

A critical assessment of the goal replacement hypothesis for habitual behaviour

David Luque^{*1,2} & Sara Molinero^{1,2}

¹*Universidad Autónoma de Madrid, Spain*

²*Universidad de Málaga, Spain*

Learning how to obtain rewards (e.g., food) is important for survival. Behavioural and neuroscience research have suggested that reward learning reflects the operation of two distinct neuro-cognitive systems: the *goal-directed and habit systems* (Balleine & O'Doherty, 2010). Recently, this dichotomy has been challenged by authors proposing that, what we thought were habitual responses, are better understood as goal-directed actions (Kruglanski & Szumowska, 2020). This letter is a critical assessment of this hypothesis, which in our opinion can hardly explain the results of recent studies. Also, we suggest that the debate about the goal-directedness of habits is probably insoluble from an experimental point of view and almost irrelevant from an applied/clinical perspective. A more fruitful approach might be to analyse to what extent a behaviour is *controllable*—regardless if it is purposeless or not.

Critics of the habit literature focus their analysis on the standard definition of habits¹: Habits are responses (R) that are triggered by a stimulus (S) regardless the ongoing value of its outcome (O). This outcome insensitivity is thought to be achieved by extensive reward training (Tricomi et al., 2009). As Kruglanski & Szumowska (2020) has recently pointed out,

***Corresponding author:** Correspondence concerning this article should be addressed to David Luque, Departamento de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, 28049 Madrid, Spain. E-mail: david.luque@gmail.com. **Conflict of interest.** We have no conflicts of interests to disclose. DL and SM was supported by grant PGC2018-094694-B-I00 (AEI / FEDER, UE) and grant 2017-T1/SOC-5147. (Comunidad de Madrid). **Author note:** David Luque <https://orcid.org/0000-0002-3457-9204>; Sara Molinero <https://orcid.org/0000-0003-1065-2988>.

¹ In fields other than experimental psychology (health psychology, marketing, sports science, etc) habits are understood as behaviours that are frequently repeated.

there is not much evidence of outcome-insensitive behaviour in humans (see also, de Wit et al., 2018). There are, however, studies in which highly-trained responses persist despite being at odd with ongoing goals (e.g., Neil et al., 2011). For these cases, Kruglanski & Szumowska (2020) and others (De Houwer et al., 2017; Moors et al., 2019; Moors & Houwer, 2017) came out with the *goal replacement hypothesis*: They claim that intrusions of highly-trained responses are not produced by the activation of outcome-insensitive habits but by the activation of alternative goals —also called *hidden goals*.

The goal replacement hypothesis might work in some scenarios, e.g., those related with overeating/drinking when you are satiated (social goals might lead you to keep consuming) but is hard to reconcile with others. Consider, for instance, the recent article by Hardwick et al. (2019). In this study, two groups of participants (more vs less extensive practice) learned several S-R associations; then some of these S-R associations were remapped. These authors showed that intrusions of the original S-R mapping were more frequent after overtraining and when forcing participants to respond very rapidly. The goal replacement hypothesis should explain why an alternative goal, either than solve the ongoing task in an efficient way, was activated during test, and what the participants tried to achieve by that goal. It has been claimed the activation of hidden goals aim to solve the discrepancy between the stimulus (S) and the ongoing goal (Moors & De Houwer, 2017), but in Hardwick's experiments it is not clear what that discrepancy might be.

We can apply the same analysis to one of our recent studies (Luque et al., 2020). Conceptually, this study shared some features with Hardwick et al. (2019): we also manipulated the amount of training and assessed habit formation in a partial reversal test, while participants had time pressure for responding. We found that the original S-R links interfered with correct goal-directed reversals, as it was evident in response time data. This effect was especially strong after overtraining. Notably, accuracy data (proportion of correct reversals) did not change with training, even in some cases, improved. From the standard habit theory this result is straightforward: S-R links became stronger with training, but not strong enough to overrule the goal-directed reversal. Because the original R and the reversal R shared the same triggering S, the original S-R link was automatically activated during test and produced a measurable interference effect over the correct goal-directed reversal. As in the case of Hardwick's results, there is not a clear way to explain these results by using the goal replacement principle. Moreover, because we measured S-R links from the interference produced over *correct reversals*, we know that the actual goal of the participants was to make the correct reversal responses, avoiding the original one. The goal replacement

hypothesis should explain why and how the participants would activate any additional goal during test, an additional goal which was only counterproductive for them.

As implausible as it might sound, we can surely force the goal replacement account for explaining the results from Hardwick et al. and Luque et al. Possible *ad hoc* explanations might involve automatic activations of hidden goals which, for some reason, would activate a wrong response (see Moors & De Houwer, 2017). For explaining Luque et al. results, we need to assume that this counterproductive goal can be kept activated at the same time than the main goal of doing the task just fine. This might sound forced—but is not impossible. What leads us to the bottom-line problem: we can always think ad hoc explanations based in unobserved goals for explaining almost any pattern of results. In other words, the goal replacement is a theory which is extremely hard to be proven wrong, if not directly unfalsifiable.

Then, are really habits purposeless behaviours? An unfalsifiable theory is not very useful for science but is not necessarily wrong, what might lead the whole field to be stuck in an unsolvable problem. Fortunately, if habits are truly goal independent might not be that relevant. Proving that a behaviour is goal-independent is not only an extremely hard task from an experimental point of view, but it is also almost irrelevant from an applied/clinical perspective. Think, for instance, in an OCD patient who cannot help to wash repeatedly their hands. Is this behaviour truly purposeless or not? We can expend the rest of our lives trying to figure out if this behaviour is automatically triggered by an S or by the activation of an automatic goal instead. However, what is important to know is whether that behaviour can be *controlled* under certain circumstances, and/or if we can modify this pattern of behaviour in a stable way.

Studying behaviour controllability is not only more relevant from an applied/clinical perspective, it also might help us to escape the theoretical and practical complexities of studying hidden goals and purposeless behaviours. We can operatively establish that a response is hard or easy to control from the degree of interference that the response produces in an incompatible ongoing goal. In other words, hard to control responses are those which are still active even when they produce counterproductive consequences for the participants. Importantly, by studying hard to control responses we are probably tapping in the same cognitive and brain areas that are usually studied under the label of “habits” (see Dezfouli & Balleine, 2012), but avoiding unfruitful debates about what a habit truly is and/or if they are really purposeless.

REFERENCES

- Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35(1), 48–69. <https://doi.org/10.1038/npp.2009.131>
- De Houwer, J., Tanaka, A., Moors, A., & Tibboel, H. (2017). Kicking the Habit: Why evidence for habits in humans might be overestimated. *Motivation Science*, 1–27. <https://doi.org/10.1037/mot0000065>
- de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A., Robbins, T. W., Gasull-Camos, J., Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, 147(7), 1043–1065. <http://dx.doi.org/10.1037/xge0000402>
- Dezfouli, A., Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35, 1036–1051. <https://doi.org/10.1111/j.1460-9568.2012.08050.x>
- Hardwick, R. M., Forrence, A. D., Krakauer, J. W., & Haith, A. M. (2019). Time- dependent competition between goal-directed and habitual response preparation. *Nature Human Behaviour*, 3(12), 1252–1262. <https://doi.org/10.1038/s41562-019-0725-0>
- Kruglanski, A. W., & Szumowska, E. (2020). Habitual Behavior Is Goal-Driven. *Perspectives on Psychological Science*, 15(5), 1256–1271. <https://doi.org/10.1177/1745691620917676>
- Luque, D., Molinero, S., Watson, P., López, F. J., & Le Pelley, M. E. (2020). Measuring habit formation through goal-directed response switching. *Journal of Experimental Psychology: General*, 149(8), 1449–1459. <https://doi.org/10.1037/xge0000722>
- Moors, A., Fini, C., Everaert, T., Bardi, L., Bossuyt, E., Kuppens, P., & Brass, M. (2019). The role of stimulus-driven versus goal-directed processes in fight and flight tendencies measured with motor evoked potentials induced by Transcranial Magnetic Stimulation. *PLoS ONE*, 14(5), 1–22. <https://doi.org/10.1371/journal.pone.0217266>
- Moors, A., & De Houwer, J. (2017). The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*, 9(4), 310–331. <https://doi.org/10.1177/1754073916669595>
- Neal, D. T., Wood, W., Wu, M., & Kurlander, D. (2011). The pull of the past: When do habits persist despite conflict with motives?. *Personality and Social Psychology Bulletin*, 37(11), 1428–1437. <https://doi.org/10.1177/0146167211419863>
- Tricomi, E., Balleine, B. W., O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, 29, 2225–2232. <https://doi.org/10.1111/j.1460-9568.2009.06796.x>