxford University Press.

Blackburn, S. (2008). *The Oxford dictionary of philosophy*. Oxford University Press.

Blackwood, N. J., Howard, R. J., Bentall, R. P., & Murray, R. M. (2001). Cognitive neuropsychiatric models of persecutory delusions. *The American journal of psychiatry*, *158*(4), 527–539. https://doi.org/10.1176/appi.ajp.158.4.527

Block, N. (1995). The mind as the software of the brain. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science* (pp. 377–425). The MIT Press.

Block, N. J., & Fodor, J. A. (1972). What psychological states are not. *Philosophical Review, 81*(2), 159–181. https://doi.org/10.2307/2183991

Bolton, D. (2013). What is mental illness? In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 434-450). Oxford University Press.

Bolton, D., & Gillett, G. (2019). *The biopsychosocial model of health and disease: New philosophical and scientific developments*. Springer Nature. https://doi.org/10.1007/978-3-030-11899-0

Boorse, C. (1975). On the distinction between disease and illness. *Philosophy & Public Affairs, 5*(1), pp. 49-68. https://www.jstor.org/stable/2265020

Boorse C. (1997). A Rebuttal on Health. In J.M. Humber & R.F. Almeder (Eds.), *What Is Disease?* (pp. 1 – 134). Humana Press. https://doi.org/10.1007/978-1-59259-451-1_1

Boorse, C. (2014). A second rebuttal on health. *Journal of Medicine and Philosophy*, *39*(6), 683-724. https://doi.org/10.1093/jmp/jhu035

Bordonaba, D., Fernández-Castro, V., & Torices, J. R. (Eds.). (2022). *The Political Turn in Analytic Philosophy: Reflections on Social Injustice and Oppression*. Walter de Gruyter GmbH & Co KG.

Borgoni, C. (2014). Dissonance and Irrationality: A Criticism of The In-Between Account of Dissonance Cases. *Pacific Philosophical Quarterly*, *97*(1), 48-57. https://doi.org/10.1111/papq.12039

Borgoni, C. (2019). Authority and Attribution: The Case of Epistemic Injustice in Self-Knowledge. *Philosophia, 47*, 293-301. https://doi.org/10.1007/s11406-018-0002-x

Borgoni, C. (forthcoming). First-person authority and the social aspects of self-knowledge. In J. Lackey, & A. McGlynn (Eds.), *The Oxford Handbook of Social Epistemology*. Oxford University Press.

Borsboom, D., Cramer, A. O., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*. https://doi.org/10.1017/S0140525X17002266

Bortolotti, L. (2010). *Delusions and other irrational beliefs*. Oxford University Press.

Bortolotti L. (2011) Double Bookkeeping in Delusions: Explaining the Gap between Saying and Doing. In J.H. Aguilar, A.A. Buckareff, & K. Frankish (Eds.), *New Waves in Philosophy of Action*. Palgrave Macmillan. https://doi.org/10.1057/9780230304253_12

Bortolotti L. (2012). In Defence of Modest Doxasticism About Delusions. *Neuroethics*, *5*(1), 39–53. https://doi.org/10.1007/s12152-011-9122-8

Bortolotti, L. (2018). *Delusions in context* (p. 130). Springer Nature.

Brakoulias, V., Langdon, R., Sloss, G., Coltheart, M., Meares, R., & Harris, A. (2008). Delusions and reasoning: A study involving cognitive behavioural therapy. *Cognitive Neuropsychiatry, 13*(2), 148–165. https://doi.org/10.1080/13546800801900587

Brandom, R. (2000). *Articulating reasons*. Harvard University Press.

Brisch, R., Saniotis, A., Wolf, R., Bielau, H., Bernstein, H. G., Steiner, J., ... & Gos, T. (2014). The role of dopamine in schizophrenia from a neurobiological and evolutionary perspective: old fashioned, but still in vogue. *Frontiers in psychiatry*, *5*, 47. https://doi.org/10.3389/fpsyt.2014.00047

Broad, C. D. (1925). *The Mind and its Place in Nature*. Kegan Paul, Trench, Trubner & Co.

Broome, M. R., Bortolotti, L., & Mameli, M. (2010). Moral responsibility and mental illness: a case study. *Cambridge quarterly of healthcare ethics: The international journal of healthcare ethics committees*, *19*(2), 179–187. https://doi.org/10.1017/S0963180109990442

Brown, J. R., & Fehige, Y. (2022). Thought Experiments. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2022 ed.). Metaphysics Research Lab, Stanford University https://plato.stanford.edu/archives/spr2022/entries/thought-experiment/

Brown, E., Shrestha, M., & Gray, R. (2021). The safety and efficacy of acceptance and commitment therapy against psychotic symptomatology: a systematic review and meta-analysis. *Revista brasileira de psiquiatria (Sao Paulo, Brazil : 1999)*, *43*(3), 324–336. https://doi.org/10.1590/1516-4446-2020-0948

Burns, C. E., Heiby, E. M., & Tharp, R. G. (1983). A verbal behavior analysis of auditory hallucinations. *The Behavior Analyst, The Behavior Analyst*, *6*, *6*(2, 2), 133, 133–143. https://doi.org/10.1007/BF03392392

Byrne, A. (1998). Interpretivism. *European Review of Philosophy*, *3*, 199-223.

Calero-Elvira, A., Froján-Parga, M. X., Ruiz-Sancho, E. M., & Alpañés-Freitag, M. (2013). Descriptive study of the Socratic method: evidence for verbal shaping. *Behavior therapy*, *44*(4), 625–638. https://doi.org/10.1016/j.beth.2013.08.001

Carel, H., & Kidd, I. J. (2014). Epistemic injustice in healthcare: a philosophial analysis. *Medicine, health care, and philosophy*, *17*(4), 529–540. https://doi.org/10.1007/s11019-014-9560-2

Caro, M., & Macarthur, D. (Eds.). (2004). *Naturalism in question*. Harvard University Press.

Caro, M., & Macarthur, D. (2022). Introduction. In M. Caro, & D. Macarthur (Eds.). *The Routledge Handbook of Liberal Naturalism* (pp. 1-4). Routledge.

Carr, J. E., & Britton, L. N. (1999). Idiosyncratic effects of noncontingent reinforcement on problematic speech. *Behavioral Interventions, 14*, 37-43.

Carruthers, P. (1996). Simulation and self-knowledge: A defence of the theory-theory. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 22-38). Cambridge University Press.

Carruthers, P. (2013). On Knowing Your Own Beliefs: A Representationalist Account. In N. Nottelman (Ed.) *New essays on belief* (pp. 145-165). Palgrave Macmillan.

Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge University Press.

Chadwick, P. D. J., & Lowe, C. F. (1990). Measurement and modification of delusional beliefs. *Journal of Consulting and Clinical Psychology, 58*(2), 225–232. https://doi.org/10.1037/0022-006X.58.2.225

Chadwick, P. D. J., & Lowe, C. F. (1994). A cognitive approach to measuring and modifying delusions. *Behaviour Research and Therapy, 32*(3), 355–367. https://doi.org/10.1016/0005-7967(94)90133-3

Chapman, R. (2020). Defining neurodiversity for research and practice. In H. B. Rosqvist, N. Chown & A. Stenning (Eds.), *Neurodiversity studies: A new critical paradigm* (pp. 218-220) Routledge.

Chemero, A. (2009). *Radical embodied cognitive science*. MIT Press.

Chiesa, M. (1994). *Radical behaviorism: The philosophy and the science*. Authors Cooperative.

Chong, H. Y., Teoh, S. L., Wu, D. B., Kotirum, S., Chiou, C. F., & Chaiyakunapruk, N. (2016). Global economic burden of schizophrenia: a systematic review. *Neuropsychiatric disease and treatment*, *12*, 357–373. https://doi.org/10.2147/NDT.S96649

Chrisman, M. (2007). From epistemic contextualism to epistemic expressivism. *Philosophical Studies*, *135*(2), 225-254. https://www.jstor.org/stable/40208748

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78*(2), 67-90. https://doi.org/10.2307/2025900

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7-19. https://www.jstor.org/stable/3328150

Clutton, P. (2018). A new defence of doxasticism about delusions: The cognitive phenomenological defence. *Mind & Language, 33*(2), 198–217. https://doi.org/10.1111/mila.12164

Cocchetti, C., Ristori, J., Romani, A., Maggi, M., & Fisher, A. D. (2020). Hormonal Treatment Strategies Tailored to Non-Binary Transgender Individuals. *Journal of clinical medicine*, *9*(6), 1609. https://doi.org/10.3390/jcm9061609

Cohen, B. M. (Ed.). (2018). *Routledge international handbook of critical mental health*. Routledge.

Coliva, A. (2016). *The Varieties of Self-Knowledge*. Palgrave Macmillan.

Colombo, A., Bendelow, G., Fulford, B., & Williams, S. (2003). Evaluating the influence of implicit models of mental disorder on processes of shared decision making within community-based multi-disciplinary teams. *Social science & medicine (1982)*, *56*(7), 1557–1570. https://doi.org/10.1016/s0277-9536(02)00156-9

Coltheart M. (2007). Cognitive neuropsychiatry and delusional belief. *Quarterly journal of experimental psychology (2006)*, *60*(8), 1041–1062. https://doi.org/10.1080/17470210701338071

Coltheart, M., Langdon, R., & McKay, R. (2011). Delusional belief. *Annual review of psychology*, *62*, 271–298. https://doi.org/10.1146/annurev.psych.121208.131622

Cooper, D. (1967). *Psychiatry and anti-psychiatry*. Tavistock Publications.

Cooper, R. (2014). *Diagnosing the Diagnostic and Statistical Manual of Mental Disorders*. Karnac Books.

Cooper, R. (2017). Where's the problem? Considering Laing and Esterson's account of schizophrenia, social models of disability, and extended mental disorder. *Theoretical medicine and bioethics*, *38*(4), 295-305. https://doi.org/10.1007/s11017-017-9413-0

Cooper, J. O., Heron, T. E., & Heward, W. L. (2019). *Applied Behavior Analysis (3rd Edition).* Pearson Education.

Cornman, J. W. (1968). On the elimination of'sensations' and sensations. *The Review of Metaphysics*, 15-35. https://www.jstor.org/stable/20124744

Corradini, A. & O'Connor, T. (2010). *Emergence in science and philosophy.* Routledge.

Craddock, N., Antebi, D., Attenburrow, M. J., Bailey, A., Carson, A., Cowen, P., Craddock, B., Eagles, J., Ebmeier, K., Farmer, A., Fazel, S., Ferrier, N., Geddes, J., Goodwin, G., Harrison, P., Hawton, K., Hunter, S., Jacoby, R., Jones, I., Keedwell, P., ... Zammit, S. (2008). Wake-up call for British psychiatry. *The British journal of psychiatry: the journal of mental science*, *193*(1), 6–9. https://doi.org/10.1192/bjp.bp.108.053561

Cramer, H., Lauche, R., Haller, H., Langhorst, J., & Dobos, G. (2016). Mindfulness- and Acceptance-based Interventions for Psychosis: A Systematic Review and Meta-analysis. *Global advances in health and medicine*, *5*(1), 30–43. https://doi.org/10.7453/gahmj.2015.083

Crichton, P., Carel, H., & Kidd, I. J. (2017). Epistemic injustice in psychiatry. *BJPsych bulletin*, *41*(2), 65–70. https://doi.org/10.1192/pb.bp.115.050682

Critchfield, T. S., & Rehfeldt, R. A. (2019). Engineering emergent learning with nonequivalence relations. In J. O. Cooper, T. E. Heron, & W. L. Heward (Eds.), *Applied Behavior Analysis* (pp. 497-526). New Jersey: Pearson Education.

Currie, G. (2000). Imagination, delusion and hallucinations. *Mind & Language, 15*(1), 168–183. https://doi.org/10.1111/1468-0017.00128

Currie, G., & Jureidini, J. (2001). Delusion, Rationality, Empathy: Commentary on Martin Davies et al.: *Philosophy, Psychiatry, & Psychology 8*(2), 159-162. doi:10.1353/ppp.2001.0006

Curry, D. S. (2020). Interpretivism and norms. *Philosophical studies*, *177*(4), 905-930. https://doi.org/10.1007/s11098-018-1212-6

Cuthbert, B. N. (2014). The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, *13*(1), 28-35. https://doi.org/10.1002/wps.20087

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine*, *11*(1), 1-8. https://doi.org/10.1186/1741-7015-11-126

David, A. S., & Halligan, P. W. (1996). Editorial. *Cognitive neuropsychiatry*, *1*(1), 1–4. https://doi.org/10.1080/135468096396659

David, A. S., & Halligan, P. W. (2000). Cognitive neuropsychiatry: Potential for progress. *The Journal of neuropsychiatry and clinical neurosciences*, *12*(4), 506-510. https://doi.org/10.1038/35058586

Davidson, D. (1984). First person authority. *Dialectica*, 101-111. https://www.jstor.org/stable/42970507

Davidson, D. (1986). A Coherence Theory of Truth and Knowledge. In E. Lepore (Ed.), *Truth and Interpretation: perspectives on the philosophy of Donald Davidson* (pp. 307-19). Basil Blackwell.

Davidson, D. (1991). Three varieties of knowledge. *Royal Institute of Philosophy Supplements*, *30*, 153-166. https://doi.org/10.1017/S1358246100007748

Davidson, D. (2001). Mental Events. In D. Davidson (Ed.), *Essays on Actions and Events* (pp. 207-229). Oxford University Press. (Original work published 1970).

Davies, M., & Egan, A. (2013). Delusion: Cognitive approaches—Bayesian inference and compartmentalization. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 689–727). Oxford University Press.

Davis, J. R., Wallace, C. J., Liberman, R. P., & Finch, B. E. (1976). The use of brief isolation to suppress delusional and hallucinatory speech. *Journal of Behavior Therapy and Experimental Psychiatry, 7*(3), 269–275. https://doi.org/10.1016/0005-7916(76)90012-4

De Haan, S. (2020a). An enactive approach to psychiatry. *Philosophy, Psychiatry, & Psychology*, *27*(1), 3-25. https://doi.org/10.1353/ppp.2020.0001

De Haan, S. (2020b). Bio-psycho-social interaction: an enactive perspective. *International Review of Psychiatry*, 1-7. https://doi.org/10.1080/09540261.2020.1830753

De Haan, S. (2020c). *Enactive psychiatry*. Cambridge University Press.

De Haan, S. (2021). Two Enactive Approaches to Psychiatry: Two Contrasting Views on What it Means to Be Human. *Philosophy, Psychiatry, & Psychology*, *28*(3), 191-196.

De Jaegher, H. (2013). Embodiment and sense-making in autism. *Frontiers in Integrative Neuroscience, 7,* Article 15. https://doi.org/10.3389/fnint.2013.00015

Deacon, B. (2013). The biomedical model of mental disorder: A critical analysis of its validity, utility, and effects on psychotherapy research. *Clinical psychology review*, *33*(7), 846-861. https://doi.org/10.1016/j.cpr.2012.09.007

Deacon, B., & Baird, G. (2009). The chemical imbalance explanation of depression: Reducing blame at what cost? *Journal of Social and Clinical Psychology*, *28*(4), 415-435.

Deacon, B., & McKay, D. (2015). The biomedical model of psychological problems: A call for critical dialogue. *The Behavior Therapist, 38*, 231-235

DeLeon, I. G., Arnold, K. L., Rodriguez-Catter, V., & Uy, M. L. (2003). Covariation between bizarre and nonbizarre speech as a function of the content of verbal attention. *Journal of Applied Behavior Analysis*, *36*(1), 101–104. https://doi.org/10.1901/jaba.2003.36-101

Demeter, T. (2013). Mental Fictionalism: The Very Idea. *The Monist, 96*(4), 483–504. https://doi.org/10.5840/monist201396422

Dennett, D. C. (1987). True believers: The intentional strategy and why it works. In D. C. Dennett (Ed.). *The Intentional Stance* (pp. 13-42). MIT Press. (Original work published 1979).

Descartes, R. (2008). *Meditations on First Philosophy, with selections from the objections and replies* (M. Moriarty, Ed., Trans.). Oxford University Press. (Original work published 1641).

Diez-Alegría, C., Vázquez, C., Nieto-Moreno, M., Valiente, C., & Fuentenebro, F. (2006). Personalizing and externalizing biases in deluded and depressed patients: are attributional biases a stable and specific characteristic of delusions?. *The British journal of clinical psychology*, *45*(Pt 4), 531–544. https://doi.org/10.1348/014466505X86681

Dings, R. (2020). Psychopathology, phenomenology and affordances. *Phenomenology & Mind*, *18*, 56-66. https://doi.org/10.13128/pam-1804

Dixon, M. R., Benedict, H., & Larson, T. (2001). Functional analysis and treatment of inappropriate verbal behavior. *Journal of Applied Behavior Analysis*, *34*(3), 361–363. https://doi.org/10.1901/jaba.2001.34-361

Dobson, K. S., & Dozois, D. J. A. (2010). Historical and philosophical bases of the cognitive-behavioral therapies. In K. S. Dobson (Ed.), *Handbook of cognitive-behavioral therapies* (pp. 3–38). Guilford Press.

Dougher M. J., (2004). *Clinical Behavior Analysis.* Context Press.

Dougher, M. J., & Hayes, S. C. (2004). Clinical Behavior Analysis. In M. J. Dougher (Ed.), *Clinical Behavior Analysis* (pp. 11-25). Context Press.

Drayson, Z. (2009). Embodied cognitive science and its implications for psychopathology. *Philosophy, Psychiatry, & Psychology*, *16*(4), 329-340. https://doi.org/10.1353/ppp.0.0261

Drożdżowicz A. (2021). Epistemic injustice in psychiatric practice: epistemic duties and the phenomenological approach. *Journal of medical ethics*, medethics-2020-106679. Advance online publication. https://doi.org/10.1136/medethics-2020-106679

Eells, T. D. (2007). History and current status of psychotherapy case formulation. In T. D. Eells (Ed.)., *Handbook of psychotherapy case formulation* (pp. 3-32). Guilford Publications.

Egan, A. (2008). Imagination, delusion, and self-deception. In T. Bayne & J. Fernandez (eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation (Macquarie Monographs in Cognitive Science)* (pp. 263-280). Psychology Press.

El-Mallakh, R., Gao, Y., & Robert, R. (2011). Tardive dysphoria: The role of long-term antidepressant use in inducing chronic depression. *Medical Hypotheses, 76*(6), 769-773.

Ellis, A. (1958). Rational Psychotherapy. *The Journal of General Psychology, 59*(1), 35-49. http://dx.doi.org/10.1080/00221309.1958.9710170

Ellis, A. (1962). *Reason and emotion in psychotherapy.* Lyle Stuart.

Ellis, H. D. (1998). Cognitive neuropsychiatry and delusional misidentification syndromes: an exemplary vindication of the new discipline. *Cognitive Neuropsychiatry*, *3*(2), 81-89. https://doi.org/10.1080/135468098396170

Ellis, H. D., & Young, A. W. (1990). Accounting for delusional misidentifications. *The British journal of psychiatry: the journal of mental science*, *157*, 239–248. https://doi.org/10.1192/bjp.157.2.239

Engel, G. L. (1960). A unified concept of health and disease. *Perspectives in biology and medicine*, *3*(4), 459-485.

Engel, G. L. (1977). The need for a new medical model: a challenge for biomedicine. *Science*, *196*(4286), 129-136. https://doi.org/10.1126/science.847460

Engel, G. L. (1978). The biopsychosocial model and the education of health professionals. *Annals of the New York Academy of Sciences*, *310*(1), 169-181. https://doi.org/10.1111/j.1749-6632.1978.tb22070.x

Engel, G. L. (1980). The clinical application of the biopsychosocial model. *American Journal of Psychiatry, 13*(5), 535-544.

Engel, G. L. (1997). From biomedical to biopsychosocial: Being scientific in the human domain. *Psychosomatics*, *38*(6), 521-528. https://doi.org/10.1016/S0033-3182(97)71396-3

Epstein, R., Lanza, R. P. & Skinner, B. F. (1980). Symbolic communication between two pigeons, (Columba livia domestica). *Science, 207*(4430), 543-545. https://doi.org/10.1126/science.207.4430.543

Epstein, R., Lanza, R. P. & Skinner, B. F. (1981). "Self-awareness" in the pigeon. *Science*, *212*(4495), 695-696. https://doi.org/10.1126/science.212.4495.695

Eysenck, H. J. (1959). Learning theory and behaviour therapy. *Journal of Mental Science, 105,* 61–75. https://doi.org/10.1192/bjp.105.438.61

Eysenck, H. J. (1960). Personality and behaviour therapy. *Proceedings of the Royal Society of Medicine, 53*(7), 504-508. https://doi.org/10.1177/003591576005300705

Eysenck, H. J. (Ed.). (1964). *Experiments in behaviour therapy*. Compton Printing.

Eysenck, H. J. (1972). Behavior therapy is behavioristic. *Behavior Therapy, 3*(4), 609–613. https://doi.org/10.1016/S0005-7894(72)80011-X

Eysenck, H. J. (1993). Forty years on: The outcome problem in psychotherapy revisited. In T. R. Giles (Ed.), *Handbook of effective psychotherapy* (pp. 3–20). Plenum Press. https://doi.org/10.1007/978-1-4615-2914-9_1

Eysenck, H. J., & Rachman, S. (2013). *The Causes and Cures of Neurosis (Psychology Revivals): An introduction to modern behaviour therapy based on learning theory and the principles of conditioning*. Routledge. (Original work published 1965). https://doi.org/10.4324/9780203766767

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Jr, Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of general psychiatry*, *26*(1), 57–63. https://doi.org/10.1001/archpsyc.1972.01750190059011

Feigl, H. (1958). The 'mental'and the 'physical'. *Minnesota studies in the philosophy of science*, *2*(2), 370-497. https://conservancy.umn.edu/handle/11299/184614

Fernandez Castro, V. (2017a). Regulation, normativity and folk psychology. *Topoi*, *39*(1), 57-67. https://doi.org/10.1007/s11245-017-9511-7

Fernandez Castro, V. (2017b). The expressive function of folk psychology. *Filosofia Unisinos/Unisinos Journal of Philosophy*, *18*(1), 36-46. https://doi.org/10.4013/fsu.2017.181.05

Fernández-Castro, V., & Heras-Escribano, M. (2020). Social Cognition: a normative approach. *Acta Analytica*, *35*(1), 75-100. https://doi.org/10.1007/s12136-019-00388-y

Fernández-Costa, D., Gómez-Salgado, J., Fagundo-Rivera, J., Martín-Pereira, J., Prieto-Callejero, B., & García-Iglesias, J. J. (2020). Alternatives to the use of mechanical restraints in the management of agitation or aggressions of psychiatric patients: A scoping review. *Journal of clinical medicine*, *9*(9), 2791.

Ferster, C. B. (1966). Animal behavior and mental illness. *The Psychological Record*, *16*(3), 345-356. https://doi.org/10.1007/BF03393678

Ferster, C. B. (1972). An experimental analysis of clinical phenomena. *The Psychological Record, 22*(1), 1–16. https://doi.org/10.1007/BF03394059

Ferster, C. B. (1973). A functional analysis of depression. *American Psychologist, 28*(10), 857–870. https://doi.org/10.1037/h0035605

Ferster, C. B., & DeMyer, M. K. (1962). A method for the experimental analysis of the behavior of autistic children. *American Journal of Orthopsychiatry, 32*(1), 89–98. https://doi.org/10.1111/j.1939-0025.1962.tb00267.x

Feuillet, L., Dufour, H., & Pelletier, J. (2007). Brain of a white-collar worker. *The Lancet*, *370*(9583), 262. https://doi.org/10.1016/S0140-6736(07)61127-1

Feyerabend, P. K. (1963a). Comment: Mental events and the brain. *The Journal of Philosophy*, *60*(11), 295-296. https://doi.org/10.2307/2023030

Feyerabend, P. K. (1963b). Materialism and the mind-body problem. *The Review of Metaphysics*, 49-66. https://www.jstor.org/stable/20123984

Field, H. (2009). Epistemology without metaphysics. *Philosophical studies*, *143*(2), 249-290. https://www.jstor.org/stable/27734403

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT press.

Fodor, J. A. (2006). The Language of Thought: First Approximations. In J. Bermúdez (Ed.), Philosophy of Psychology: Contemporary Readings (pp. 101-126). Routledge.

Follette, W. C., Naugle, A. E., & Callaghan, G. M. (1996). A radical behavioral understanding of the therapeutic relationship in effecting change. *Behavior therapy*, *27*(4), 623-641. https://doi.org/10.1016/S0005-7894(96)80047-5

Forsdyke, D. R. (2015). Wittgenstein's certainty is uncertain: Brain scans of cured hydrocephalics challenge cherished assumptions. *Biological Theory, 10*(4), 336-342. https://doi.org/10.1007/s13752-015-0219-x

Foster, C., Startup, H., Potts, L., & Freeman, D. (2010). A randomised controlled trial of a worry intervention for individuals with persistent persecutory delusions. *Journal of behavior therapy and experimental psychiatry*, *41*(1), 45–51. https://doi.org/10.1016/j.jbtep.2009.09.001

Foucault, M. (1965). *Madness and Civilization: A History of Insanity in the Age of Reason*. (R. Howard, Trans.). Random House. (Original work published 1961).

Frankish, K. (2009). Delusion: a two-level framework. In M. Broome, & L. Bortolotti (Eds.), *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives* (pp. 269–28). Oxford University Press.

Frankish, K. (2012). Delusions, levels of belief, and non-doxastic acceptances. *Neuroethics*, *5*(1), 23-27.

Frápolli, M. J., & Navarro-Laespada, L. (2021). I am large, I contain multitudes: Epistemic pragmatism, testimonial injustice and positive intersectionalism. *Daimon Revista Internacional de Filosofia, 84*. 115-129. https://doi.org/10.6018/daimon.481931

Frápolli, M. J. & Villanueva, N. (2012), Minimal Expressivism. *Dialectica, 66*, 471-487. https://doi.org/10.1111/1746-8361.12000

Frápolli, M. J., & Villanueva, N. (2013). Frege, Sellars, Brandom: expresivismo e inferencialismo semánticos. En Pérez Chico (coord.) *Perspectivas en la filosofía del lenguaje* (pp. 583-617). Prensas Universitarias de Zaragoza.

Freeman D. (2007). Suspicious minds: the psychology of persecutory delusions. *Clinical psychology review*, *27*(4), 425–457. https://doi.org/10.1016/j.cpr.2006.10.004

Freeman D. (2011). Improving cognitive treatments for delusions. *Schizophrenia research*, *132*(2-3), 135–139. https://doi.org/10.1016/j.schres.2011.08.012

Freeman, D., Dunn, G., Startup, H., Pugh, K., Cordwell, J., Mander, H., Černis, E., Wingham, G., Shirvell, K., & Kingdon, D. (2015). Effects of cognitive behaviour therapy for worry on persecutory delusions in patients with psychosis (WIT): A parallel, single-blind, randomised controlled trial with a mediation analysis. *The Lancet Psychiatry, 2*(4), 305–313. https://doi.org/10.1016/S2215-0366(15)00039-5

Freeman, D., & Garety, P. (2006). Delusions. In J. E. Fisher, & W. T. O'Donohue (Eds.), *Practitioner's guide to evidence-based psychotherapy* (pp. 205-213). Springer Science + Business Media. https://doi.org/10.1007/978-0-387-28370-8

Freeman, D., & Garety, P. (2014). Advances in understanding and treating persecutory delusions: a review. *Social psychiatry and psychiatric epidemiology*, *49*(8), 1179–1189. https://doi.org/10.1007/s00127-014-0928-7

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *The British journal of clinical psychology*, *41*(Pt 4), 331–347. https://doi.org/10.1348/014466502760387461

Freeman, D., Pugh, K., Dunn, G., Evans, N., Sheaves, B., Waite, F., Cernis, E., Lister, R., & Fowler, D. (2014). An early Phase II randomised controlled trial testing the effect on persecutory delusions of using CBT to reduce negative cognitions about the self: the potential benefits of enhancing self confidence. *Schizophrenia research*, *160*(1-3), 186–192. https://doi.org/10.1016/j.schres.2014.10.038

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia.* Lawrence Erlbaum Associates, Inc.

Froján-Parga, M. X. (2011). ¿Por qué funcionan los tratamientos psicológicos? *Clínica y salud*, *22*(3), 201-204.

Froján-Parga, M. X. (Coord.) (2020). *Análisis funcional de la conducta humana: concepto, metodología y aplicaciones*. Pirámide.

Froján-Parga, M. X., Alpañés-Freitag, M., Calero-Elvira, A., & Vargas-De La Cruz, I. (2010a). Una concepción conductual de la motivación en el proceso terapéutico. *Psicothema*, 556-561.

Froján-Parga, M. X., Calero-Elvira, A., & Montaño-Fidalgo, M. (2011). Study of the Socratic method during cognitive restructuring. *Clinical psychology & psychotherapy*, *18*(2), 110–123. https://doi.org/10.1002/cpp.676

Froján-Parga, M. X., Calero-Elvira, A., Pardo-Cebrián, R., & Núñez de Prado-Gordillo, M. (2018). Verbal change and cognitive change: Conceptual and methodological analysis for the study of cognitive restructuring using the Socratic dialog. *International Journal of Cognitive Therapy, 11*(2), 200–221. https://doi.org/10.1007/s41811-018-0019-8

Froján Parga, M. X., Montaño Fidalgo, M., & Calero Elvira, A. (2006). ¿Por qué la gente cambia en terapia? Un estudio preliminar. *Psicothema*, *18*(4), 797–803.

Froján-Parga, M. X., Montaño, M., & Calero, A. (2010b). Therapists' verbal behavior analysis: a descriptive approach to the psychotherapeutic phenomenon. *The Spanish journal of psychology*, *13*(2), 914–926. https://doi.org/10.1017/s1138741600002560

Froján-Parga, M. X., Montaño-Fidalgo, M., Calero-Elvira, A., Soler, Á. G., Fernández, Á. G., & Ruiz-Sancho, E. M. (2008). Sistema de categorización de la conducta verbal del terapeuta. *Psicothema*, *20*(4), 603-609.

Froján-Parga, M. X., Núñez de Prado-Gordillo, M., Álvarez-Iglesias, A., & Alonso-Vega, J. (2019). Functional Behavioral Assessment-based interventions on adults' delusions, hallucinations and disorganized speech: A single case meta-analysis. *Behaviour research and therapy*, *120*, 103444. https://doi.org/10.1016/j.brat.2019.103444

Froján Parga, M. X., Núñez de Prado Gordillo, M., & de Pascual Verdú, R. (2017). Cognitive techniques and language: A return to behavioral origins. *Psicothema*, *29*(3), 352–357. https://doi.org/10.7334/psicothema2016.305

Froján-Parga, M. X., Ruiz-Sancho, E. M., & Calero-Elvira, A. (2016). A theoretical and methodological proposal for the descriptive assessment of therapeutic interactions. *Psychotherapy research : journal of the Society for Psychotherapy Research*, *26*(1), 48–69. https://doi.org/10.1080/10503307.2014.935518

Fuchs, T. (2007). Psychotherapy of the lived space: a phenomenological and ecological concept. *American Journal of psychotherapy*, *61*(4), 423-439. https://doi.org/10.1176/appi.psychotherapy.2007.61.4.423

Fuchs, T. (2009). Embodied cognitive neuroscience and its consequences for psychiatry. *Poiesis & Praxis*, *6*(3-4), 219-233. https://doi.org/10.1007/s10202-008-0068-9

Fulford K. W. M. (2011). Bringing together values-based and evidence-based medicine: UK Department of Health Initiatives in the 'Personalization' of Care. *Journal of evaluation in clinical practice*, *17*(2), 341–343. https://doi.org/10.1111/j.1365-2753.2010.01578.x

Fulford, K. W. M., & Colombo, A. (2004). Six models of mental disorder: a study combining linguistic-analytic and empirical methods. *Philosophy, Psychiatry, & Psychology*, *11*(2), 129–144.

Fulford, K. W. M., Davies, M., Gipps, R., Graham, G., Sadler, J., Stanghellini, G., & Thornton, T. (Eds.). (2013a). *The Oxford handbook of philosophy and psychiatry*. Oxford University Press.

Fulford, K. W. M., Davies, M., Gipps, R. G., Graham, G., Sadler, J. Z., Stanghellini, G., & Thornton, T (2013b). The Next Hundred Years. Watching our P's and Q's. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 1–14). Oxford University Press.

Fulford, K. W. M., & Thornton, T. (2017). Delusions: A Project in Understanding. In T. Schramme, & S. Edwards (Eds.), *Handbook of the Philosophy of Medicine*, 557–576.

Fulford, K. W. M. & Van Staden, C. W. (2013). Values-Based Practice: Topsy-Turvy Take-Home Messages from Ordinary Language Philosophy (and a Few Next Steps). In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 385–412). Oxford University Press.

Gallagher S. (2008). Direct perception in the intersubjective context. *Consciousness and cognition*, *17*(2), 535–543. https://doi.org/10.1016/j.concog.2008.03.003

Gallagher, S. (2009). Delusional realities. In M. Broome, & L. Bortolotti (Eds.), *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives* (pp. 245–268). Oxford University Press.

Gallagher, S., & Varga, S. (2015). Social cognition and psychopathology: a critical overview. *World Psychiatry*, *14*(1), 5–14. https://doi.org/10.1002/wps.20173

García-Montes, J. M., & Pérez-Álvarez, M. (2005). Fundamentación experimental y primeras aplicaciones clínicas de la Terapia de Aceptación y Compromiso (ACT) en el campo de los síntomas psicóticos. *Revista Latinoamericana de Psicología*, *37*(2), 379–393.

Garety, P. (1991). Reasoning and delusions. The British Journal of Psychiatry, 159(Suppl 14), 14–18. https://doi.org/10.1192/S0007125000296426

Garety, P. A., Bebbington, P., Fowler, D., Freeman, D., & Kuipers, E. (2007). Implications for neurobiological research of cognitive models of psychosis: a theoretical paper. *Psychological medicine*, *37*(10), 1377–1391. https://doi.org/10.1017/S003329170700013X

Garety, P., Fowler, D., Kuipers, E., Freeman, D., Dunn, G., Bebbington, P., ... & Jones, S. (1997). London–East Anglia randomised controlled trial of cognitive–behavioural therapy for

psychosis: II: Predictors of outcome. *The british journal of psychiatry*, *171*(5), 420–426. https://doi.org/10.1192/bjp.171.5.420

Garety, P. A., & Freeman, D. (1999). Cognitive approaches to delusions: A critical review of theories and evidence. *British Journal of Clinical Psychology, 38*(2), 113–154. https://doi.org/10.1348/014466599162700

Garety, P. A., & Freeman, D. (2013). The past and future of delusions research: from the inexplicable to the treatable. *The British Journal of Psychiatry*, *203*(5), 327-333. https://doi.org/10.1192/bjp.bp.113.126953

Garety, P. A., Kuipers, E., Fowler, D., Freeman, D., & Bebbington, P. E. (2001). A cognitive model of the positive symptoms of psychosis. *Psychological medicine*, *31*(2), 189-195. https://doi.org/10.1017/S0033291701003312

Garety, P., Waller, H., Emsley, R., Jolley, S., Kuipers, E., Bebbington, P., Dunn, G., Fowler, D., Hardy, A., & Freeman, D. (2015). Cognitive mechanisms of change in delusions: An experimental investigation targeting reasoning to effect change in paranoia. *Schizophrenia Bulletin, 41*(2), 400–410. https://doi.org/10.1093/schbul/sbu103

Gaudiano, B. A. (Ed.). (2015). *Incorporating Acceptance and Mindfulness Into the Treatment of Psychosis: Current Trends and Future Directions*. New York: Oxford University Press.

Gaudiano, B. A., Busch, A. M., Wenze, S. J., Nowlan, K., Epstein-Lubow, G., & Miller, I. W. (2015). Acceptance-based Behavior Therapy for Depression With Psychosis: Results From a Pilot Feasibility Randomized Controlled Trial. *Journal of psychiatric practice*, *21*(5), 320–333. https://doi.org/10.1097/PRA.0000000000000092

Gaudiano, B. A., Ellenberg, S., Ostrove, B., Johnson, J., Mueser, K. T., Furman, M., & Miller, I. W. (2020). Feasibility and Preliminary Effects of Implementing Acceptance and Commitment Therapy for Inpatients With Psychotic-Spectrum Disorders in a Clinical Psychiatric Intensive Care Setting. *Journal of cognitive psychotherapy*, *34*(1), 80–96. https://doi.org/10.1891/0889-8391.34.1.80

Gaudiano, B. A., & Herbert, J. D. (2006). Acute treatment of inpatients with psychotic symptoms using Acceptance and Commitment Therapy: pilot results. *Behaviour research and therapy*, *44*(3), 415–437. https://doi.org/10.1016/j.brat.2005.02.007

Gaudiano, B. A., Herbert, J. D., & Hayes, S. C. (2010). Is it the symptom or the relation to it? Investigating potential mediators of change in acceptance and commitment therapy for psychosis. *Behavior therapy*, *41*(4), 543–554. https://doi.org/10.1016/j.beth.2010.03.001

Gaudiano, B. A., Nowlan, K., Brown, L. A., Epstein-Lubow, G., & Miller, I. W. (2013). An open trial of a new acceptance-based behavioral treatment for major depression with

psychotic features. *Behavior modification*, *37*(3), 324–355. https://doi.org/10.1177/0145445512465173

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet (London, England)*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

GBD 2019 Diseases and Injuries Collaborators (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet (London, England)*, *396*(10258), 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Gerrans, P. (2004). Cognitive architecture and the limits of interpretationism. *Philosophy, Psychiatry, & Psychology*, *11*(1), 43–48.

Ghaemi, S. N. (2009). The rise and fall of the biopsychosocial model. *The British Journal of Psychiatry*, *195*(1), 3–4. https://doi.org/10.1192/bjp.bp.109.063859

Ghaemi, S. N. (2010). *The rise and fall of the biopsychosocial model*. The Johns Hopkins University Press.

Gibbard, A. (1986). An expressivistic theory of normative discourse. *Ethics*, *96*(3), 472–485. https://www.jstor.org/stable/2381066

Gibson, J. J. (2015). *The ecological approach to visual perception*. Psychology Press. (Original work published 1979).

Giladi, P. (Ed.). (2019). *Responses to naturalism: Critical perspectives from idealism and pragmatism*. Routledge.

Glackin, S. N., Roberts, T., & Krueger, J. (2021). Out of our heads: Addiction and psychiatric externalism. *Behavioural Brain Research*, *398*, 112936. https://doi.org/10.1016/j.bbr.2020.112936

Goddard, M. J. (2014). Critical Psychiatry, Critical Psychology, and the Behaviorism of B. F. Skinner. *Review of General Psychology*, *18*(3), 208–215. https://doi.org/10.1037/gpr0000012

Goffman, E. (1963). *Behavior in public places*. The Free Press.

Goldstone, E., Farhall, J., & Ong, B. (2011). Life hassles, experiential avoidance and distressing delusional experiences. *Behaviour research and therapy*, *49*(4), 260–266. https://doi.org/10.1016/j.brat.2011.02.002

González-Pardo, H., & Pérez-Álvarez, M. (2007). *La invención de los trastornos mentales. ¿Escuchando al fármaco o al paciente?* Alianza.

Gordon, R. M. (1996). 'Radical' simulationism. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 11-21). Cambridge University Press.

Graham, G. (2010). Are the deluded believers? Are philosophers among the deluded? *Philosophy, Psychiatry, & Psychology, 17*(4), 337–339. https://doi.org/10.1353/ppp.2010.0033

Graham, G. (2010). *The disordered mind: An introduction to philosophy of mind and mental illness*. Routledge.

Graham, G., & Stephens, G. L. (Eds.). (1994). *Philosophical psychopathology*. MIT Press.

Greenwood, J., D. (2015). Neobehaviorism, radical behaviorism, and problems of behaviorism. In J. D. Greenwood (Ed.), *A Conceptual History of Psychology: Exploring the Tangled Web* (pp. 410-453). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107414914.012

Guinther, P. M., & Dougher, M. J. (2013). From behavioral research to clinical therapy. In *APA handbook of behavior analysis, Vol. 2: Translating principles into practice.* (pp. 3-32). American Psychological Association.

Halligan, P. W., & David, A. S. (2001). Cognitive neuropsychiatry: towards a scientific psychopathology. *Nature reviews. Neuroscience*, *2*(3), 209–215. https://doi.org/10.1038/35058586

Hamilton, A. (2006). 11 Against the belief model of delusion. In M. C. Chung, K. W. M. Fulford, & G. Graham (eds.) *Reconceiving schizophrenia* (217-234). OUP.

Hanley, G. P., Iwata, B. A., & McCord, B. E. (2003). Functional analysis of problem behavior: A review. *Journal of Applied Behavior Analysis, 36*(2), 147–185. https://doi.org/10.1901/jaba.2003.36-147

Haslam, N. (2013). Reliability, validity, and the mixed blessings of operationalism. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 987-1002). Oxford University Press.

Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior therapy*, *35*(4), 639-665. https://doi.org/10.1016/S0005-7894(04)80013-3

Hayes, S. C. (2016). Why Contextual Behavioral Science exists. An Introduction to Part I. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.). *The Wiley handbook of contextual behavioral science* (pp. 9-16). John Wiley & Sons.

Hayes, S. C. (2021). Contextual Behavioral Science as a Distinct Form of Behavioral Research and Practice. In D. Zilio & K. Carrara (Eds.), *Contemporary Behaviorisms in Debate* (pp. 239-255). Springer.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational Frame Theory: A Post-Skinnerian account of human language and cognition*. New York: Plenum Press.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2003). Behavior analysis, relational frame theory, and the challenge of human language and cognition: A reply to the commentaries on Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition. *The Analysis of Verbal Behavior*, *19*(1), 39-54. https://doi.org/10.1007/BF03392981

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual behavioral science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science, 1*(1-2), 1–16. https://doi.org/10.1016/j.jcbs.2012.09.004

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (1999). *Acceptance and commitment therapy: An experiential approach to behavior change.* Guilford Press.

Haynes, S. N., & Geddy, P. (1973). Suppression of psychotic hallucinations through time-out. *Behavior Therapy, 4*(1), 123–127. https://doi.org/10.1016/S0005-7894(73)80083-8

Hempel, C. G. (1965). Fundamentals of taxonomy. In C. G. Hempel, *Aspects of scientific explanation* (pp. 137 - 154). Free Press.

Heras-Escribano, M. (2019). *The philosophy of affordances*. Palgrave Macmillan.

Heras-Escribano, M., Noble, J., & De Pinedo, M. (2015). Enactivism, action and normativity: a Wittgensteinian analysis. *Adaptive Behavior*, *23*(1), 20-33. https://doi.org/10.1177/1059712314557364

Heras-Escribano, M., & Pinedo-García, M. D. (2018). Naturalism, non-factualism, and normative situated behaviour. *South African Journal of Philosophy*, *37*(1), 80-98. https://doi.org/10.1080/02580136.2017.1422633

Hoffman, G. A. (2016). Out of our skulls: How the extended mind thesis can extend psychiatry. *Philosophical Psychology*, *29*(8), 1160-1174. https://doi.org/10.1080/09515089.2016.1236369

Hohwy, J., & Rajan, V. (2012). Delusions as forensically disturbing perceptual inferences. *Neuroethics*, *5*(1), 5-11. https://doi.org/10.1007/s12152-011-9124-6

Hohwy, J., & Rosenberg, R. (2005). Cognitive neuropsychiatry: conceptual, methodological and philosophical perspectives. *The world journal of biological psychiatry: the official journal of the World Federation of Societies of Biological Psychiatry*, *6*(3), 192–197. https://doi.org/10.1080/15622970510029867

Horner, R. H., Albin, R. W., & Mank, D. M. (1989). Effects of undesirable, competing behaviors on the generalization of adaptive skills. A case study. *Behavior Modification*, *13*(1), 74–90. https://doi.org/10.1177/01454455890131005

Hull, C. L. (1945). The place of innate individual and species differences in a natural-science theory of behavior. *Psychological Review,* *52*(2), 55–60. https://doi.org/10.1037/h0056383

Hume, D. (1896). *A Treatise of Human Nature* (L. A. Selby-Bigge, Ed.). Oxford: Clarendon Press (Original work published 1739-1740). https://oll.libertyfund.org/title/bigge-a-treatise-of-human-nature

Humphreys, P. (1997). Emergence, not supervenience. *Philosophy of science*, *64*, S337-S345.

Hurl, K., Wightman, J., Haynes, S. N., & Virues-Ortega, J. (2016). Does a pre-intervention functional assessment increase intervention effectiveness? A meta-analysis of within-subject interrupted time-series studies. *Clinical psychology review*, *47*, 71-84. https://doi.org/10.1016/j.cpr.2016.05.003

Hurley, S. (2001). Perception and action: alternative views. *Synthese, 29*, 3-40. https://doi.org/10.1023/A:1012643006930

Hutto, D. (2022). Relaxed naturalism: a liberating philosophy of nature. In M. Caro, & D. Macarthur (Eds.). *The Routledge Handbook of Liberal Naturalism* (pp. 165-176). Routledge.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. MIT press.

Hutton, P., Wood, L., Taylor, P. J., Irving, K., & Morrison, A. P. (2014). Cognitive behavioural therapy for psychosis: rationale and protocol for a systematic review and meta-analysis. *Psychosis*, *6*(3), 220-230. https://doi.org/10.1080/17522439.2013.825005

Hyland, P., & Boduszek, D. (2012). Resolving a difference between cognitive therapy and rational emotive behaviour therapy: Towards the development of an integrated CBT model of psychopathology. Mental Health Review Journal. 17(2). 104-116. https://doi.org/10.1108/13619321211270425

Insel, T. R. (2013). Transforming diagnosis. *NIHM's website*. http://psychrights.org/2013/130429NIMHTransformingDiagnosis.htm

Insel T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *The American journal of psychiatry*, *171*(4), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science*, *348*(6234), 499-500. https://doi.org/10.1126/science.aab2358

Insel, T. R., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. https://doi.org/10.1176/appi.ajp.2010.09091379

Institute of health Metrics and Evaluation (IHME). Global Health Data Exchange (GHDx). http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/27a7644e8ad28e739382d31e77589dd7 (Accessed 28/01/2022)

Irvine, E., & Sprevak, M. (2020). Eliminativism about consciousness. In U. Kriegel (Ed.), *Oxford Handbook of the Philosophy of Consciousness* (pp. 348–370) Oxford University Press.

Isaacs, W., Thomas, J., & Goldiamond, I. (1960). Application of operant conditioning to reinstate verbal behavior in psychotics. *Journal of Speech & Hearing Disorders, 25,* 8–12. https://doi.org/10.1044/jshd.2501.08

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of applied behavior analysis*, *27*(2), 197-209. (Original work published 1982) https://doi.org/10.1901/jaba.1994.27-197

Iwata, B. A., Pace, G. M., Dorsey, M. F., Zarcone, J. R., Vollmer, T. R., Smith, R. G., ... & Willis, K. D. (1994). The functions of self-injurious behavior: An experimental-epidemiological analysis. *Journal of applied behavior analysis*, *27*(2), 215-240. https://doi.org/10.1901/jaba.1994.27-215

Jacob, P. (2019). Intentionality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/intentionality

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., Gortner, E., & Prince, S. E. (1996). A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting and Clinical Psychology, 64*(2), 295–304. https://doi.org/10.1037/0022-006X.64.2.295

James, W. (1981). *The principles of psychology*. Harvard University Press. (Original work published 1890).

Jansen, J. E., Gleeson, J., Bendall, S., Rice, S., & Alvarez-Jimenez, M. (2020). Acceptance- and mindfulness-based interventions for persons with psychosis: A systematic review

and meta-analysis. *Schizophrenia research*, *215*, 25–37. https://doi.org/10.1016/j.schres.2019.11.016

Jaspers, K. (1963). *General psychopathology*. University of Chicago Press. (Original work published 1913).

Jauhar, S., McKenna, P. J., Radua, J., Fung, E., Salvador, R., & Laws, K. R. (2014). Cognitive–behavioural therapy for the symptoms of schizophrenia: systematic review and meta-analysis with examination of potential bias. *The British Journal of Psychiatry*, *204*(1), 20–29. https://doi.org/10.1192/bjp.bp.112.116285

Jimenez, J. M., Todman, M., Perez, M., Godoy, J. F., & Landon-Jimenez, D. V. (1996). The behavioral treatment of auditory hallucinatory responding of a schizophrenic patient. *Journal of behavior therapy and experimental psychiatry*, *27*(3), 299–310. https://doi.org/10.1016/s0005-7916(96)00033-x

Jones, C., Hacker, D., Meaden, A., Cormac, I., Irving, C. B., Xia, J., ... & Chen, J. (2018a). Cognitive behavioural therapy plus standard care versus standard care plus other psychosocial treatments for people with schizophrenia. *Cochrane Database of Systematic Reviews*, (11). https://doi.org/10.1002/14651858.CD007964.pub2

Jones, C., Hacker, D., Meaden, A., Cormac, I., Irving, C. B., Xia, J., ... & Chen, J. (2018b). Cognitive behavioural therapy plus standard care versus standard care plus other psychosocial treatments for people with schizophrenia. *Cochrane Database of Systematic Reviews*, (11). https://doi.org/10.1002/14651858.CD008712.pub3

Kalis A. (2019). No Intentions in the Brain: A Wittgensteinian Perspective on the Science of Intention. *Frontiers in psychology*, *10*, 946. https://doi.org/10.3389/fpsyg.2019.00946

Kallmann, F. J. (1946). The genetic theory of schizophrenia; an analysis of 691 schizophrenic twin index families. *The American journal of psychiatry*, *103*(3), 309–322. https://doi.org/10.1176/ajp.103.3.309

Kaney, S., & Bentall, R. P. (1989). Persecutory delusions and attributional style. *British Journal of Medical Psychology, 62*(2), 191–198. https://doi.org/10.1111/j.2044-8341.1989.tb02826.x

Kandel, E. R. (2005). *Psychiatry, psychoanalysis, and the new biology of mind*. American Psychiatric Pub.

Kanfer, F. H., & Saslow, G. (1965). Behavioral analysis: An alternative to diagnostic classification. *Archives of General Psychiatry, 12*(6), 529–538. https://doi.org/10.1001/archpsyc.1965.01720360001001

Karasu, T. B. (1982). Psychotherapy and pharmacotherapy: Toward an integrative model. *The American Journal of Psychiatry, 139*(9), 1102–1113. https://doi.org/10.1176/ajp.139.9.1102

Kazdin, A. E. (1982). History of behavior modification. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 3-32). Springer.

Kendell, (1975). The Concept of Disease and its Implications for Psychiatry. *British Journal of Psychiatry, 127*, 305-315.

Kendell, R. E. (2004). The myth of mental illness. In J. A. Schaler (Ed.). (2004). *Szasz under fire: A psychiatric abolitionist faces his critics* (pp. 29-48). Open Court Publishing.

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15*(1), 5-12. https://doi.org/10.1002/wps.20292

Kendler, K. S., & Parnas, J. (Eds.). (2015). *Philosophical issues in psychiatry: Explanation, phenomenology, and nosology*. JHU Press.

Kendler, K. S., & Schaffner, K. F. (2011). The dopamine hypothesis of schizophrenia: An historical and philosophical analysis. *Philosophy, Psychiatry, & Psychology, 18*(1), 41–63. https://doi.org/10.1353/ppp.2011.0005

Keshavan, M., Clementz, B., Pearlson, G., Sweeney, J., & Tamminga, C. (2013). Reimagining psychoses: an agnostic approach to diagnosis. *Schizophrenia Research, 146* (1-3), 10-16.

Keshavan, M., Nasrallah, H., & Tandon, R. (2011). Schizophrenia, "Just the Facts" 6. Moving ahead with the schizophrenia concept: from the elephant to the mouse. *Schizophrenia research*, *127*(1-3), 3-13. doi:10.1016/j.schres.2011.01.011

Kety, S. S., Rosenthal, D., Wender, P. H., & Schulsinger, F. (1968). The types and prevalence of mental illness in the biological and adoptive families of adopted schizophrenics. In D. Rosenthal, S. S. Kety (Eds.), *The Transmission of Schizophrenia* (pp. 345-362). Pergamon Press.

Kety, S. S., Rosenthal, D., Wender, P. H., & Schulsinger, F. (1971). Mental illness in the biological and adoptive families of adpoted schizophrenics. *The American journal of psychiatry*, *128*(3), 302–306. https://doi.org/10.1176/ajp.128.3.302

Khoury, B., Lecomte, T., Gaudiano, B. A., & Paquin, K. (2013). Mindfulness interventions for psychosis: a meta-analysis. *Schizophrenia research*, *150*(1), 176–184. https://doi.org/10.1016/j.schres.2013.07.055

Kidd, I. J., Medina, J., & Pohlhaus Jr, G. (Eds.). (2017). *The Routledge handbook of epistemic injustice*. Taylor & Francis.

Kilbride, M., Byrne, R., Price, J., Wood, L., Barratt, S., Welford, M., & Morrison, A. P. (2013). Exploring service users' perceptions of cognitive behavioural therapy for psychosis: a user led study. *Behavioural and Cognitive Psychotherapy*, *41*(1), 89-102. https://doi.org/10.1017/S1352465812000495

Kim, J. (1992). 'Downward Causation' in Emergentism and Nonreductive Physicalism. In A. Beckermann, H. Flohr, & J. Kim (Eds.), *Emergence or Reduction?*. De Gruyter

Kim, J. (1993). The Nonreductivist's Troubles with Mental Causation. In E. Sosa (Ed.), *Supervenience and mind* (pp. 336-357). Cambridge University Press.

Kim, J. (2011). *Philosophy of mind*. Westview Press.

Kinderman, P., & Bentall, R. P. (1997). Causal attributions in paranoia and depression: Internal, personal, and situational attributions for negative events. *Journal of Abnormal Psychology, 106*(2), 341–345. https://doi.org/10.1037/0021-843X.106.2.341

Kingdon, D. G., & Turkington, D. (1991). A role for cognitive-behavioural strategies in schizophrenia? *Social Psychiatry and Psychiatric Epidemiology*, *26*(3), 101-103. https://doi.org/10.1007/BF00782948

Kiverstein, J., & Clark, A. (2009). Introduction: Mind embodied, embedded, enacted: One church or many? *Topoi, 28*(1), 1–7. https://doi.org/10.1007/s11245-008-9041-4

Klerman, G. L. (1978). The evolution of a scientific nosology. In J. C. Shershow (Ed.). *Schizophrenia, Science and Practice* (pp. 248). Harvard University Press.

Knapp, P., & Beck, A. T. (2008). Cognitive therapy: Foundations, conceptual models, applications and research. *Revista Brasileira de Psiquiatria, 30*(Suppl2), S54-S64. https://doi.org/10.1590/S1516-44462008000600002

Kohlenberg, R. J., Bolling, M. Y., Kanter, J. W., & Parker, C. R. (2002). Clinical behavior analysis: Where it went wrong, how it was made good again, and why its future is so bright. *The Behavior Analyst Today, 3*(3), 248-253. http://dx.doi.org/10.1037/h0099988

Kohlenberg, R. J., & Tsai, M. (1991). *Functional analytic psychotherapy: Creating intense and curative therapeutic relationships*. New York: Plenum Press.

Kohlenberg, R. J., Tsai, M., & Dougher, M. J. (1993). The dimensions of clinical behavior analysis. *The Behavior Analyst*, *16*(2), 271-282. https://doi.org/10.1007/BF03392636

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., ... & Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, *126*(4), 454- 477. https://doi.org/10.1037/abn0000258

Kotov, R., Krueger, R. F., & Watson, D. (2018). A paradigm shift in psychiatric classification: The Hierarchical Taxonomy Of Psychopathology (HiTOP). *World Psychiatry, 17*(1), 24–25. https://doi.org/10.1002/wps.20478

Kotov, R., Jonas, K. G., Carpenter, W. T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K., Reininghaus, U., Slade, T., South, S. C., Sunderland, M., Waszczuk, M. A., Widiger, T. A., Wright, A., Zald, D. H., Krueger, R. F., Watson, D., & HiTOP Utility

Workgroup (2020). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, *19*(2), 151–172. https://doi.org/10.1002/wps.20730

Kriegel, U. (2015). *The varieties of consciousness*. Oxford University Press.

Kripke, S. (1982). *Wittgenstein on rules and private language.* Harvard University Press.

Krueger, J. (2020). Schizophrenia and the scaffolded self. *Topoi*, *39*(3), 597-609. https://doi.org/10.1007/s11245-018-9547-3

Krueger, J. (2021). Enactivism, other minds, and mental disorders. *Synthese*, *198*(1), 365-389. https://doi.org/10.1007/s11229-019-02133-9

Krueger, J., & Colombetti, G. (2018). Affective affordances and psychopathology. *Affective affordances and psychopathology*, 221-246. https://doi.org/10.1400/266056

Krueger, J., & Maiese, M. (2018). Mental institutions, habits of mind, and an extended approach to autism. *Thaumàzein Rivista di Filosofia*, *6*, 10-41. https://doi.org/10.13136/thau.v6i0.90

Kuipers, E., Garety, P., Fowler, D., Dunn, G., Bebbington, P., Freeman, D., & Hadley, C. (1997). London–East Anglia randomised controlled trial of cognitive–behavioural therapy for psychosis: I: Effects of the treatment phase. *The British Journal of Psychiatry*, *171*(4), 319-327. https://doi.org/10.1192/bjp.171.4.319

Kupfer, D. J., First, M. B., Regier, D. A. (2002). *A Research Agenda For DSM-V*. American Psychiatric Association.

Kvaale, E. P., Haslam, N., & Gottdiener, W. H. (2013). The 'side effects' of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. *Clinical psychology review*, *33*(6), 782-794. https://doi.org/10.1016/j.cpr.2013.06.002

Lacasse, J., & Leo, J. (2015) Challenging the narrative of chemical imbalance: A look at the evidence. En B. Probst (Ed.), *Critical Thinking in Clinical Diagnosis and Assessment* (pp. 275-282). New York: Springer.

Lader, M., Tylee, A., & Donoghue J. (2009). Withdrawing benzodiazepines in primary care. *CNS Drugs, 23*, 19-34.

Lancaster, B. M., LeBlanc, L. A., Carr, J. E., Brenske, S., Peet, M. M., & Culver, S. J. (2004). Functional analysis and treatment of the bizarre speech of dually diagnosed adults. *Journal of Applied Behavior Analysis*, *37*(3), 395–399. https://doi.org/10.1901/jaba.2004.37-395

Laing, R. (2010). *The divided self: An existential study in sanity and madness*. Penguin. (Original work published 1960).

Laing, R. D., & Cooper, D. G. (1964). *Reason and violence: A decade of Sartre's philosophy, 1950-1960*. Tavistock Publications.

Langdon, R., McKay, R., & Coltheart, M. (2008). The cognitive neuropsychological understanding of persecutory delusions. In D. Freeman, R. Bentall, P. Garety (Eds.), *Persecutory delusions: assessment, theory, and treatment* (pp. 221-236). OUP.

Laws K. R. (2016). Commentary: Does Cognitive Behavior Therapy for psychosis (CBTp) show a sustainable effect on delusions? A meta-analysis. *Frontiers in psychology*, *7*, 59. https://doi.org/10.3389/fpsyg.2016.00059

Layng, T. V., & Andronis, P. T. (1984). Toward a functional analysis of delusional speech and hallucinatory behavior. *The Behavior analyst*, *7*(2), 139–156. https://doi.org/10.1007/BF03391897

Lazare, A. (1973). Hidden conceptual models in clinical psychiatry. *New England Journal of Medicine*, *288*(7), 345-351. https://doi.org/10.1056/NEJM197302152880705

Lazarus, A. A. (1958). New methods in psychotherapy: A case study. *South African Medical Journal*, *32*(26), 660-663. https://journals.co.za/doi/pdf/10.10520/AJA20785135_39134

Lazarus, A. A. (1968). Variations in desensitization therapy. *Psychotherapy: Theory, Research & Practice, 5*(1), 50–52. https://doi.org/10.1037/h0088651

Lazarus, A. A. (1977). Has behavior therapy outlived its usefulness? *American Psychologist, 32*(7), 550–554. https://doi.org/10.1037/0003-066X.32.7.550

Lazarus, A. A. & Rachman, S. (1957). The use of systematic desensitization in psychotherapy. *South African Medical Journal*, *31*(37), 934-937. https://journals.co.za/doi/pdf/10.10520/AJA20785135_45107

Ledwidge, B. (1978). Cognitive behavior modification: A step in the wrong direction? *Psychological Bulletin, 85*(2), 353–375. https://doi.org/10.1037/0033-2909.85.2.353

Lehmann, H. E., & Hanrahan, G. E. (1954). Chlorpromazine; new inhibiting agent for psychomotor excitement and manic states. *A.M.A. archives of neurology and psychiatry*, *71*(2), 227–237.

Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, *21*(1), 133–145. https://doi.org/10.1002/wps.20941

Leoni, F. (2013). From Madness to Mental Illness: Psychiatry and Biopolitics in Michel Foucault. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini,

& T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 85-94). Oxford University Press.

Lewin, R. (1980). Is Your Brain Really Necessary? *Science, 210*(4475), 1232-1234. https://www.science.org/doi/10.1126/science.7434023

Lewis, D. K. (1966). An argument for the identity theory. *The Journal of Philosophy, 63*(1), 17-25. https://doi.org/10.2307/2024524

Lewis, D. K. (1980). Mad pain and Martian pain. In N. Block (ed.), *Readings in the Philosophy of Psychology* (pp. 216-222). Harvard University Press.

Lincoln, T. M., & Peters, E. (2019). A systematic review and discussion of symptom specific cognitive behavioural approaches to delusions and hallucinations. *Schizophrenia research*, *203*, 66-79. https://doi.org/10.1016/j.schres.2017.12.014

Lindsley, O. R. (1956). Operant conditioning methods applied to research in chronic schizophrenia. *Psychiatric Research Reports*, *5*, 118–139. http://bindel.verio.com/wb_fluency.org/LabResearch/Lindsley1956Operant.pdf

Lindsley, O. R. (1959). Reduction in rate of vocal psychotic symptoms by differential positive reinforcement. *Journal of the Experimental Analysis of Behavior, 2*(269), 269. http://bindel.verio.com/wb_fluency.org/LabResearch/Lindsley1959.pdf

Lindsley, O. R. (1962). Operant conditioning methods in diagnosis. In *Psychosomatic Medicine: The First Hahnemann Symposium. Philadelphia: Lea & Febiger* (pp. 41-54). http://bindel.verio.com/wb_fluency.org/LabResearch/Lindsley1962a.pdf

Lindsley, O. R. (1963). Free-operant conditioning and psychotherapy. *Current psychiatric therapies*, *3*, 47-56. http://bindel.verio.com/wb_fluency.org/LabResearch/Lindsley1963c.pdf

Lindsley, O. R. (1964). Characteristics of the behavior of chronic psychotics as revealed by free-operant conditioning methods. In C. M. Franks (Ed.), *Conditioning Techniques in Clinical Practice and Research* (pp. 231-254). Springer.

Link, B. G., Cullen, F. T., Struening, E., Shrout, P. E., & Dohrenwend, B. P. (1989). A modified labeling theory approach to mental disorders: An empirical assessment. *American sociological review*, 400-423. https://doi.org/10.2307/2095613

López-Silva, P. (2016). The typology problem and the doxastic approach to delusions. *Filosofia Unisinos/Unisinos Journal of Philosophy*, *17*(2), 202-211. https://doi.org/10.4013/fsu.2016.172.15

López-Silva, P. (2021). La marca de la psicosis: hacia una síntesis del problema tipológico de los delirios. *Revista Colombiana de Psiquiatría*.

Louise, S., Fitzpatrick, M., Strauss, C., Rossell, S. L., & Thomas, N. (2018). Mindfulness- and acceptance-based interventions for psychosis: Our current understanding and a meta-analysis. *Schizophrenia research*, *192*, 57–63. https://doi.org/10.1016/j.schres.2017.05.023

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that everyone has won and all must have prizes? *Archives of general psychiatry*, *32*(8), 995-1008. 10.1001/archpsyc.1975.01760260059004

Lycan, W. G., & Pappas, G. S. (1972). What is eliminative materialism? *Australasian Journal of Philosophy*, *50*(2), 149-159. https://doi.org/10.1080/00048407212341181

Lynch, D., Laws, K. R., & McKenna, P. J. (2010). Cognitive behavioural therapy for major psychiatric disorder: does it really work? A meta-analytical review of well-controlled trials. *Psychological medicine*, *40*(1), 9-24. https://doi.org/10.1017/S003329170900590X

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological review*, *55*(2), 95-107. https://doi.org/10.1037/h0056029

Mace, F. C. (1994). The significance and future of functional-analysis methodologies. Journal of Applied Behavior Analysis, 27(2), 385–392. https://doi.org/10.1901/jaba.1994.27-385

Mace, F. C., Lalli, J. S., & Lalli, E. P. (1991). Functional analysis and treatment of aberrant behavior. *Research in Developmental Disabilities*, *12*(2), 155–180. https://doi.org/10.1016/0891-4222(91)90004-C

Mace, F. C., Webb, M. E., Sharkey, R. W., Mattson, D. M., & Rosen, H. S. (1988). Functional analysis and treatment of Bizarre speech. *Journal of Behavior Therapy and Experimental Psychiatry*, *19*(4), 289–296. https://doi.org/10.1016/0005-7916(88)90060-2

Madden, G. J., Hanley, G. P., & Dougher, M. J. (2016). Clinical behavior analysis. In J. C. Norcross, G. R. VandenBos, D. K. Freedheim, & M. M. Domenech Rodríguez (Eds.). *APA handbook of clinical psychology: Roots and branches, Vol. 1* (pp. 351-368). American Psychological Association.

Maher, B. (2005). Delusional thinking and cognitive disorder. *Integrative Physiological & Behavioral Science*, *40*(3), 136-146. (Original work published in 1974). https://doi.org/10.1007/BF03159710

Mahoney, M. J. (1977a). Cognitive therapy and research: A question of questions. *Cognitive therapy and Research*, *1*(1), 5-16.

Mahoney, M. J. (1977b). Reflections on the cognitive-learning trend in psychotherapy. *American Psychologist, 32*(1), 5–13. https://doi.org/10.1037/0003-066X.32.1.5

Mahoney, M. J., & Kazdin, A. E. (1979). Cognitive behavior modification: Misconceptions and premature evacuation. *Psychological Bulletin, 86*(5), 1044–1049 https://doi.org/10.1037/0033-2909.86.5.1044

Manne, K. (2020). *Entitled: How male privilege hurts women*. Crown.

Marková, I. S., & Berrios, G. E. (2012). Epistemology of psychiatry. *Psychopathology*, *45*(4), 220–227. https://doi.org/10.1159/000331599

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. D. Reidel.

McDonough, M. R., Johnson, A. K., & Waters, B. J. (2017). Reducing attention maintained bizarre speech in an adult woman. *Behavior Analysis: Research and Practice, 17*(2), 197-203. http://dx.doi.org/10.1037/bar0000059

McDowell, J. (1994). The content of perceptual experience. *The Philosophical Quarterly 44*(175), 190–205. https://doi.org/10.2307/2219740

McDowell, J, (2004). Naturalism in the Philosophy of Mind. In M. Caro, & D. Macarthur (Eds.), *Naturalism in question* (pp. 91-105). Harvard University Press.

McDowell, J. (1996). *Mind and world*. Harvard University Press.

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137-156). Springer.

McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, *18*(2), 259-281. https://doi.org/10.1080/13869795.2015.1032331

McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, *199*(1), 1039-1063. https://doi.org/10.1007/s11229-020-02760-7

McKay, R. (2012). Delusional inference. *Mind & Language*, *27*(3), 330-355. https://doi.org/10.1111/j.1468-0017.2012.01447.x

McLaughlin, B. P. (2008). The rise and fall of British emergentism. In M. A. Bedau & P. Humphreys (Eds.), *Emergence: Contemporary readings in philosophy and science* (pp. 19–59). MIT Press. (Original work published 1992) https://doi.org/10.7551/mitpress/9780262026215.003.0003

McLeod, H. J. (2009). ACT and CBT for psychosis: Comparisons and contrasts. In J. T. Blackledge, J. Ciarrochi, & F. P. Deane (Eds.), *Acceptance and commitment therapy: Contemporary theory, research and practice* (pp. 263–279). Australian Academic Press.

Mehl, S., Schlier, B., & Lincoln, T. M. (2018). Does CBT for psychosis have an impact on delusions by improving reasoning biases and negative self-schemas? *Zeitschrift für Psychologie,226*(3), 152-163. https://doi.org/10.1027/2151-2604/a000335

Mehl, S., Werner, D., & Lincoln, T. M. (2015). Does Cognitive Behavior Therapy for psychosis (CBTp) show a sustainable effect on delusions? A meta-analysis. *Frontiers in psychology*, *6*, 1450. https://doi.org/10.3389/fpsyg.2015.01450

Mehl, S., Werner, D., & Lincoln, T. M. (2019). "Does Cognitive Behavior Therapy for psychosis (CBTp) show a sustainable effect on delusions? A meta-analysis": Corrigendum. *Frontiers in Psychology, 10,* Article 1868. https://doi.org/10.3389/fpsyg.2019.01868

Meichenbaum, D. (1977). Cognitive-Behavior Modification: An Integrative Approach. Plenum. http://dx.doi.org/10.1007/978-1-4757-9739-8

Menary, R. (2010). Introduction to the special issue on 4E cognition. *Phenomenology and the Cognitive Sciences, 9*(4), 459–463. https://doi.org/10.1007/s11097-010-9187-6

Middleton, H., & Moncrieff, J. (2019). Critical psychiatry: a brief overview. *BJPsych Advances*, *25*(1), 47-54. https://doi.org/10.1192/bja.2018.38

Miller-Tate A. J. (2019). Contributory injustice in psychiatry. *Journal of medical ethics*, *45*(2), 97–100. https://doi.org/10.1136/medethics-2018-104761

Miyazono, K., & Bortolotti, L. (2014). The causal role argument against doxasticism about delusions. *Avant: Trends in Interdisciplinary Studies*, *5*(3), 30-50. https://doi.org/10.26913/50302014.0112.0003

Molnar, G. (2003). *Powers: A study in metaphysics* (S. Mumford, Ed.). Clarendon Press.

Moncrieff, J. (2015a). Antipsychotic maintenance treatment: time to rethink? *PLoS medicine*, *12*(8), e1001861. https://doi.org/10.1371/journal.pmed.1001861

Moncrieff, J. (2015b). The myths and realities of drug treatment for mental disorders. *The Behavior Therapist, 38*, 214-218. https://www.madinamerica.com/wp-content/uploads/2015/11/Behavior-Therapist-Oct-2015.pdf

Monestès, J. L., Villatte, M., Stewart, I., & Loas, G. (2014). Rule-based insensitivity and delusion maintenance in schizophrenia. *The Psychological Record*, *64*(2), 329-338. https://doi.org/10.1007/s40732-014-0029-8

Montaño-Fidalgo, M., Martínez-Sánchez, H., Froján-Parga, M. X., & Calero-Elvira, A. (2013). The role of verbal behavior in the analysis of the therapeutic process. *Conductual, 1*(2), 62-72. https://conductual.com/articulos/The%20role%20of%20verbal%20behavior%20in%20the%20analysis%20of%20the%20therapeutic%20process.pdf

Moore, G. E. (1922). *Principia ethica*. Cambridge University Press (Original work published 1903)

Moore, J. (1981). On mentalism, methodological behaviorism, and radical behaviorism. *Behaviorism, 9*(1), 55–77. https://www.jstor.org/stable/27758972

Moore, J. (2001). On distinguishing methodological from radical behaviorism. *European Journal of Behavior Analysis*, *2*(2), 221–244. https://doi.org/10.1080/15021149.2001.11434196

Moore, J. (2008). *Conceptual foundations of radical behaviorism*. Sloan Publishing.

Moore, J. (2009). Why the radical behaviorist conception of private events is interesting, relevant, and important. *Behavior and Philosophy*, 21–37. https://www.jstor.org/stable/41472420

Moore, J. (2013). Methodological behaviorism from the standpoint of a radical behaviorist. *The Behavior Analyst*, *36*(2), 197–208. https://doi.org/10.1007/BF03392306

Morrison, A., Renton, J., Dunn, H., Williams, S., & Bentall, R. (2004). *Cognitive therapy for psychosis: A formulation-based approach*. Routledge.

Murphy, D. (2009). Psychiatry and the concept of disease as pathology. In M. Broome & L. Bortolotti (Eds.), *Psychiatry as cognitive neuroscience: philosophical perspectives*, (pp. 103–117). Oxford University Press.

Murphy, D. (2012). The folk epistemology of delusions. *Neuroethics*, *5*(1), 19–22. https://doi.org/10.1007/s12152-011-9125-5

Murphy, D. (2013). The medical model and the philosophy of science. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 966–986). Oxford University Press.

Murphy, D. (2020). Philosophy of psychiatry. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/psychiatry/

National Institute of Mental Health (2019 November 19). *NIMH Research Domain Criteria (RDoC) Initiative: Development and Environment in RDoC Workshop -Proceedings and Thematic Summary*. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/resources/nimh-research-domain-criteria-rdoc-initiative-development-and-environment-in-rdoc-workshop-proceedings-and-thematic-summary

Naeem, F., Farooq, S., & Kingdon, D. (2014). Cognitive behavioural therapy (brief versus standard duration) for schizophrenia. *The Cochrane database of systematic reviews*, (4), CD010646. https://doi.org/10.1002/14651858.CD010646.pub2

Naeem, F., Khoury, B., Munshi, T., Ayub, M., Lecomte, T., Kingdon, D., & Farooq, S. (2016). Brief cognitive behavioral therapy for psychosis (CBTp) for schizophrenia: Literature review and meta-analysis. *International Journal of Cognitive Therapy*, *9*(1), 73–86.

National Institute for Health and Care Excellence (2009). Schizophrenia: Core Interventions in the Treatment and Management of Schizophrenia in Primary and Secondary Care

(Update). *NICE Clinical Guidelines, No. 82*. https://www.ncbi.nlm.nih.gov/books/NBK11681/

National Institute for Health and Care Excellence (2014). Psychosis and schizophrenia in adults: prevention and management. *NICE Clinical Guidelines, No. 178*. https://www.ncbi.nlm.nih.gov/books/NBK555203/

National Institute of Mental Health (2019 November 19). *NIMH Research Domain Criteria (RDoC) Initiative: Development and Environment in RDoC Workshop — Proceedings and Thematic Summary*. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/resources/nimh-research-domain-criteria-rdoc-initiative-development-and-environment-in-rdoc-workshop-proceedings-and-thematic-summary

Newen, A., De Bruin, L., & Gallagher, S. (Eds.). (2018). *The Oxford handbook of 4E cognition*. Oxford University Press.

Nielsen, K. (2021a). Comparing two enactive perspectives on mental disorder. *Philosophy, Psychiatry, & Psychology*, *28*(3), 175–185. 10.1353/ppp.2021.0028

Nielsen, K. (2021b). Comparing Two Enactive Perspectives. *Philosophy, Psychiatry, & Psychology*, *28*(3), 197–200. 10.1353/ppp.2021.0031

Nielsen, K., & Ward, T. (2018). Towards a new conceptual framework for psychopathology: Embodiment, enactivism, and embedment. *Theory & Psychology, 28*(6), 800–822. https://doi.org/10.1177/0959354318808394

Nielsen, K., & Ward, T. (2020). Mental disorder as both natural and normative: Developing the normative dimension of the 3e conceptual framework for psychopathology. *Journal of Theoretical and Philosophical Psychology, 40*(2), 107–123. https://doi.org/10.1037/teo0000118

Nöe, A. (2001). Experience and the active mind. *Synthese, 129*, 41–60. https://doi.org/10.1023/A:1012695023768

Nöe, A. (2004). *Action in perception*. The MIT Press.

Nottelmann, N. (Ed.). (2013). *New Essays on Belief: Constitution, Content and Structure*. Springer.

Nydegger, R. V. (1972). The elimination of hallucinatory and delusional behavior by verbal conditioning and assertive training: A case study. *Journal of Behavior Therapy and Experimental Psychiatry*, *3*(3), 225–227. https://doi.org/10.1016/0005-7916(72)90080-8

O'Connor, T. (1994). Emergent properties. *American Philosophical Quarterly*, *31*(2), 91–104. https://www.jstor.org/stable/20014490

O'Connor, T. (2020). Emergent Properties. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/properties-emergent

O'Connor, T., & Wong, H. Y. (2005). The metaphysics of emergence. *Noûs*, *39*(4), 658-678. https://www.jstor.org/stable/3506115

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939-1031. https://doi.org/10.1017/S0140525X01000115

Öst L. G. (2014). The efficacy of Acceptance and Commitment Therapy: an updated systematic review and meta-analysis. *Behaviour research and therapy*, *61*, 105–121. https://doi.org/10.1016/j.brat.2014.07.018

Oswald, I. (1962). Induction of illusory and hallucinatory voices with considerations of behaviour therapy. *Journal of Mental Science*, *108*(453), 196-212. https://doi.org/10.1192/bjp.108.453.196

Pankey, J., & Hayes, S. C. (2003). Acceptance and commitment therapy for psychosis. *International Journal of Psychology & Psychological Therapy, 3*(2), 311–328.

Parent, T. (2013). In the mental fiction, mental fictionalism is fictitious. *The Monist*, *96*(4), 605-621. https://doi.org/10.5840/monist201396428

Pascual-Verdú, R., & Trujillo-Sánchez, C. T. (2018). Estudio de la relación entre las verbalizaciones motivadoras y el seguimiento de instrucciones en la terapia psicológica A study of the relation between motivational utterances and instruction compliance. *Revista Clínica Contemporánea*, *9*(e14), 1-11.

Pascual-Verdú, R., Trujillo, C., Gálvez, E., Andrés-López, N., Castaño-Hurtado, R., & Froxán-Parga, M. X. (2019). Sistema ACOVEO: una propuesta funcional para el análisis de la interacción verbal en terapia. *Conductual, 7*(2), 69-82.

Pedersen, N. J. L. L., & Wright, C. (2018). Pluralist Theories of Truth. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/truth-pluralist/

Peele, S. (2015). Why neurobiological models can't contain mental disorder and addiction. *The Behavior Therapist*, *38*(7), 218-222.

Pereira, G. L., Hernández, A. R., de Pascual-Verdú, R., & Froxán-Parga, M. X. (2019). Los procesos de condicionamiento clásico en la interacción verbal terapéutica. *Revista Mexicana de Análisis de la Conducta*, *45*(1), 90-110. http://dx.doi.org/10.5514/rmac.v45.i1.70870

Pérez-Álvarez, M. (1996). *La psicoterapia desde el punto de vista conductista*. Biblioteca nueva.

Pérez-Álvarez, M. (2004). *Contingencia y drama: la psicología según el conductismo*. Minerva Ediciones.

Pérez-Álvarez, M. (2011). *El mito del cerebro creador: cuerpo, conducta y cultura*. Alianza.

Pérez-Álvarez, M. (2012). Third-generation therapies: Achievements and challenges. *International Journal of Clinical and Health Psychology*, *12*(2), 291-310. https://www.redalyc.org/articulo.oa?id=33723643008

Pérez-Álvarez, M., & García-Montes, J. M. (2007). The Charcot effect: The invention of mental illnesses. *Journal of Constructivist Psychology*, *20*(4), 309-336. https://doi.org/10.1080/10720530701503843

Pérez-Navarro, E. (2021). The way things go: moral relativism and suspension of judgment. *Philosophical Studies*, 1-16. https://doi.org/10.1007/s11098-021-01650-z

Pérez-Navarro, E., Fernández-Castro, V., González de Prado-Salas, J., & Heras-Escribano, M. (2019). Not expressivist enough: Normative disagreement about belief attribution. *Res Philosophica*, *96*(4), 409-430. https://doi.org/10.11612/resphil.1794

Pescosolido, B., Martin, J., Long, J., Medina, T., Phelan, J., & Link, B. (2010). "A disease like any other"? A decade of change in public reactions to schizophrenia, depression, and alcohol dependence. *American Journal of Psychiatry*, *167*(11), 1321-1330. doi:10.1176/appi.ajp.2010.09121743

Peterson, S. M. & Neef, A. N. (2019). Functional Behavior Assessment. In J. O. Cooper, T. E. Heron, & W. L. Heward (Eds.), *Applied Behavior Analysis* (pp. 678-706). New Jersey: Pearson Education.

Pilgrim, C. (2019). Equivalence-based instruction. In J. O. Cooper, T. E. Heron, & W. L. Heward (Eds.), *Applied Behavior Analysis* (pp. 442-496). New Jersey: Pearson Education.

Pilgrim, D. (2015). The biopsychosocial model in health research: Its strengths and limitations for critical realists. *Journal of Critical Realism*, *14*(2), 164-180. https://doi.org/10.1179/1572513814Y.0000000007

Pinedo-García, M. (2014). ¡No es un algo, pero tampoco es una nada! Mente y normatividad. *Análisis, 1*(1), 121-160. https://doi.org/10.26754/ojs_arif/a.rif.20141980

Pinedo-García, M. (2020). Ecological psychology and enactivism: A normative way out from ontological dilemmas. *Frontiers in Psychology*, *11*, 1637. https://doi.org/10.3389/fpsyg.2020.01637

Pinedo-García, M., & Noble, J. (2008). Beyond persons: extending the personal/subpersonal distinction to non-rational animals and artificial agents. *Biology & Philosophy*, *23*(1), 87–100. https://doi.org/10.1007/s10539-007-9077-7

Pitt, L., Kilbride, M., Nothard, S., Welford, M., & Morrison, A. P. (2007). Researching recovery from psychosis: a user-led project. *Psychiatric Bulletin*, *31*(2), 55–60. https://doi.org/10.1192/pb.bp.105.008532

Place, U. T. (1956). Is consciousness a brain process? *British journal of psychology*, *47*(1), 44–50. https://doi.org/10.1111/j.2044-8295.1956.tb00560.x

Place, U. T. (1988). Thirty years on—Is consciousness still a brain process? *Australasian Journal of Philosophy*, *66*(2), 208–219. https://doi.org/10.1080/00048408812343291

Price, H. (2004). Naturalism without representationalism. In M. Caro, & D. Macarthur (Eds.), *Naturalism in question* (pp. 71–88). Harvard University Press.

Price, H. (2011). *Naturalism without mirrors*. Oxford University Press.

Price, H., Blackburn, S., Brandom, R., Horwich, P., & Williams, M. (2013). *Expressivism, pragmatism and representationalism*. Cambridge University Press.

Poland, J., & Von Eckardt, B. (2013). Mapping the domain of mental illness. In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 735–752). Oxford University Press.

Porcher, J. E. (2016). Delusion as a folk psychological kind. *Filosofia Unisinos/Unisinos Journal of Philosophy*, *17*(2), 212–226.

Porcher, J. E. (2018). The Doxastic Status of Delusion and the Limits of Folk Psychology. In *Schizophrenia and Common Sense* (pp. 175–190). Springer, Cham.

Porcher, J. E. (2019). Double bookkeeping and doxasticism about delusion. *Philosophy, Psychiatry, & Psychology*, *26*(2), 111–119.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112. https://doi.org/10.1016/j.jsp.2018.02.003

Putnam, H. (1975). The nature of mental states. In H. Putnam (Ed.) *Mind, language and reality: Philosophical Papers Volume 2* (pp. 429–440). Cambridge University Press. (Original work published 1967).

Qaseem, A., Barry, M. J., & Kansagara, D. (2016). Nonpharmacologic versus pharmacologic treatment of adult patients with major depressive disorder: a clinical practice

guideline from the American College of Physicians. *Annals of internal medicine*, *164*(5), 350-359.

Rachlin, H. (1977a). A review of MJ Mahoney's Cognition and Behavior Modification. *Journal of Applied Behavior Analysis*, *10*(2), 369-374. https://doi.org/10.1901/jaba.1977.10-369

Rachlin, H. (1977b). Reinforcing and punishing thoughts. *Behavior Therapy, 8*(4), 659–665. https://doi.org/10.1016/S0005-7894(77)80196-2

Rachman, S. (1958). Objective psychotherapy: some theoretical considerations. *South African Medical Journal*, *32*(1), 19-21. https://journals.co.za/doi/pdf/10.10520/AJA20785135_44034

Rachman, S. (1959). The treatment of anxiety and phobic reactions by systematic desensitization psychotherapy. *The Journal of Abnormal and Social Psychology, 58*(2), 259–263. https://doi.org/10.1037/h0040150

Radden, J. (2010). *On delusion*. Routledge.

Ramberg, B. (2000). Post-ontological philosophy of mind: Rorty versus Davidson. En R. Brandom (Ed.) *Rorty and His Critics* (pp. 351-369). Blackwell Publishers.

Ramsey, W. (2020). Eliminative materialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/materialism-eliminative/

Ramsey, W., Stich, S., & Garan, J. (1990). Connectionism, eliminativism, and the future of folk psychology. In D. J. Cole, J. H. Fetzer & T. L. Rankin (Eds.) *Philosophy, Mind, and Cognitive Inquiry* (pp. 117-144). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-1882-5_5

Rehfeldt, R. A., & Chambers, M. R. (2003). Functional analysis and treatment of verbal perseverations displayed by an adult with autism. Journal of Applied Behavior Analysis, 36(2), 259–261. https://doi.org/10.1901/jaba.2003.36-259

Reimer, M. (2010). Only a philosopher or a madman: Impractical delusions in philosophy and psychiatry. *Philosophy, Psychiatry, & Psychology, 17*(4), 315–328. https://doi.org/10.1353/ppp.2010.0028

Reimer, M. (2012). Davidsonian holism in recent philosophy of psychiatry. In G. Preyer (Ed.), *Donald Davidson on Truth, Meaning, and the Mental* (pp. 249-269). Oxford University Press.

Ritunnano, R. (2022). Overcoming Hermeneutical Injustice in Mental Health: A Role for Critical Phenomenology. *Journal of the British Society for Phenomenology*, 1-18. https://doi.org/10.1080/00071773.2022.2031234

Roberts, T., Krueger, J., & Glackin, S. (2019). Psychiatry beyond the brain: Externalism, mental health, and autistic spectrum disorder. *Philosophy, Psychiatry, & Psychology*, *26*(3), 51-68. https://doi.org/10.1353/ppp.2019.0030

Roessler, B. (2015). Autonomy, self-knowledge, and oppression. In M. A. L. Oshana (Ed.), *Personal Autonomy and Social Oppression: Philosophical Perspectives* (pp. 68-84). Routledge.

Röhricht, F., Gallagher, S., Geuter, U., & Hutto, D. D. (2014). Embodied cognition and body psychotherapy: The construction of new therapeutic environments. *Sensoria: A Journal of Mind, Brain & Culture*, *10*(1), 11-20. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.948.9895&rep=rep1&type=pdf

Rorty, R. (1965). Mind-body identity, privacy, and categories. *The Review of Metaphysics*, 24-54. https://www.jstor.org/stable/20124096

Rorty, R. (1970). In Defense of Eliminative Materialism. *The Review of Metaphysics*, *24*(1), 112-121. https://www.jstor.org/stable/20125726

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton University Press.

Rose, D., Buckwalter, W., & Turri, J. (2014). When words speak louder than actions: Delusion, belief, and the power of assertion. *Australasian Journal of Philosophy*, *92*(4), 683-700. http://dx.doi.org/10.2139/ssrn.3649259

Rosenfarb, I. S. (2013). A functional analysis of schizophrenia. *The Psychological Record, 63*(4), 929–946. https://doi.org/10.11133/j.tpr.2013.63.4.013

Ruiz-Sancho, E., Froján-Parga, M. X., & Galván-Domínguez, N. (2015). Verbal interaction patterns in the clinical context: a model of how people change in therapy. *Psicothema*, *27*(2), 99–107. https://doi.org/10.7334/psicothema2014.119

Rummel-Kluge, C., Komossa, K., Schwarz, S., Hunger, H., Schmid, F., Lobos, C. A., Kissling, W., Davis, J. M., & Leucht, S. (2010). Head-to-head comparisons of metabolic side effects of second generation antipsychotics in the treatment of schizophrenia: a systematic review and meta-analysis. *Schizophrenia research*, *123*(2-3), 225–233. https://doi.org/10.1016/j.schres.2010.07.012

Russell, B. (1913). *Theory of Knowledge, The 1913 Manuscript*. (E. R. Eames & K. Blackwell, Eds.). Routledge.

Rutherford, A. (2003). Skinner Boxes for Psychotics: Operant Conditioning at Metropolitan State Hospital. *The Behavior Analyst, 26*(2), 267–279. https://doi.org/10.1007/BF03392081

Ryle, G. (2009). *The concept of mind* (J. Tanney, Ed.). Routledge. (Original work published 1949).

Salzinger, K., Portnoy, S., & Feldman, R. S. (1964). Experimental manipulation of continuous speech in schizophrenic patients. *The Journal of Abnormal and Social Psychology, 68*(5), 508–516. https://doi.org/10.1037/h0040445

Sánchez-Curry, D. (2020). Interpretivism and norms. *Philosophical Studies*, *177*(4), 905-930. https://doi.org/10.1007/s11098-018-1212-6

Sass, L. A. (1994). *The paradoxes of delusion: Wittgenstein, Schreber and the schizophrenic mind*. Cornell University Press.

Sass, L. A. (2004). Some reflections on the (analytic) philosophical approach to delusion. *Philosophy, Psychiatry, & Psychology*, *11*(1), 71-80.

Sass, L. A. (2014). Delusions and double book-keeping. In T. Fuchs, T. Breyer, & C. Mundt (Eds.), *Karl Jaspers' philosophy and psychopathology* (pp. 125–148). Springer.

Sass, L. A., & Pienkos, E. (2013). Delusion: The phenomenological approach. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 632–657). Oxford University Press.

Savitt, S. (1975). Rorty's disappearance theory. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 28*(6), 433-436. https://www.jstor.org/stable/4319001

Schaler, J. A. (Ed.). (2004). *Szasz under fire: A psychiatric abolitionist faces his critics*. Open Court Publishing.

Scheff, T. J. (1999). *On being mentally ill: A sociological theory*. Aldine de Gruyter (Original work published 1966).

Scheff, T. J. (1974). The labelling theory of mental illness. *American sociological review*, 444-452. https://doi.org/10.2307/2094300

Schnaitter, R. (1984). Skinner on the" mental" and the" physical". *Behaviorism*, *12*(1), 1-14. https://www.jstor.org/stable/27759032

Schneider, S. M., & Morris, E. K. (1987). A history of the term radical behaviorism: From Watson to Skinner. *The Behavior Analyst, 10*(1), 27–39. https://doi.org/10.1007/BF03392404

Schomerus, G., Schwahn, C., Holzinger, A., Corrigan, P. W., Grabe, H. J., Carta, M. G., & Angermeyer, M. C. (2012). Evolution of public attitudes about mental illness: A systematic review and meta-analysis. *Acta Psychiatrica Scandinavica*, *125*(6), 440-452. https://doi.org/10.1111/j.1600-0447.2012.01826.x

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, *36*(2), 249-275. https://doi.org/10.1111/1468-0068.00370

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, *117*(2), 245-273.

Schwitzgebel, E. (2012). Mad belief? *Neuroethics*, *5*(1), 13-17. https://doi.org/10.1007/s12152-011-9127-3

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelman (Ed.) *New essays on belief* (pp. 75-99). Palgrave Macmillan. https://doi.org/10.1057/9781137026521_5

Schwitzgebel, E. (2021). The Pragmatic Metaphysics of Belief. In C. Borgoni, D. Kindermann, & A. Onofri (Eds.), *The Fragmented Mind*. Oxford University Press. https://doi.org/10.1093/oso/9780198850670.003.0015

Scull, A., & Schulkin, J. (2009). Psychobiology, Psychiatry, and Psychoanalysis: The Intersecting Careers of Adolf Meyer, Phyllis Greenacre, and Curt Richter. *Medical History, 53*(1), 5-36. https://doi.org/10.1017/S002572730000329X

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science, 1*(19), 253-329.

Sellars, W. (1963). Philosophy and the scientific image of man. En W. Sellars (Ed.) *Empiricism and the Philosophy of Mind* (pp. 1-40). Routledge & Kegan Paul Ltd.

Shah, P., & Mountain, D. (2007). The medical model is dead-long live the medical model. *The British journal of psychiatry: the journal of mental science*, *191*, 375–377. https://doi.org/10.1192/bjp.bp.107.037242

Shapiro, L. (2007). *The correspondence between Princess Elisabeth of Bohemia and Rene Descartes*. University of Chicago Press.

Shapiro, L. (2021). Elisabeth, Princess of Bohemia. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/elisabeth-bohemia/

Shawyer, F., Farhall, J., Mackinnon, A., Trauer, T., Sims, E., Ratcliff, K., Larner, C., Thomas, N., Castle, D., Mullen, P., & Copolov, D. (2012). A randomised controlled trial of acceptance-based cognitive behavioural therapy for command hallucinations in psychotic disorders. *Behaviour research and therapy*, *50*(2), 110–121. https://doi.org/10.1016/j.brat.2011.11.007

Shawyer, F., Farhall, J., Thomas, N., Hayes, S. C., Gallop, R., Copolov, D., & Castle, D. J. (2017). Acceptance and commitment therapy for psychosis: randomised controlled trial. *The British journal of psychiatry: the journal of mental science*, *210*(2), 140–148. https://doi.org/10.1192/bjp.bp.116.182865

Sherman, J. A. (1965). Use of reinforcement and imitation to reinstate verbal behavior in mute psychotics. *Journal of Abnormal Psychology, 70*(3), 155–164. https://doi.org/10.1037/h0022148

Shorter, E. (1998). *A History of Psychiatry: From the Era of the Asylum to the Age of Prozac*. John Wiley & Sons.

Sidman, M. (2009). Equivalence relations and behavior: An introductory tutorial. *The Analysis of verbal behavior*, *25*(1), 5-17. https://doi.org/10.1007/BF03393066

Silberstein, M., & McGeever, J. (1999). The search for ontological emergence. *The philosophical quarterly*, *49*(195), 201-214. https://www.jstor.org/stable/2660261

Singer, J. (1999). Why can't you be normal for once in your life? From a problem with no name to the emergence of a new category of difference. In M. Corker & S. French. *Disability discourse* (pp. 59-70). Open University Press

Sitko, K., Bewick, B. M., Owens, D., & Masterson, C. (2020). Meta-analysis and meta-regression of cognitive behavioral therapy for psychosis (CBTp) across time: the effectiveness of CBTp has improved for delusions. *Schizophrenia Bulletin Open*, *1*(1), sgaa023. https://doi.org/10.1093/schizbullopen/sgaa023

Skinner, B. F. (1936). The verbal summator and a method for the study of latent speech. *The Journal of Psychology: Interdisciplinary and Applied, 2*, 71–107. https://doi.org/10.1080/00223980.1936.9917445

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, *52*(5), 270–277. https://doi.org/10.1037/h0062535

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review, 57*(4), 193–216. https://doi.org/10.1037/h0054367

Skinner, B. F. (1953). *Science and Human Behavior.* MacMillan.

Skinner, B. F. (1957). *Verbal behavior.* Appleton-Century-Crofts.

Skinner, B. F. (1963). Behaviorism at fifty. *Science, 140*(3570), 951–958. https://doi.org/10.1126/science.140.3570.951

Skinner, B. F. (1969), *Contingencies of reinforcement*. Appleton-Century-Crofts

Skinner, B. F. (1971). *Beyond Freedom and Dignity*. Vintage Book.

Skinner, B. F. (1974). *About behaviorism.* Knopf.

Skinner, B. F. (1977). Why I am not a cognitive psychologist. *Behaviorism*, *5*(2), 1-10. https://www.jstor.org/stable/27758892

Skinner, B. F. (1981). Selection by consequences. *Science*, *213*(4507), 501–504. https://doi.org/10.1126/science.7244649

Skinner, B. F. (1984). Coming to terms with private events. *Behavioral and Brain Sciences*, *7*(4), 572–581. https://doi.org/10.1017/S0140525X00027400

Skinner, B. F. (1990). Can psychology be a science of mind? *American Psychologist, 45*(11), 1206–1210. https://doi.org/10.1037/0003-066X.45.11.1206

Slade, P. D. (1972). The effects of systematic desensitization on auditory hallucinations. *Behaviour Research and Therapy, 10*(1), 85–91. https://doi.org/10.1016/0005-7967(72)90013-7

Smart, J. J. (1959). Sensations and brain processes. The Philosophical Review, *68(2)*, 141–156. https://doi.org/10.2307/2182164

Smart, J. J. (2017). The Mind/Brain Identity Theory. In E. N. Zalta (Ed.*), The Stanford encyclopedia of philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/mind-identity/

Sneddon, A. (2002). Towards externalist psychopathology. *Philosophical Psychology*, *15*(3), 297–316. https://doi.org/10.1080/0951508021000006102

Spitzer, R. L., Endicott, J., & Franchi, J. A. M. (2018). Medical and mental disorder: Proposed definition and criteria. *Annales Médico-psychologiques, revue psychiatrique, 176*(7), 656–665. https://doi.org/10.1016/j.amp.2018.07.004 (Original work published 1978).

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: rationale and reliability. *Archives of general psychiatry*, *35*(6), 773–782. https://doi.org/10.1001/archpsyc.1978.01770300115013

Sprevak, M. (2011). Neural sufficiency, reductionism, and cognitive neuropsychiatry. *Philosophy, Psychiatry, & Psychology*, *18*(4), 339–344. doi:10.1353/ppp.2011.0057

Srinivasan, A. (2015). Normativity without cartesian privilege. Philosophical Issues, 25, 273–299.

Srinivasan, A. (2020). Radical externalism. *Philosophical Review*, *129*(3), 395–431. https://doi.org/10.1215/00318108-8311261

Srinivasan, A. (2021). *The Right to Sex*. Bloomsbury Publishing.

Staddon J. (2021) Theoretical Behaviorism. In D. Zilio, & K. Carrara (Eds.) *Contemporary Behaviorisms in Debate* (pp. 79-95). Springer, Cham. https://doi.org/10.1007/978-3-030-77395-3_7

Stephens, G. L., & Graham, G. (2006). The delusional stance. In M. C. Chung, K. W. M. Fulford, & G. Graham (eds.) *Reconceiving schizophrenia* (193–216). OUP.

Stewart, C., Stewart, I., & Hughes, S. (2016). A contextual behavioral approach to the study of (persecutory) delusions. *Journal of Contextual Behavioral Science*, *5*(4), 235–246. https://doi.org/10.1016/j.jcbs.2016.09.002

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief*. the MIT press.

Stoljar, D. (2021). Physicalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/physicalism/

Stone, T., & Young, A. W. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, *12*(3-4), 327-364. https://doi.org/10.1111/j.1468-0017.1997.tb00077.x

Sturmey, P. (Ed.). (2020). *Functional analysis in clinical treatment*. Academic Press.

Sullivan-Bissett, E., Bortolotti, L., Broome, M., & Mameli, M. (2016). Moral and legal implications of the continuity between delusional and non-delusional beliefs. In G. Keil, L. Keuck, & R. Hauswald (Eds.), *Vagueness in psychiatry*. Oxford University Press.

Szasz, T. S. (1960). The myth of mental illness. *American psychologist*, *15*(2), 113-118. https://doi.org/10.1037/h0046535

Szasz, T. (1974). *The myth of mental illness: Foundations of a theory of personal conduct*. Harper & Row. (Original work published 1961).

Szasz T. S. (1976). Schizophrenia: the sacred symbol of psychiatry. *The British journal of psychiatry: the journal of mental science*, *129*, 308–316. https://doi.org/10.1192/bjp.129.4.308

Szasz, T. (1991). Against behaviorism. A review of BF Skinner's *About behaviorism*. *Psychological Notes 5*, 1-2. (Original work published 1974). http://libertarian.co.uk/sites/default/lanotepdf/psycn005.pdf

Szasz, T. (2003). *Pharmacracy: Medicine and politics in America*. Syracuse University Press.

Szasz, T. (2009). *Antipsychiatry: quackery squared*. Syracuse University Press.

Szasz, T. (2011). The myth of mental illness: 50 years later. *The Psychiatrist*, *35*(5), 179-182. https://doi.org/10.1192/pb.bp.110.031310

Tabb, K. (2015). Psychiatric progress and the assumption of diagnostic discrimination. *Philosophy of Science*, *82*(5), 1047-1058. https://doi.org/10.1086/683439

Tabb, K. (2017). Philosophy of psychiatry after diagnostic kinds. *Synthese*, *196*(6), 2177-2195. https://doi.org/10.1007/s11229-017-1659-6

Tabb, K. (2020). Should Psychiatry Be Precise? Reduction, Big Data, and Nosological Revision in Mental Health Research. In K. Kendler, J. Parnas, & P. Zachar (Eds.), *Levels of Analysis in Psychopathology: Cross-Disciplinary Perspectives* (pp. 308-334). Cambridge University Press. https://doi.org/10.1017/9781108750349.028

Tandon, R. (2013). Schizophrenia and other psychotic disorders in DSM-5: Clinical implications of revisions from DSM-IV. *Clinical schizophrenia & related psychoses*, *7*(1), 16-19.

Tanney, J. (2009). Rethinking Ryle: a critical discussion of The Concept of Mind. In G. Ryle (J. Tanney, Ed.), *The concept of mind* (pp. ix – lvii). Routledge.

Thornton, T. (2007). *Essential philosophy of psychiatry*. Oxford University Press.

Thornton, T. (2013). Clinical judgment, tacit knowledge, and recognition in psychiatric diagnosis. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini, & T. Thornton (Eds.), *The Oxford handbook of philosophy and psychiatry* (pp. 1047–1062). Oxford University Press.

Thornton, T. (2014). Values based practice and authoritarianism. In M. Loughlin (ed.) *Debates in Values-based Practice: arguments for and against* (pp. 50-61). Cambridge University Press.

Tolman, E. C. (1928). Purposive behavior. *Psychological Review, 35*(6), 524–530. https://doi.org/10.1037/h0070770

Tonarelli, S. B., Pasillas, R., Alvarado, L., Dwivedi, A., & Cancellare, A. (2016). Acceptance and commitment therapy compared to treatment as usual in psychosis: A systematic review and meta-analysis. *Journal of Psychiatry*, *19*(3), 2-5. http://dx.doi.org/10.4172/2378-5756.1000366

Tonneau, F. (2001). Equivalence relations: A critical analysis. *European Journal of Behavior Analysis*, *2*(1), 1-33. https://doi.org/10.1080/15021149.2001.11434165

Travis, R., & Sturmey, P. (2008). A Review of Behavioral Interventions for Psychotic Verbal Behavior in People With Intellectual Disabilities. *Journal of Mental Health Research in Intellectual Disabilities*, *1*(1), 19-33. https://doi.org/10.1080/19315860701686963

Travis, R., & Sturmey, P. (2010). Functional analysis and treatment of the delusional statements of a man with multiple disabilities: A four-year follow-up. Journal of Applied Behavior Analysis, 43(4), 745–749. https://doi.org/10.1901/jaba.2010.43-745

Tumulty, M. (2011). Delusions and dispositionalism about belief. *Mind & language*, *26*(5), 596-628. https://doi.org/10.1111/j.1468-0017.2011.01432.x

Tumulty, M. (2012). Delusions and not-quite-beliefs. *Neuroethics*, *5*(1), 29-37. https://doi.org/10.1007/s12152-011-9126-4

Turner, D. T., Burger, S., Smit, F., Valmaggia, L. R., & van der Gaag, M. (2020). What constitutes sufficient evidence for case formulation–driven CBT for psychosis? Cumulative meta-analysis of the effect on hallucinations and delusions. *Schizophrenia bulletin*, *46*(5), 1072-1085. https://doi.org/10.1093/schbul/sbaa045

Turner, D. T., van der Gaag, M., Karyotaki, E., & Cuijpers, P. (2014). Psychological interventions for psychosis: a meta-analysis of comparative outcome studies. *American Journal of Psychiatry*, *171*(5), 523-538. https://doi.org/10.1176/appi.ajp.2013.13081159

Udachina, A., Thewissen, V., Myin-Germeys, I., Fitzpatrick, S., O'kane, A., & Bentall, R. P. (2009). Understanding the relationships between self-esteem, experiential avoidance, and paranoia: structural equation modelling and experience sampling studies. *The Journal of nervous and mental disease*, *197*(9), 661–668. https://doi.org/10.1097/NMD.0b013e3181b3b2ef

Udachina, A., Varese, F., Myin-Germeys, I., & Bentall, R. P. (2014). The role of experiential avoidance in paranoid delusions: an experience sampling study. *The British journal of clinical psychology*, *53*(4), 422–432. https://doi.org/10.1111/bjc.12054

Upthegrove, R. (2018). Delusional beliefs in the clinical context. In L. Bortolotti (Ed.), *Delusions in context* (pp. 1-34). Palgrave Macmillan, Cham.

Van der Gaag, M., Valmaggia, L. R., & Smit, F. (2014). The effects of individually tailored formulation-based cognitive behavioural therapy in auditory hallucinations and delusions: a meta-analysis. *Schizophrenia research*, *156*(1), 30-37. https://doi.org/10.1016/j.schres.2014.03.016

Van Oudenhove, L., & Cuypers, S. (2014). The relevance of the philosophical 'mind–body problem'for the status of psychosomatic medicine: a conceptual analysis of the biopsychosocial model. *Medicine, Health Care and Philosophy*, *17*(2), 201-213. https://doi.org/10.1007/s11019-013-9521-1

Van Praag, H. M. (1972). Biologic psychiatry in perspective: the dangers of sectarianism in psychiatry. V. Some inferred trends. *Comprehensive psychiatry*, *13*(5), 401-410. https://doi.org/10.1016/0010-440X(72)90081-8

Vandbakk, M., Arntzen, E., Gisnaas, A., Antonsen, V., & Gundhus, T. (2012). Effect of training different classes of verbal behavior to decrease aberrant verbal behavior. *The Analysis of verbal behavior*, *28*(1), 137–144. https://doi.org/10.1007/BF03393115

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind*. MIT Press.

Varela, F. & Thompson, E. (2003). Neural synchrony and the unity of mind: A neurophenomenological perspective. In A. Cleeremans (Ed.), *The Unity of Consciousness*. Oxford University Press.

Varga, S. (2015). *Naturalism, interpretation, and mental disorder*. Oxford University Press.

Varga, S. (2017). Mental disorder between naturalism and normativism. *Philosophy Compass*, *12*(6), e12422. https://doi.org/10.1111/phc3.12422

Vargas, J. S. (1991). BF Skinner—The last few days. *Journal of the Experimental Analysis of Behavior*, *55*(1), 1. https://doi.org/10.1901/jeab.1991.55-1

Veiga-Martínez, C., Pérez-Álvarez, M., & García-Montes, J. M. (2008). Acceptance and commitment therapy applied to treatment of auditory hallucinations. *Clinical Case Studies*, *7*(2), 118-135. https://doi.org/10.1177/1534650107306291

Vilardaga, R., Hayes, S. C., Levin, M. E., & Muto, T. (2009). Creating a strategy for progress: A contextual behavioral science approach. *The Behavior Analyst, 32*(1), 105–133. https://doi.org/10.1007/BF03392178

Villanueva, N. F. (2014). Know thyself: A tale of two theses and two theories. *Teorema: Revista Internacional de Filosofía*, *33*(3), 49-66. https://www.jstor.org/stable/26370102

Villanueva, N. (2018). Expresivismo y semántica. En D. Pérez Chico (coord.), *Cuestiones de la filosofía del lenguaje* (pp. 437-469). Prensas de la Universidad de Zaragoza.

Villanueva, N. (2019). Descripciones y estados mentales. In J. J. Acero (Ed.), *Guía Comares de Wittgenstein* (pp. 145-170). Comares.

Von Bertalanffy, L. (1950). An outline of general system theory. *British Journal for the Philosophy of Science, 1,* 134-165. https://doi.org/10.1093/bjps/I.2.134

Von Bertalanffy, L. (1968). *General system theory*. Geroge Braziller.

Wakefield, J. C. (1992). Disorder as harmful dysfunction: a conceptual critique of DSM-III-R's definition of mental disorder. *Psychological review*, *99*(2), 232-247. https://doi.org/10.1037/0033-295X.99.2.232

Wakefield, J. C. (2007). The concept of mental disorder: diagnostic implications of the harmful dysfunction analysis. *World Psychiatry*, *6*(3), 149-156. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2174594/

Wakefield, S., Roebuck, S., & Boyden, P. (2018). The evidence base of acceptance and commitment therapy (ACT) in psychosis: A systematic review. *Journal of contextual behavioral science*, *10*, 1-13. https://doi.org/10.1016/j.jcbs.2018.07.001

Walter, H. (2013). The third wave of biological psychiatry. *Frontiers in Psychology*, *4*, 582. https://doi.org/10.3389/fpsyg.2013.00582

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158–177. https://doi.org/10.1037/h0074428

Watson, J. B. (1919). *Psychology: From the standpoint of a behaviorist*. JB Lippincott.

Westra, E., & Carruthers, P. (2018). Theory of mind. In T.K. Shackelford & V.A. Weekes-Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science*. Springer Dordrecht https://doi.org/10.1007/978-3-319-16999-6_2376-1

Whitaker, R. (2011). *Anatomy of an Epidemic: Magic Bullets, Psychiatric Drugs, and the Astonishing Rise of Mental Illness in America*. Crown Publishers.

White, R., Gumley, A., McTaggart, J., Rattrie, L., McConville, D., Cleare, S., & Mitchell, G. (2011). A feasibility study of Acceptance and Commitment Therapy for emotional dysfunction following psychosis. *Behaviour research and therapy*, *49*(12), 901–907. https://doi.org/10.1016/j.brat.2011.09.003

Wilder, D. A., Masuda, A., O'Connor, C., & Baham, M. (2001). Brief functional analysis and treatment of bizarre vocalizations in an adult with schizophrenia. *Journal of Applied Behavior Analysis, 34*(1), 65–68. https://doi.org/10.1901/jaba.2001.34-65

Wilder, D. A., White, H., & Yu, M. L. (2003). Functional analysis and treatment of bizarre vocalizations exhibited by an adult with schizophrenia: A replication and extension. *Behavioral Interventions, 18*(1), 43–52. https://doi.org/10.1002/bin.128

Wilder, D. A., Wong, S. E., Hodges, A. C., & Ertel, H. M. (2020). Schizophrenia and other psychotic disorders. In P. Sturmey (Ed.), *Functional analysis in clinical treatment* (pp. 315-338). Academic Press.

Wittgenstein, L. (1958). *Philosophical Investigations* (G.E.M. Anscombe, Trans.). Blackwell (Original work published 1953).

Wittgenstein, L. (1961). *Notebooks, 1914-1916* (G. E. M. Anscombe, G. H. von Wright, Eds., G. E. M. Anscombe, Trans.). Blackwell

Wittgenstein, L. (1969). *On certainty* (G. E. M. Anscombe, G. H. von Wright, Eds., D. Paul, & , G. E. M. Anscombe, Trans.). Blackwell.

Wittgenstein, L. (1974). *Philosophical grammar* (Rhees, Ed., Kenny, Trans.) Blackwell.

Wittgenstein, L. (1980a). *Remarks on the Philosophy of Psychology Vol 1.* (G. E. M. Anscombe, G. H. von Wright, Eds., D. Paul, &, G. E. M. Anscombe, Trans.). Blackwell.

Wittgenstein, L. (1980b). *Remarks on the Philosophy of Psychology Vol 2.* (G. H. von Wright, & H. Nyman, Eds., C. G. Luckhardt, & M. A. E. Aue, Trans.). Blackwell.

Wittgenstein, L. (1982). *Last Writings on the Philosophy of Psychology*, *Vol 1.* (G. H. von Wright, & H. Nyman, Eds., C. G. Luckhardt, & M. A. E. Aue, Trans.). Blackwell.

Wittgenstein, L. (1992). *Last Writings on the Philosophy of Psychology*, *Vol 2.* (G. H. von Wright, & H. Nyman, Eds., C. G. Luckhardt, & M. A. E. Aue, Trans.). Blackwell.

Wittgenstein, L. (2001). *Tractatus Logico-Philosophicus* (D. F. Pears & B. F. McGuinness, Trans.). Routledge (Original work published 1921).

Wolpe, J. (1952). Objective psychotherapy of the neuroses. *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde*, *26*(42), 825–829.

Wolpe, J. (1954). Reciprocal inhibition as the main basis of psychotherapeutic effects. *A.M.A. Archives of Neurology and Psychiatry, 72,* 205–226. https://doi.org/10.1001/archneurpsyc.1954.02330020073007

Wolpe, J. (1959). Psychotherapy Based on the Principle of Reciprocal Inhibition. In A. Burton, (Ed). *Case studies in counseling and psychotherapy* (pp. 353-381). Prentice-Hall.

Wolpe, J. (1978). Cognition and causation in human behavior and its therapy. *American Psychologist, 33*(5), 437–446. https://doi.org/10.1037/0003-066X.33.5.437

Wong, S. E. (2006). Behavior Analysis of Psychotic Disorders: Scientific Dead End or Casualty of the Mental Health Political Economy? *Behavior and Social Issues, 15*(2), 152–177. https://doi.org/10.5210/bsi.v15i2.365

Wong, S. E. (1996). Psychosis. In M. A. Mattaini & B. A. Thyer (Eds.), *Finding solutions to social problems: Behavioral strategies for change* (pp. 319–343). American Psychological Association. https://doi.org/10.1037/10217-012

Wong, S. E. (2014). A critique of the diagnostic construct schizophrenia. *Research on Social Work Practice, 24*(1), 132–141. https://doi.org/10.1177/1049731513505152

Wood, L., Burke, E., & Morrison, A. (2015). Individual cognitive behavioural therapy for psychosis (CBTp): a systematic review of qualitative literature. *Behavioural and Cognitive Psychotherapy*, *43*(3), 285-297. https://doi.org/10.1017/S1352465813000970

World Health Organization. (2022). *Schizophrenia*. https://www.who.int/news-room/fact-sheets/detail/schizophrenia

Wright, C. (1998). Self-knowledge: the Wittgensteinian legacy. *Royal Institute of Philosophy Supplements*, *43*, 101-122.

Wykes, T., Steel, C., Everitt, B., & Tarrier, N. (2008). Cognitive behavior therapy for schizophrenia: effect sizes, clinical models, and methodological rigor. *Schizophrenia bulletin*, *34*(3), 523-537. https://doi.org/10.1093/schbul/sbm114

Young, A. W. (1999). Delusions. *The Monist*, *82*(4), 571–589. http://www.jstor.org/stable/27903656

Zawidzki, T. W. (2008). The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations*, *11*(3), 193-210. https://doi.org/10.1080/13869790802239235

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Zettle, R. D., & Hayes, S. C. (1982). Rule governed behavior: A potential theoretical framework for cognitive behavior therapy. In P. C. Kendall (Ed.), *Advances in cognitive behavioral research and therapy* (pp. 73-118). New York: Academic.

Zettle, R. D., Hayes, S. C., Barnes-Holmes, D., & Biglan, A. (2016). *The Wiley handbook of contextual behavioral science*. John Wiley & Sons.

Zhong, L. (2019). Taking emergentism seriously. *Australasian Journal of Philosophy*, *98*(1), 31-46. https://doi.org/10.1080/00048402.2019.1589547

Zilio, D., & Carrara, K. (Eds.). (2021). *Contemporary Behaviorisms in Debate*. Springer.

Zimmermann, G., Favrod, J., Trieu, V. H., & Pomini, V. (2005). The effect of cognitive behavioral treatment on the positive symptoms of schizophrenia spectrum disorders: a meta-analysis. *Schizophrenia research*, *77*(1), 1-9. https://doi.org/10.1016/j.schres.2005.02.018