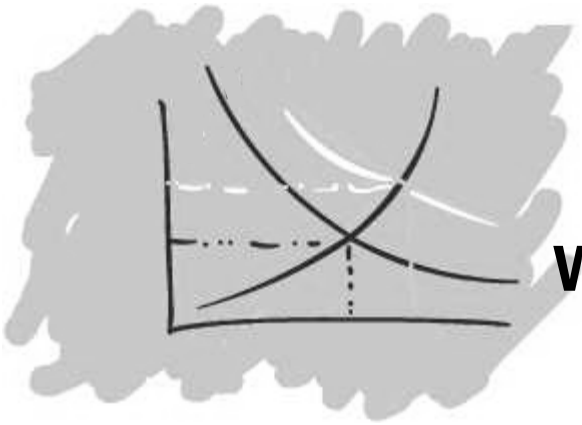
I.S.S.N: 1885-6888



ECONOMIC ANALYSIS WORKING PAPER SERIES

On Approval and Disapproval: Theory and Experiments

♦

Raúl López Pérez y Marc Vorsatz

Working Paper 8/2009

UAM
UNIVERSIDAD AUTONOMA
DE MADRID

DEPARTAMENTO DE ANÁLISIS ECONÓMICO:
TEORÍA ECONÓMICA E HISTORIA ECONÓMICA

On Approval and Disapproval: Theory and Experiments

Raúl López-Pérez * Marc Vorsatz†

December 14, 2009

Abstract

Prior studies have shown that selfish behavior is reduced when co–players have the opportunity to approve/disapprove a player's choice, even if that has no consequences on the player's material payoff. Using a prisoner's dilemma, we experimentally study the causes of this phenomenon, which seems crucial to understand compliance with social norms. Our data is consistent with a model based on the assumption that people feel badly if they expect to be disapproved by others. Furthermore, we find suggestive evidence in line with the following assumptions: (i) People become more aware about the others opinion if feedback is available, and (ii) even if the feedback is ex post and has no effect on their ex ante expectations about disapproval, people prefer not to receive negative feedback.

Keywords: Approval, disapproval, non–material rewards/sanctions, social norms.

JEL-Numbers: A13, C72, D64, Z13.

"Compared with the contempt of mankind, all other evils are easily supported"

(from: The theory of moral sentiments, by Adam Smith)

*Corresponding author. Universidad Autónoma de Madrid, Facultad de Ciencias Económicas, Departamento de Análisis Económico, Cantoblanco, 28049 Madrid, Spain. Email: raul.lopez@uam.es

†Fundación de Estudios de Economía Aplicada–FEDEA, Calle Jorge Juan 46, 28001 Madrid, Spain. Email: mvorsatz@fedea.es.

1 Introduction

This paper uses theory and experiments to study why *non–material* rewards and sanctions affect behavior. We term rewards/sanctions to be non–material if they do not alter the material welfare of the rewarded/sanctioned agent, but affect her emotional state. Frequently conveyed by means of verbal or facial expressions, they transmit our approval/disapproval of the others' behavior or personal qualities (for this reason, we will often refer to them simply as *feedback*). Thus, examples of non–material sanctions include social disapproval, humiliation, insults, peer pressure, public embarrassment, and social ostracism.[1]

Exploring why such feedback affects behavior is interesting for several reasons. First, compliance with social norms, which is decisive for the functioning of societies (Homans 1961, Arrow 1974, Elster 1989, and Fehr and Fischbacher 2004), seems to be promoted when people can receive feedback. This has been stressed, among others, by social scientists like Emile Durkheim and Talcott Parsons. Second, non–material rewards and sanctions have two appealing features: They are often less costly to apply than their material counterparts and their use is apparently less constrained by social norms. The use of material sanctions by private parties, in constrast, is often regarded as morally reprehensible –*e.g.*, many people believe that parents should never punish physically their children.[2]

Consistent with the idea that non–material sanctions/rewards promote compliance with norms, a growing experimental literature has provided evidence that this type of feedback reduces selfish behavior (we review this literature in the next section). However, it is not yet clear why this occurs. Note that recent models of other–regarding preferences (see Camerer

[1] We note that some of the literature uses the term "non–monetary punishment" as a synonym of non–material sanctions, as in Masclet et al. (2003).

[2] Of course, non–material rewards and sanctions have the disadvantage that they cannot induce compliance among those agents who are immune to social approval/disapproval.

2003 and Fehr and Schmidt 2006 for surveys) cannot explain this phenomenon: Approving or disapproving an action does not affect the distribution of material resources among the players or the (ex ante) beliefs in this regard, and the utility in most of these models depends on one (or both) of these variables.

This paper complements the literature on other–regarding preferences and aims at clarifying why approval and disapproval affect behavior. For this, we first propose a simple model based on the idea that some people feel badly if they expect their actions to be disapproved by others, whereas they feel well if they expect their actions to be approved –see Holländer (1990) and Kandel and Lazear (1992) for alternative models. For simplicity, we just refer to this as the disapproval–aversion (DA) assumption. While this DA hypothesis is important, we must stress that it cannot explain on its own why the feedback affects behavior. This is especially clear if the feedback is *ex post*: Since the expected utility of a player depends on her *ex ante* expectations about approval/disapproval and the feedback is provided after the choice, why should it have an effect on its own? To provide an explanation, we consider two additional hypotheses: First, players might be more likely to think about the other's opinions if there is a feedback stage or some other external factor that somehow forces them to put on the other player's shoes (we call this the awareness hypothesis), and second, players might be averse to *effectively* receive disapproving messages, which we call the negative information avoidance (NIA) hypothesis.

In order to test for these hypotheses and to gain further understanding about why feedback affects behavior, we use the prisoner's dilemma (PD) and three experimental treatments (control, expectations, and feedback). In the control treatment, subjects simply play a PD game. In the expectations treatment, we elicit players' expectations about approval/disapproval from their co–player *before* playing the same PD. Finally, the feedback treatment consists of

a two–stage game. In the first stage, subjects play the PD game. In the second stage (the feedback stage), they can approve/disapprove the co–player's prior choice in the PD by means of a message that has no effect on the receiver's material payoff.

Our main results are as follows. *First*, the three treatments allow us to discriminate whether the awareness and NIA hypotheses are empirically relevant. In effect, suppose to start with that the awareness factor is *individually* significant while the NIA factor is not. In that case, the rate of cooperation should be similar in the expectations and the feedback treatments (in both treatments, there is an external factor that makes players think about the approval/disapproval from the co–player), and lower in the control treatment (where there is no such external factor). If only the NIA factor was *individually* significant, in contrast, the level of cooperation should be similar in the control and expectations treatments, but higher in the feedback treatment (the only one in which players can effectively receive messages). In short, the comparisons control *vs.* expectations and expectations *vs.* feedback capture the net effect of the awareness and NIA factors, respectively. In addition, the comparison control *vs.* feedback evaluates the effect of *both* factors when they act jointly. In this respect, we only observe significant differences in the cooperation rate in the comparison control *vs.* feedback, but not on the other two comparisons. In other words, the awareness and NIA factors have a significant effect when they act *jointly*, but not when they act individually. This suggests that both forces are necessary to understand why the feedback increases cooperation.

Second, we can also test the DA hypothesis. In effect, the model predicts that if a player is sufficiently sensitive to being approved and disapproved, and always expects cooperation to be relatively more approved than defection, then that player should cooperate in the PD. The expectations treatment allows us to test this prediction. Consistent with our model, we find that the average cooperator expects cooperation to be relatively more approved than

defection. Interestingly, this is not necessarily true for those players who defect.

Third, our feedback treatment provides some further insights on approval/disapproval. On one hand, we observe that cooperation is relatively more approved than defection. However, cooperators provide a more positive/negative feedback than defectors if their co–player cooperates/defects. Joined with our second result above, this means that cooperators do not only expect cooperation to be approved, but also that they approve this kind of behavior. A similar result is not found for the average defector.

The remainder of the paper proceeds as follows. The next section reviews some related literature, while Section 3 develops the model of disapproval–aversion. In Section 4, we present the experimental design and procedures. Section 5 provides experimental evidence on the prisoner's dilemma in our three treatments. We conclude in Section 6. The experimental instructions as well as some additional parametric estimations are relegated to the appendices.

2 Related Literature

As we have noted, one can distinguish between material and non–material rewards/sanctions. We first mention briefly some of the experimental literature on material rewards and sanctions, which is rather rich –consult Fehr and Schmidt (2006) for a more extensive survey. For instance, Ostrom et al. (1992) show that cooperation can be sustained in repeated social dilemmas by adding a punishment stage –see also Fehr and Gächter (2000)– and that the use of sanctions is inversely related to its cost and positively correlated with its effectiveness. Falk et al. (2005) study the determinants of the occurrence of punishment and argue that retaliation against unfair behavior is the main explanatory factor (although they stress the existence of other, secondary factors). Sefton et al. (2007) compare one–to–one reward and sanction mechanisms in a repeated voluntary contribution mechanism (VCM) and find that

contributions in the punishment treatment are rather stable over time. In contrast, contributions in the rewards treatment are initially high but markedly decrease in later rounds, maybe because the use of rewards significantly decays over time as well. In the context of a common pool resource (CPM) game, Vyrastekova and van Soest (2008) find also that one–to–one rewards are not effective in sustaining cooperation. However, this is not true when the impact ratio is $1:3$. As in the case of sanctions, therefore, this suggests that the impact ratio is crucial to understand the efficiency of rewards.

The evidence on non–material sanctions/rewards is less abundant, although increasing. For instance, Masclet et al. (2003) study a repeated VCM with a punishment stage and compare one treatment with material sanctions with another one with non–material sanctions, which were implemented through the assignment of costless (and non–costly) "disapproval points". They report that both types of sanctions were effective in promoting contributions to the public good, although material sanctions were more effective over time. In addition, Noussair and Tucker (2005) show that contributions to the public good are higher when both types of sanctions are present than if just one of the two types is available. Rege and Telle (2004) report that contributions to a one–shot public good game played among strangers are significantly higher when individual contributions are made public at the end of the game, thus facilitating social approval/disapproval –in contrast, Gächter and Fehr (1999) do not observe this phenomenon in a 10–period public good game. Peeters and Vorsatz (2009) suggest that players in a repeated public good game with partners matching use non–material rewards/sanctions as a signaling device for the next round, which sometimes increases contributions. Dugar (2008) examines a coordination game with several Pareto–ranked equilibria and two treatments: One with only non–material sanctions and another one with only non–material rewards. He reports a gradual convergence towards the most efficient

equilibrium in the first treatment, but the opposite result in the approval treatment. Hence, approval and disapproval have asymmetric behavioral effects on coordination. Furthermore, Ellingsen and Johannesson (2008) show that the average giving in one–shot dictator games increases if recipients can send an open–form, written message to the dictator after observing her choice. Finally, Xiao and Houser (2009) report the same phenomenon even if dictators are not required to read the message. Based on previous literature, they suggest that players could form their expectations about approval and disapproval in a self–serving manner, and find their data consistent with the hypothesis that self–serving biases are diminished if players can receive ex post feedback.

3 Disapproval–Aversion: A Toy Model

This section offers a simple model based on the idea that people feel well/badly if their choices are approved/disapproved by another person. The model applies to any two–player, extensive form game. Let $N = \{1, 2\}$ denote the set of players and $s = (s_1, s_2)$ a profile of pure strategies belonging to the strategy space $S = S_1 \times S_2$. Let $u(s) = (u_1(s), u_2(s))$ be the vector of utilities that ensues if s is played. The corresponding vector of *monetary* payoffs is denoted by $\pi(s) = (\pi_1(s), \pi_2(s))$. Players act rational (and there is common knowledge thereof), in the sense that they maximize their expected utility given their beliefs about the other player's choices.

We posit that players care about their own monetary payoff but also about whether the other player approves or disapproves their behavior. More precisely, let $I_i^j(s) \in \{-1, 0, 1\}$ be an indicator function that takes the value -1 if strategy s_i of player i is disapproved by player j ($j \neq i$) at $s \in S$, value 0 if that behavior is neither approved nor disapproved by player j, and value 1 if it is approved by player j. Given this, we assume that the utility

function of player i is $u_i(s) = \pi_i(s) + \gamma_i\, I_i^j(s)$, where $\gamma_i \geq 0$. In other words, a player suffers a psychological cost of γ_i if the other person thinks badly of her behavior, while she gets a psychological reward of the same amount if her co–player approves her behavior.[3] For brevity, we just refer to this assumption by saying that players are *disapproval–averse* (DA). The psychological intuition behind this assumption is twofold: First, humans construct their self–image partly by resorting to others' opinions (see Festinger, 1954), and second, a negative self–image triggers negative emotions like shame, inferiority feelings, or low self–esteem, while a positive self–image triggers positive feelings like pride.

We also assume that the population consists of a continuum of DA types who differ in their γ_i. In particular, the model includes the case $\gamma_i = 0$. Of course, this corresponds to the *homo economicus*, selfish type whose utility only depends on her own monetary payoff. We also assume that a player's type is private information and denote by $\rho(\gamma)$ the commonly known probability that player i has a γ_i of *at least* γ. Obviously, $\rho(0) = 1$ and $\lim_{\gamma\to\infty} \rho(\gamma) \to 0$.

To illustrate some implications of the DA hypothesis, we consider first some games where players are anonymous and cannot communicate in any manner –hence, they cannot exchange verbal or written messages, or even observe facial expressions. In these games, obviously, players are uncertain about the other player's opinion –*i.e.*, about the specific value of $I_i^j(s)$. Nevertheless, we assume that they have, for any possible strategy profile, some expectations on whether the co–player (privately) approves her choice. Given such expectations, players maximize their expected utility $\mathbb{E}_i(u(s)) = \mathbb{E}(\pi_i(s)) + \gamma_i\, \mathbb{E}(I_i^j(s))$, where $\mathbb{E}(I_i^j(s))$ denotes the ex ante expectation on $I_i^j(s_i, s_j)$, or the *expected approval rate* of s_i (given that j plays s_j).[4]

[3] Assuming that approval and disapproval have the same absolute effect on utility is an obvious simplification. In fact, one possible interpretation of the evidence from Dugar (2008) is that people are relatively less sensitive to approval than to disapproval. This could be incorporated into the model by making parameter γ_i dependent on the value of the indicator $I_i^j(s)$, and more precisely, smaller when $I_i^j(s) = 1$. Given that this assumption is not essential for our posterior analysis, we have nevertheless opted for the symmetric specification.

[4] The previous notation implicitly indicates that the probability that some player j approves a given choice

We start with a simple example: A binary dictator game in which the dictator i chooses between the (dictator, dummy) monetary allocations (200,120) and (160,160). To insist, note that the dictator is in this game uncertain about the dummy's opinion. Let then $\alpha_{200} \in [-1,1]$ and $\alpha_{160} \in [-1,1]$ be the expected approval rates of the respective choices. Given this, the dictator picks the egalitarian allocation (160,160) if, and only if, $160+\gamma_i\,\alpha_{160} > 200+\gamma_i\,\alpha_{200}$, which is equivalent to $(\alpha_{160}-\alpha_{200})\,\gamma_i > 40$. In other words, allocation (160,160) will be chosen only if γ_i is large enough and this choice is expected to be approved more often than the selfish choice ($\alpha_{160}-\alpha_{200}>0$).

As a second example of a game without communication (now more suited to our posterior experimental analysis), consider the one–shot prisoner's dilemma game (PD) of Figure 1, where *monetary* payoffs satisfy $t > c > d > 0$ and $2c > t$.

	Cooperate	Defect
Cooperate	(c,c)	$(0,t)$
Defect	$(t,0)$	(d,d)

Figure 1: Prisoner's dilemma game.

Since defection strictly dominates cooperation in monetary terms, it is clear that a selfish player i should always defect. A player with $\gamma_i > 0$, however, cooperates under certain conditions. To show this, let $\alpha_{CD} \in [-1,1]$ be the expected approval rate of a player who cooperates when the other player defects (α_{CC}, α_{DD}, and α_{DC} are analogously defined). If the co–player is expected to cooperate with probability δ, it follows that the average expected approval rate of a player who cooperates, or *expected approval rate of cooperation* $\alpha_C(\delta)$, equals $\alpha_C(\delta) = \delta\,\alpha_{CC} + (1-\delta)\,\alpha_{CD}$. Analogously, her *expected approval rate of defection*

does not depends on her γ_j so that we do not need to condition $\mathbb{E}(I_i^j(s))$ on the distribution of γ in the population. While this assumption simplifies the exposition, it is not necessary for our results.

equals $\alpha_D(\delta) = \delta\alpha_{DC} + (1-\delta)\alpha_{DD}$. Under the hypothesis that cooperation is always expected to be more approved than defection, Proposition 1 determines the non–empty set of types who cooperate in the PD.

Proposition 1 *If $\alpha_C(\delta) > \alpha_D(\delta)$ for any $\delta \in [0,1]$, there exists a $\gamma^* > 0$ so that player i contributes in the PD game if, and only if, $\gamma_i \geq \gamma^*$.*

Proof: In what follows, we show that the types with $\gamma_i \geq \gamma^*$ find cooperation strictly dominant under this condition. In effect, provided that the co-player cooperates with probability δ, it follows that player i's expected utility from cooperation is equal to

$$\delta\,(c + \gamma_i\,\alpha_{CC}) + (1-\delta)\,\gamma_i\,\alpha_{CD} = \delta\,c + \gamma_i\,\alpha_C(\delta).$$

On the other hand, if she defects her expected utility is

$$\delta\,(t + \gamma_i\,\alpha_{DC}) + (1-\delta)\,(d + \gamma_i\,\alpha_{DD}) = \delta\,t + (1-\delta)\,d + \gamma_i\,\alpha_D(\delta).$$

Consequently, cooperation is optimal if and only if

$$\delta\,c + \gamma_i\,\alpha_C(\delta) \geq \delta\,t + (1-\delta)\,d + \gamma_i\,\alpha_D(\delta) \Leftrightarrow \gamma_i\,(\alpha_C(\delta) - \alpha_D(\delta)) \geq \delta\,(t-c-d) + d.$$

Now, since $\alpha_C(\delta) - \alpha_D(\delta) > 0$ for any δ by assumption, the previous expression can be rewritten as

$$\gamma_i \;\geq\; \frac{\delta\,(t-c-d) + d}{\alpha_C(\delta) - \alpha_D(\delta)} \equiv \gamma^*.$$

Consequently, cooperation is optimal for player i if, and only if, $\gamma_i \geq \gamma^*$. It follows that a proportion $\delta = \rho(\gamma^*)$ of players cooperate in the PD. □

In other words: If a player always expects cooperation to be relatively more approved than defection, she cooperates if her γ_i is sufficiently high. Conversely, if a player does not

expect $\alpha_C(\delta) > \alpha_D(\delta)$ for any δ, she will defect independently of her type. This suggests that belief heterogeneity may play a role in explaining behavioral differences. Finally, one can prove that there may exist several equilibria if $\alpha_C(\delta) > \alpha_D(\delta)$ for some but not for all δ.[5]

Suppose now that players can provide feedback in the PD. Should that foster cooperation? Since players can obtain information about approval/disapproval in that case, one might be tempted to believe that our model predicts a positive effect on cooperation, *even* if players communicate after all other choices have been made. However, we can show with a simple example that *disapproval–aversion alone predicts no change in behavior*. The example corresponds to a game with two stages: In the first stage, the PD game of Figure 1 is played; in the second stage, each player observes the co–player's prior choice and sends then a message expressing approval or disapproval of that choice. In other words, players can non–materially punish or reward the co–player after playing the PD game. Applying our model to this game is direct: Since ex post messages cannot alter the expected approval rates of cooperation and defection $\alpha_C(\delta)$ and $\alpha_D(\delta)$, it follows that they cannot affect the ex ante expected utility either (they may affect the ex post utility, but this is irrelevant here). As a result, adding the feedback stage has no effect on the cooperation rate.

However, we know from the experimental literature that the addition of a feedback stage increases the overall cooperation rate. To explain this phenomenon, we complement the model with two possible factors. First, it could be that some agents do not become aware about the co–player's opinions unless they are somehow reminded about this by an external factor; for instance, the mere availability of the feedback might help agents to focus on the other player's opinions. This implies that some agents might act as if their γ_i was large only if

[5] A formal proof is available on request. In any case, the experimental evidence provided later indicates that the average *cooperative* player expects $\alpha_C(\delta) > \alpha_D(\delta)$ for any δ. Hence, Proposition 1 applies to them.

their awareness was increased by some external factor, but not otherwise. More precisely, we model this idea as follows:

HYPOTHESIS 1 (Awareness): *The probability $\rho(\gamma^*)$ increases if an external factor makes players think about the other player's opinion.*

Second, players might not behave as standard expected utility maximizers when choosing under uncertainty, but feel *regret* (Loomes and Sugden, 1982). This means that if a player makes an uncertain choice and the consequence of that choice happens to be bad, she feels badly by reflecting on how much better her position would have been had she chosen differently. For instance, a player might regret her choice if it is disapproved by another player. Consequently, a player who feels regret should prefer a lottery where she is disapproved with some probability but she is *not* informed about the resolution of the uncertainty to another lottery that only differs from the former in that she is informed *ex post* whether she is actually disapproved –*e.g.*, by receiving a disapproving message. We model this idea by implicitly assuming that players become more sensitive (*i.e.*, increase their γ_i) if they can receive feedback about others' opinions. More precisely, we have the following negative information avoidance (NIA) hypothesis:

HYPOTHESIS 2 (NIA): *The probability $\rho(\gamma^*)$ increases if players can receive ex post information about others' approval/disapproval.*

To summarize, our model is based on three key hypotheses that might explain why individuals cooperate in the PD (the DA hypothesis) and why feedback affects behavior (the DA hypothesis joint with the awareness and/or NIA hypothesis). Our experiment will shed some light on the empirical validity of these assumptions. Before presenting our design, however, we briefly compare our model with some alternative specifications that seek to introduce ap-

proval/disapproval in the utility function. This comparison can further clarify the insights behind our model.

Two seminal papers have formalized the idea that people care about approval/disapproval from others. To start with, Höllander (1990) considers a setting in which n identical agents must decide how much of their endowment to contribute to a collective good. Agents obtain utility from their private consumption, the collective good, and the approval from other agents. In this regard, Höllander posits that approval is a continuous variable that depends positively on the agent's contribution to the collective good b and on the comparative value $b - c$, where c denotes the average contribution. That is, agents are more approved the more they contribute, and receive additional approval if they contribute above the average.

In turn, Kandel and Lazear (1992) consider a team of n workers who can put some costly effort e_i to produce a joint output $f(e_1, e_2, ..., e_n)$, which is later distributed equally among the workers. Workers are identical and their utility depends on their share of output, the cost of exerting effort, and a psychological cost or peer pressure P. Several formalizations of this cost are discussed. If P depends negatively on the worker's own effort, then the worse one feels the less one contributes. In addition, the cost might increase if the worker exerts less effort than the average worker, or if she is monitored by other workers.

There are several differences between these models and ours. First, we provide a general, game–theoretical model for two–player games. Among other things, this allows us to explicitly model communication and observability in detail.[6] Second, our model permits a priori any pattern of approval and disapproval. The previous models, in contrast, hypothesize

[6]We note a subtle point in this regard. When we think about monitoring/observing another player, we are likely to think about being physically present while that player makes her choice. In this setting, players can provide feedback through verbal or facial communication. In many games, however, one can observe another player's action just by observing the final output, even if one is not physically present. Since it is more difficult to provide feedback in this latter case, the two settings are qualitatively different. Our model recognizes this difference by means of the NIA hypothesis.

that people are relatively more approved/disapproved when they behave more/less cooperatively than the average player. While this hypothesis might be in principle an appealling one (particularly in multiple–player games), our experimental data shows that people are rather heterogeneous with regard to what they approve/disapprove: We will see, for instance, that many defectors in the prisoner's dilemma disapprove cooperation. This leads to our third point: Heterogeneity, which our model allows for, seems crucial to understand actual behavioral patterns.[7] Note finally that the model by Höllander (1990) cannot explain why the availability of feedback decreases selfish behavior, for the same reasons why our DA hypothesis alone cannot explain it either. None of them makes any assumption akin to the awareness or NIA hypotheses.

To finish, it can be worthy to compare the idea of disapproval–aversion with that of guilt–aversion –as in Battigalli and Dufwenberg (2007). Guilt–averse players suffer a cost if they choose what the co–player does not expect. In contrast, disapproval–averse players suffer a cost if they do what the co–player disapproves (*even* if the co–player expects that choice). Another key difference is that models of guilt–aversion posit that preferences depend *directly* on beliefs and hence are based on the Psychological Game Theory of Geanakoplos et al. (1989). This is not the case in our model: People care about being approved/disapproved, and their expectations in this regard affect *expected* utility. Note also that guilt–aversion cannot explain why ex post feedback affects behavior.

4 Experimental Design and Procedures

Our experimental study consists of three treatments (*control*, *expectations*, and *feedback*). In the *control treatment*, participants are randomly and anonymously matched into pairs and

[7] While Proposition 1 emphasizes heterogeneity in the sensitivity to the feedback (parameter γ_i), it also suggests that heterogeneity in the expectations about approval/disapproval is important as well.

play a single round of the PD presented in Figure 2. In order to avoid the terms "cooperate" and "defect", actions are respectively labelled X and Y instead. The numbers in the payoff matrix are in terms of ECU (experimental currency units), where 20 ECU equals 1 Euro.

	Cooperate	Defect
Cooperate	$(180, 180)$	$(80, 260)$
Defect	$(260, 80)$	$(100, 100)$

Figure 2: Specification of the PD game.

In the *expectations treatment*, subjects also play the previous PD. Before making their choice, however, they are asked what they expect their co–player will think about their choice for any possible outcome of the game. More precisely, each subject is asked four questions like the following one: "If I choose X and the other participant chooses X, he/she believes that my choice was (1) good, (2) neither good nor bad, (3) bad." Subjects are not paid for answering these questions.

Finally, in the *feedback treatment* subjects play the PD game of Figure 2 as well, but after observing the outcome of the game, they are given the opportunity to send one costly message to the other player (the cost of sending the message is 10 ECU). More precisely, each subject can choose one of the following three messages: "Your choice was (1) good, (2) neither good nor bad, (3) bad". To implement the feedback, we employ the strategy method; that is, subjects are asked for any possible contingency in the PD whether they want to pay the message fee and, independently of their answer, which message they would send in case they decided to pay the fee (of course, the co–player would only receive the message if the fee had been paid). We opted for the strategy method in order to maximize the amount of statistical evidence. Since the experiment was computerized and subjects were not warned

about the arrival of a message (if any) in their screens, they could not avoid reading it.

We chose the PD for our experiment due to its simplicity. Further, our three treatments permit us to explore the empirical validity of our three key hypotheses. To start, the DA hypothesis can be tested with the expectations treatment, as our analysis in Proposition 1 predicts a certain correlation between expectations and cooperation (a player should defect if cooperation is not expected to be approved more than defection, but she may cooperate otherwise). Further, the net effect of the awareness and NIA hypotheses can be evaluated by pairwise comparisons of our treatments. In effect, awareness (but not NIA) could explain a significantly higher cooperation rate in expectations *vs.* control, while an increase in cooperation from expectations to feedback would be consistent with NIA.

We conducted the experiment, which was programmed within the z-Tree toolbox provided by Fischbacher (2007), in the computer laboratory at Maastricht University. Since all students from the Faculty of Economics and Business Administration have an e–mail account associated with their student ID, we promoted the experiment mainly via electronic newsletters and gave students the opportunity to register online for their preferred session. In total, 180 undergraduates participated in the experiment. Since some students did not show up, 58 students participated in the control, 60 in the feedback, and 62 in the expectations treatment. No student took part in more than one session.

Each session proceeded as follows. In the beginning, subjects were randomly and anonymously matched into pairs. Each subject received instruction sheets depending on the treatment, an official payment receipt, and a set of control questions.[8] Subjects could study the instructions at their own pace and eventually occurring doubts were privately clarified. The experiment started once everybody answered all control questions correctly. After making

[8]The instructions corresponding to the feedback treatment can be found in the Appendix.

their choices in the corresponding PD game of each treatment (in the subjects' instructions, this was referred to as scenario 1), subjects played in addition five dictator games with feedback (scenarios 2-6); we report the results from these games in another paper. Participants in any treatment were initially informed that they would play six games, without providing any information about the structure of the five dictator games. Moreover, subjects knew that they would be randomly re–matched with another player after making their choices in the first scenario, that their decisions in scenario 1 would not affect their payoffs in any other scenario, and that they would not receive any information regarding the decisions of any other player until the end of the experiment (note that the use of the strategy method in the feedback treatment made this possible). In theory, therefore, behavior in the PD game in each treatment could not be affected by the existence of the additional games. After playing scenario 1, subjects were effectively re-matched and received new instruction sheets for the dictator games. In each treatment, and in order to prevent income effects, only one game among the PD game and the five dictator games was randomly selected for payment (subjects knew this from the beginning). At the end of the experiment, subjects were informed about their co–player's actions in the payoff–relevant game. The average payment for the 45 minutes session was about 9 Euros.

5 Experimental Evidence

Our experimental design allows us to test several implications of the model. First of all, Proposition 1 indicates that players with a sufficiently high γ_i should cooperate if that choice is expected to be relatively more approved than defection. Consequently, we use the data from the expectations treatment to study whether the assumption $\alpha_C(\delta) - \alpha_D(\delta) > 0$ holds within the group of cooperators for all $\delta \in [0, 1]$. We have the following result.

Result 1 (Expectations & Cooperation): *In the expectations treatment, cooperation is on average expected to be significantly more approved than defection. This result still holds for the subjects who actually cooperate, but not necessarily for the subjects who actually defect. This is because actual cooperators are more likely to expect approval (disapproval) when both players cooperate (defect) potentially.*

Evidence on Result 1: We present first some aggregate data on the subjects' expectations about approval/disapproval. More precisely, Table 1 shows the proportion of subjects who expect their potential choice to be approved or disapproved, depending on the other player's potential choice –recall that we elicited the players' expectations in this regard for any possible strategy vector. For instance, the column corresponding to the strategy vector (C,C) indicates that more than 67 % of the subjects expect to be approved if they cooperate and the other player cooperates as well, while around 14 % expect to be disapproved in this case.[9] Note that these percentages do not add up to 1 because some subjects expected their choice neither to be approved nor disapproved. In each column, we also indicate the p–values of the one–sided Wilcoxon signed rank tests that compare the approval and disapproval rates for the corresponding strategy vector.

	(C,C)	(C,D)	(D,C)	(D,D)
Approval	0.6774	0.4193	0.3548	0.1935
	[0.0000]	[0.2552]	[0.0136]	[0.0322]
Disapproval	0.1451	0.5322	0.6290	0.3870

Table 1: Expectations about how the co–player evaluates one's own action (in percentages). The first of the two actions of an outcome always corresponds to the player who assesses the expectation. In brackets, the one–sided p–values of the Wilcoxon signed rank tests that compare the equality of the approval and disapproval rate for a given outcome.

[9] The latter expectations are somehow surprising. All subjects with these kind of expectations actually defect in the PD (see the next table). They might believe that cooperation is always the wrong choice (perhaps because they see it as a dominated strategy), and they expect the co–player to believe this as well.

From this, we infer the following aggregate patterns: (a) The average expectation of approval α_{CC} –*i.e.*, the difference between the percentage of people who approve and those who disapprove in the (C,C) column– is significantly larger than zero so that the average subject expects approval to be more likely than disapproval if both subjects cooperate potentially, (b) α_{DD} and α_{DC} are both significantly smaller than zero, and (c) α_{CD} does not differ significantly from zero. As the reader can easily verify, all this implies that the average subject expects cooperation to be always significantly more approved than defection.

While this aggregate data suggests some general tendencies, one must note that there exists certain heterogeneity. For instance, many subjects expect to be disapproved if they cooperate unilaterally, but a similar proportion expects to be approved in this case. So, the question arises if there is a relation between actual behavior and expectations. In particular, is it true that the average cooperator expects $\alpha_C(\delta) - \alpha_D(\delta) > 0$ for any $\delta \in [0,1]$ as demanded by Proposition 1? To answer these kind of questions we disaggregate expectations, distinguishing between the subjects who actually cooperate and those who actually defect. The following table also indicates the p–values of the one–sided Wilcoxon signed rank tests that compare the approval and disapproval rates for each strategy vector.

	Expectations of Cooperators			Expectations of Defectors		
	Approval		Disapproval	Approval		Disapproval
(C,C)	0.8888	[0.0001]	0.0000	0.5909	[0.0020]	0.2045
(C,D)	0.3888	[0.2418]	0.5555	0.4318	[0.2709]	0.5227
(D,C)	0.4444	[0.3276]	0.5555	0.3181	[0.0111]	0.6590
(D,D)	0.0555	[0.0020]	0.6111	0.2500	[0.3416]	0.2954

Table 2: Expectations for the group of cooperators and the group of defectors about how the co–player evaluates one's own action (in percentages). The first of the two actions of an outcome always corresponds to the player who assesses the expectation. In brackets, the one–sided p–values of the Wilcoxon signed rank tests that compare the equality of the approval and disapproval rate for a given outcome and group.

Table 2 shows that (a) α_{CC} is significantly greater than zero for both subgroups, (b) α_{CD} is in both groups smaller than zero but not significantly so, (c) α_{DC} is significantly smaller than zero for the group of defectors but not for the group of cooperators, and (d) α_{DD} is significantly smaller than zero for the cooperators but not for the defectors. From this, it can be concluded that the average cooperator expects cooperation to be relatively more approved than defection for any belief δ that the co–player cooperates; that is, $\alpha_C(\delta) > \alpha_D(\delta)$ for all $\delta \in [0, 1]$. Interestingly, the data and in particular the above mentioned result (d) imply that $\alpha_C(\delta) > \alpha_D(\delta)$ does not always hold for the average defector, especially if she expects the co–player to defect (*i.e.*, if δ is close to zero).

To investigate the differences between both subgroups in more detail, we first analyze whether the average expectations differ across them. Using Mann–Whitney U tests, we find that α_{CD} ($p = 0.3959$, one-sided) and α_{DC} ($p = 0.2008$, one-sided) do not differ significantly across the two subgroups. On the other hand, α_{CC} is significantly larger (*i.e.*, more positive) for the cooperators ($p = 0.0087$, one–sided), while α_{DD} is significantly smaller (*i.e.*, more negative) for the cooperators ($p = 0.0069$, one–sided). That is, cooperators are more likely to expect approval (disapproval) if both players cooperate (defect).

In a further level of disaggregation, Table 3 provides for each subgroup information on how cooperation is expected to be evaluated relative to defection keeping the co–player's choice constant. Since subjects can provide three levels of evaluation (approval, neutral, disapproval), we distinguish five possible levels of comparison $\{+2, +1, 0, -1, -2\}$; for example, the value $+2$ is obtained if a subject expects cooperation to be approved and defection to be disapproved, 0 is obtained if both actions are expected to be evaluated equally, and so on. The table also includes the p–values of the one–sided Wilcoxon signed rank tests that analyze whether the mean of a distribution is different from zero.

	Expectations of Cooperators					Expectations of Defectors				
	+2	+1	0	-1	-2	+2	+1	0	-1	-2
$\alpha_{CC} - \alpha_{DC}$	0.50	0.06	0.39	0.06	0.00	0.45	0.14	0.23	0.04	0.14
	Mean: 1.00 [0.0020]					Mean: 0.72 [0.0013]				
$\alpha_{CD} - \alpha_{DD}$	0.28	0.11	0.39	0.17	0.05	0.18	0.23	0.16	0.23	0.20
	Mean: 0.40 [0.1190]					Mean: -0.02 [0.4151]				

Table 3: Frequencies of relative expectations for the group of cooperators and the group of defectors about how cooperation is evaluated relative to defection by the co–player (keeping the co–player's action constant). The first of the two actions of an outcome always corresponds to the player who assesses the expectation. In brackets, the one–sided p–values of the Wilcoxon signed–rank tests that analyze whether the mean of the distribution is different from zero.

We see that 50 % of the cooperators expect cooperation to be approved and defection to be disapproved when the co–player is assumed to cooperate, while 45 % of the defectors share the same expectations. An interesting point is that some subjects expect cooperation to be evaluated *worse* than defection, especially when the co–player is expected to defect (see, the line $\alpha_{CD} - \alpha_{DD}$). In fact, 43 % of all actual defectors and 22 % of all actual cooperators hold that kind of expectation. Incidentally, Proposition 1 does not apply to these cooperators, as they do not always expect cooperation to be relatively more approved than defection. Thus, not all cooperative behavior seems to be motivated by disapproval–aversion. □

While the data from the expectations treatment allowed us to test an implication of the DA hypothesis, the feedback treatment provides evidence on which kind of behavior is actually approved/disapproved, and who disapproves which behavior. Note that our model is silent in this respect as it does not impose any particular feedback pattern. We find the following.

Result 2 (Feedback & Cooperation): *In the feedback treatment, cooperation is approved significantly more often than defection. This finding is more pronounced for the subjects who actually cooperate than for those who actually defect.*

Evidence on Result 2: We present first some aggregate data. Table 4 presents the hypothetical messages that subjects sent to the co–player in the feedback treatment for every possible strategy vector of the PD (recall that we used the strategy method to elicit these messages). For instance, column (C,C) of the table indicates that 65 % of the subjects would approve the other player's choice if both cooperated, while around 1.6 % would disapprove in this case (the rest of the subjects would neither approve nor disapprove). To analyze whether the approval and disapproval rates are identical for some strategy vector, we use again one–sided Wilcoxon signed rank tests. Since all p–values are smaller than 0.05, the clear tendency is that subjects approve potential cooperators and disapprove potential defectors.

	(C,C)	(C,D)	(D,C)	(D,D)
Approval rate	0.6500	0.1166	0.4833	0.1500
	[0.0000]	[0.0000]	[0.0004]	[0.0212]
Disapproval rate	0.0166	0.5833	0.2000	0.3333

Table 4: Percentages of hypothetical messages in the feedback treatment. The first of the two actions of an outcome always corresponds to the sender of the message. In brackets, the one–sided p–values of the Wilcoxon signed rank tests that compare the equality of the approval and disapproval rate for a given outcome.

Note that Table 4 is based on hypothetical messages. To actually send the message for a certain outcome of the PD game, however, a subject had to agree to pay a 10 ECU fee. One reason that motivated us to introduce this fee was that some subjects might choose randomly between messages if they were costless, and that could contaminate the analysis. By comparing the message pattern of those players who are willing to pay the fee at a certain outcome with the corresponding hypothetical pattern, we can get some evidence that subjects did not make choices randomly. Indeed our results are not much different. In case both subjects cooperate potentially, ten players are willing to pay the fee; eight of them send an approving message and the other two a neutral one. If the other player defects unilaterally, ten

players are willing to pay the fee; nine players send a negative message and one player sends a positive one. If the other player cooperates unilaterally, seven players are willing to pay the fee; four messages are positive, one is neutral, and the other two are negative. Finally, if both subjects defect potentially, three players are willing to pay the fee; one message is negative and the other two are neutral.[10] It is worth noticing that cooperators are significantly more likely than defectors (0.2173 *vs.* 0.0810) to pay the 10 ECU fee to send a message, as a Z–test shows ($p = 0.0020$, one–sided).

While there is a clear tendency in the approval/disapproval patterns, Table 4 also suggests some heterogeneity. One possible reason might be the existence of some correlation between the behavior in the PD game and the approving/disapproving behavior. To clarify this, we analyze separately the frequency of messages within the group of cooperators and the group of defectors. The relevant results are displayed in Table 5.

	Feedback from Cooperators			Feedback from Defectors		
	Approval		Disapproval	Approval		Disapproval
(C,C)	0.7821	[0.0000]	0.0000	0.5675	[0.0000]	0.0270
(C,D)	0.0434	[0.0001]	0.7391	0.1621	[0.0075]	0.4864
(D,C)	0.5217	[0.0110]	0.1304	0.4594	[0.0600]	0.2432
(D,D)	0.1304	[0.0486]	0.4347	0.1621	[0.1659]	0.2702

Table 5: Percentages of hypothetical messages from the group of cooperators and the group of defectors. The first of the two actions of an outcome always corresponds to the sender of the message. In brackets, the one–sided p-values of the Wilcoxon signed rank tests that compare the equality of the approval and disapproval rate for a given outcome and group.

We see that in both subgroups, the approval rate of cooperation is greater than zero while that of defection is smaller than zero. Only within the group of defectors, α_{DD} and

[10] Due to the limited number of observations, tests on actual messages are not very powerful. However, we observe that the ratio positive messages/negative messages is significantly greater than zero ($p = 0.0030$, one–sided) for the outcome (C,C); significantly smaller than zero ($p = 0.0067$, one–sided) for the outcome (C,D); and, not significantly different from zero ($p = 0.2420$, one–sided) for the outcome (D,C). Finally, no meaningful test can be performed for the outcome (D,D) because we only have three independent observations in that case.

α_{DC} (defined now as rates of *actual* approval) do not turn out to be significantly different from zero at the five percent significance level. If we compare the approval rates across the two subgroups, we find the following results. First, cooperators are significantly more likely to approve in case of mutual cooperation ($p = 0.0432$, one–sided Mann-Whitney U test). Further, cooperators are also more likely to disapprove defection, as α_{CD} is significantly more negative for that group ($p = 0.0228$, one–sided). On the other hand, α_{DD} and α_{DC} do not differ significantly across groups ($p = 0.1270$ for each, one–sided). This evidence indicates that, as suggested above, there exists a relation between behavior and messages. That is, cooperators are more likely to approve cooperation and disapprove defection.

We can further clarify this point and provide more detailed evidence on individual behavior with the help of Table 6, which indicates how subjects in each subgroup rank cooperation and defection by the co–player, keeping their own choice constant. The five possible rankings $\{+2, +1, 0, -1, -2\}$ have an analogous interpretation as in Table 3.

	Feedback from Cooperators					Feedback from Defectors				
	+2	+1	0	-1	-2	+2	+1	0	-1	-2
$\alpha_{CC} - \alpha_{CD}$	0.65	0.22	0.09	0.04	0.00	0.43	0.16	0.27	0.11	0.03
	Mean: 1.48 [0.0001]					Mean: 0.85 [0.0002]				
$\alpha_{DC} - \alpha_{DD}$	0.35	0.13	0.39	0.13	0.00	0.19	0.19	0.43	0.14	0.05
	Mean: 0.70 [0.0067]					Mean: 0.23 [0.0487]				

Table 6: Frequencies of relative feedback (hypothetical messages) from the group of cooperators and the group of defectors about how cooperation is evaluated relative to defection (keeping the player's action constant). The first of the two actions of an outcome always corresponds to the sender of the message. In brackets, the one–sided p–values of the Wilcoxon signed–rank tests that analyze whether the mean of the distribution is different from zero.

It can be seen from the first line, for example, that 65 % of the cooperators approve cooperation and disapprove defection (ranking +2) provided that they themselves cooperate, whereas 43 % of the defectors express these ranking in the hypothetical case that they cooper-

ate. Also, the means of the distributions indicate, as we have already seen before, that actual cooperators are more likely to rank cooperation higher than defection, especially in case they cooperate. In this sense, the results from this table confirm our previous analysis. □

So far, we have found some evidence in line with the hypothesis that some individuals are disapproval–averse. Moreover, we have observed that cooperation is more likely to be approved than defection (in particular among the cooperators) and that cooperators are more likely to provide feedback if it is costly. Also, if Tables 1 and 4 are compared, it can be observed that expectations are on average rather correct. In particular, subjects anticipate in the expectations treatment correctly that cooperators are more likely to be approved than disapproved while defectors are more likely to be disapproved than approved. In what follows, we investigate whether the possibility to send feedback increases cooperation and, if so, why that occurs. More precisely, we test here the awareness and the NIA hypotheses presented in Section 3. Each of these assumptions predicts a rise in cooperation in the feedback treatment, although their predictions differ for the expectations treatment. Thus, the idea that subjects want to avoid information about bad events (the NIA hypothesis) implies the following pattern for the cooperation rate (note that subjects do not receive information about disapproval in the expectations and control treatments):

$$\text{Feedback} > \text{Expectations} = \text{Control}.$$

In contrast, if the feedback mainly works by making subjects aware of the other player's opinion (the awareness hypothesis), we should observe the following result:

$$\text{Feedback} = \text{Expectations} > \text{Control}.$$

To understand this prediction, note that subjects in the expectations treatment must think

about their co–player's opinions because their expectations in this respect are elicited.[11] In the feedback treatment, in turn, subjects should also think about their co–player's opinions because they are explicitly told in the instructions that the co–player will afterwards have the opportunity to submit a message announcing their opinion (approval/disapproval). Our implicit assumption is that subjects become equally aware about their co–player's opinion in both treatments.

Obviously, it is also possible that both the awareness and the NIA factor play an individually significant role in explaining the feedback effect. In this case, the following pattern with respect to the cooperation rate should be observed:

$$\text{Feedback} > \text{Expectations} > \text{Control}.$$

Finally, suppose that neither factor matters on its own but they are *jointly* significant. Then, we should observe more cooperation in the feedback treatment than in the control treatment, while, at the same time, all comparisons with respect to the expectations treatment turn out to be insignificant; that is,

$$\text{Feedback} > \text{Control}; \text{Expectations} = \text{Control}; \text{Expectations}=\text{Feedback}.$$

Our final result shows that the data indeed coincides with this very last case.

Result 3 (NIA & Awareness): *The following pattern is observed: The cooperation rate in the feedback treatment is larger than cooperation rate in the expectations treatment, which, in turn, is larger than the cooperation rate in the control treatment. However, only the difference between the feedback and the control treatment is significant.*

[11] Asking players about the co–player's opinion might have a potential side effect, as they might think as well about how the co–player would choose. This is because expectations about approval/disapproval are somehow based on expectations about choices. Now, some evidence suggests that players change their behavior if they become more aware about the co–player's expected choice (see Croson, 2000). This could affect our estimation of the awareness factor, but it seems impossible at the present to separate both phenomena.

Evidence on Result 3: Figure 3 below shows the cooperation rate for each of the three treatments. It is highest in the feedback treatment (38.33 %), followed by the expectations (29.03 %) and the control treatment (22.41 %). Hence, the mere possibility to send a costly message that approves or disapproves the choice of the other player increases the overall level of cooperation. Further, asking subjects to evaluate their own behavior in the eyes of the other player (the expectations treatment) also increases cooperation with respect to the control treatment but less than the feedback treatment does. Also remember that the strictly positive cooperation rate in the control treatment is consistent with the DA hypothesis.

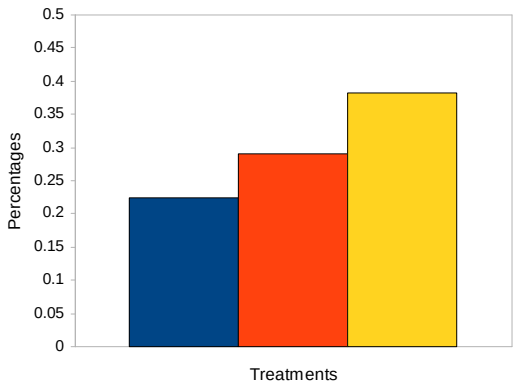


Figure 3: Levels of cooperation across treatments. To the left, the control treatment; in the middle, the expectations treatment; and to the right, the feedback treatment.

Using Mann–Whitney U tests, we obtain that the level of cooperation in the feedback treatment is significantly higher than in the control treatment (p=0.0461, one–sided). On the other hand, the level of cooperation in the expectations treatment is not significantly different from either the control treatment (p=0.2059, one–sided) or the feedback treatment (p=0.1864, one–sided). □

The fact that the cooperation rate increases across treatments suggests that both the NIA and the awareness assumption play a role. However, the effect of these forces is only significant

when they act jointly (the difference in cooperation rates is significant only between the control and feedback treatments). It seems, therefore, that the availability of approval/disapproval affects behavior because it makes players think about the other player's opinion and because players do not like to be effectively informed about disapproval.

6 Conclusion

In this paper, we have studied why individuals behave more cooperatively when their co–players have the possibility to approve/disapprove their actions. Our main working hypothesis was that some individuals are disapproval–averse, meaning that they feel badly if others think badly of their behavior. It turned out that disapproval–aversion is able to explain why some individuals cooperate in the absence of monetary incentives to do so, however, at the same time it is not sufficient to establish why the level of cooperation is higher in the presence of non–material sanctions/rewards. In effect, since feedback is often provided after choices have been made, it cannot influence ex ante expectations about disapproval (which is the decisive variable). As a possible explanation for recent experimental findings, we have therefore considered two additional factors: First, the mere availability of the feedback might help some subjects to focus on the other player's opinion, and second, people might be averse to effectively receive negative information (maybe because they feel regret in that case).

To distinguish between these factors, we ran three different experimental treatments: In the control treatment, subjects played a standard prisoner's dilemma game; in the expectations treatment, subjects were asked in addition –for every possible contingency of the game– how their co-player would think about their behavior before taking a definite action; finally, in the feedback treatment, subjects were allowed to approve/disapprove the action of their co–player. Our main experimental results have been as follows: *First*, the experimental data

is in line with the assumption that some players are disapproval–averse, because the cooperation rate in the control treatment is strictly positive and in the expectations treatment, most cooperators expect cooperation to be approved and defection to be disapproved; *second*, in the feedback treatment, actual cooperators approve/disapprove potential cooperators/defectors more than actual defectors do; and *third*, since the cooperation rate in the PD is highest in the feedback treatment, lowest in the control treatment, and reaches a medium level in the expectations treatment but only the difference between the feedback and the control treatment is significant, it is suggested that the possibility to give feedback fosters cooperation because it increases awareness *and* because players dislike receiving negative information.

The analysis here suggests several lines for future research. First of all, more experimental research is warranted; in particular, it seems important to study what kind of behaviors people actually approve/disapprove, but also how expectations about this are formed. Second, our model could be used to explain other experimental facts, like the role of anonymity in dictator games and public good games (Hoffman et al. 1994, Bohnet and Frey 1999, Dana et al. 2006, and Tadelis 2008). Further, one could incorporate additional motivations like inequity aversion (as in Fehr and Schmidt 1999) into the model in order to explain other important experimental phenomena like the use of material sanctions, conditional cooperation, *etc.* In addition, the model could also be extended to account for the fact that we particularly care about the approval or disapproval from close relatives and friends, and less from distant others (*i.e.*, the parameter γ_i could depend on the co–player's identity). Finally, one could apply the ideas here to study why awards by companies and governments affect behavior (Frey, 2007), or to explore the role of approval and disapproval in charity giving, team production, union formation, voting, or crime, to cite a varied range of problems where non–material sanctions and rewards are thought to play an important role (Gächter and Fehr 1999).

References

1. Arrow K (1974). The limits of organization. Norton & Company.

2. Battigalli P and M Dufwenberg (2007). Guilt in games. American Economic Review Papers & Proceedings 97: 170–176.

3. Bohnet I and B Frey (1999). The sound of silence in prisoner's dilemma and dictator games. Journal of Economic Behavior and Organization 38: 43–57.

4. Camerer C (2003). Behavioral game theory: Experiments in strategic interaction. Russell Sage Foundation, Princeton University Press.

5. Croson R (2000). Thinking like a game theorist: Factors affecting the frequency of equilibrium play. Journal of Economic Behavior and Organization 41: 299–314.

6. Dana J, D Cain, R Dawes and M Robyn (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. Organizational Behavior and Human Decision Processes 100: 193–201.

7. Dugar S (2008). Nonmonetary sanctions and rewards in an experimental coordination game. Mimeo.

8. Ellingsen T and M Johannesson (2008). Anticipated verbal feedback induces pro-social behavior. Evolution and Human Behavior 29: 100–105.

9. Elster J (1989). Social norms and economic theory. Journal of Economic Perspectives 3: 99–117.

10. Falk A, E Fehr and U Fischbacher (2005). Driving forces behind informal sanctions. Econometrica 7: 2017–2030.

11. Fehr E and U Fischbacher (2004). Social norms and human cooperation. Trends in Cognitive Sciences 8: 185–190.

12. Fehr E and S Gächter (2000). Cooperation and punishment in public goods experiments.

American Economic Review 90: 980–994.

13. Fehr E and K Schmidt (1999). A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114: 817–868.

14. Fehr E and K Schmidt (2006). The economics of fairness, reciprocity and altruism – Experimental evidence and new theories (eds. S Kolm and J Ythier). In: Handbook of the Economics of Giving, Altruism and Reciprocity 1, Elsevier.

15. Festinger L (1954). A theory of social comparison processes. Human Relations 7: 117–40.

16. Fischbacher U (2007). Z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10: 171–178.

17. Frey B (2007). Awards as compensation. European Management Review 4: 6–14.

18. Gächter S and E Fehr (1999). Cooperative action as a social exchange. Journal of Economic Behavior and Organization 39: 341–369.

19. Geanakoplos J, D Pearce and E Stacchetti (1989). Psychological games and sequential rationality. Games and Economic Behavior 1: 60–79.

20. Hoffman E, K McCabe, K Shachat and V Smith (1994). Preferences, property rights and anonymity in bargaining games. Games and Economic Behavior 7: 346–380.

21. Holländer H (1990). A social exchange approach to voluntary cooperation. American Economic Review 80: 1157–1167.

22. Homans G (1961). Social behavior: Its elementary forms. Harcourt, Brace & World, New York.

23. Kandel E and E Lazear (1992). Peer pressure and partnership. Journal of Political Economy 100: 801–817.

24. Loomes G and Sugden R (1982). Regret theory: An alternative theory of rational choice under uncertainty. The Economic Journal 92: 805–824.

25. Masclet D, C Noussair, S Tucker and M-C Villeval (2003). Monetary and non-monetary

punishment in the voluntary contributions mechanism. American Economic Review 93: 366–380.

26. Noussair C and S Tucker (2005). Combining monetary and social sanctions to promote cooperation. Economic Inquiry 43: 649–660.

27. Ostrom E, J Walker and R Gardner (1992). Covenants with and without a sword: Self-governance is possible. American Political Science Review 86: 404–417.

28. Peeters R and M Vorsatz (2009). Immaterial rewards and sanctions in a voluntary contribution experiment. METEOR research memorandum 09/005, Maastricht University.

29. Rege M and K Telle (2004). The impact of social approval and framing on cooperation in public good situations. Journal of Public Economics 88: 1625–1644.

30. Sefton M, R Shupp and J Walker, (2007). The effect of rewards and sanctions in provision of public goods. Economic Inquiry 45: 671–690.

31. Tadelis S (2008). The power of shame and the rationality of trust. Mimeo.

32. Vyrastekova J and D van Soest (2008). On the (in)effectiveness of rewards in sustaining cooperation. Experimental Economics 11: 53–65.

33. Xiao E and D Houser (2009). Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange. Journal of Economic Psychology 30: 393–404.

Appendix A: Probit Estimations

Here, we present a linear regression model as an alternative argument for the non–parametric Mann–Whitney U tests reported in the main text in order to analyze how expectations about approval/disapproval and the feedback provided differ across the group cooperators and the group of defectors. The dependent variable Y_i is a dummy variable that takes the value 1 if subject i cooperates and the value 0 if subject i defects. Depending on the treatment, we use for each outcome either expectations or hypothetical messages as regressors. To be more precise, the variable $X_{(s_i,s_j)}$ corresponds to the case when subject i takes action s_i and her co–player takes action s_j. Then, $X_{(s_i,s_j)}$ takes the value -1 if, in the expectations treatment, subject i expects that s_i is disapproved by j or if, in the feedback treatment, subject i hypothetically disapproves s_j. The variable takes the value 0 (+1) in case of a neutral (positive) assessment. Moreover, if ε_i is the error term for individual i (distributed normally and independently with mean zero), the model is completely specified as follows:

$$Y_i \;=\; \beta_0 \,+\, \beta_1\, X_{(C,C)} \,+\, \beta_2\, X_{(C,D)} \,+\, \beta_3\, X_{(D,C)} \,+\, \beta_4\, X_{(D,D)} \,+\, \varepsilon_i. \tag{1}$$

	Expectations	Feedback	
		Hypothetical	Actual
Constant	0.3046	-0.8051	3.2761^{**}
	(1.3828)	(1.4022)	(1.5807)
$X_{(C,C)}$	0.9244^{***}	0.2980	1.1602^{**}
	(0.3248)	(0.4199)	(0.5586)
$X_{(C,D)}$	0.0939	-0.3577	1.1785^{**}
	(0.2477)	(0.3513)	(0.5061)
$X_{(D,C)}$	0.2993	0.0752	...
	(0.2515)	(0.2233)	...
$X_{(D,D)}$	-0.5132^{**}	0.0592	-0.3744
	(0.2651)	(0.2704)	(0.8474)
R^2	0.1925	0.0632	0.1692

Table 7: Probit Maximum Likelihood estimation on the decision of whether or not to contribute in the expectations and the feedback treatment. The first of the two actions of an outcome always corresponds to player who assesses the belief (who sends the message). Standard errors are in parenthesis. Errors are robust to heteroskedasticity. *** indicates significance at the 1–percent level. ** indicates significance at the 5–percent level. * indicates significance at the 10–percent level.

Table 7 shows the results of the Probit Maximum Likelihood estimations of Equation 1 with the errors being robust to heteroskedasticity. We also controlled for age and gender, but neither of the two variables turned out to be significant. Consequently, we eliminated them from our final specification. Regarding the expectations treatment, it can be observed that those subjects who expect to be more positively evaluated after outcome (C,C) are also more likely to cooperate (the sign of β_1 is positive and highly significant). Similarly, the subjects who believe to be more positively evaluated after outcome (D,D) are more likely to defect (the sign of β_4 is negative and significant). Finally, the expectations related to the outcomes (C,D) and (D,C) do not influence the probability that a subject cooperates. Consequently, the conclusions drawn from the linear regression model coincide with the Mann–Whitney U tests reported at the end of Result 1.

We also see that hypothetical message are no indicator of whether a subject cooperates in the feedback treatment. Since the Mann–Whitney U tests at the end of Result 2 evidence that cooperators provide a more positive feedback than defectors after outcome (C,C) and a more negative feedback after outcome (C,D), the two approaches yield clearly different results. However, we can recover the original insights if we consider actual messages instead of hypothetical ones; that is, we use as regressors dummy variables that take the value 1 if subject i actually sent a message to j after outcome (s_i, s_j) and the value 0 otherwise.

The last column of Table 7 shows that subjects who are more prone to send a message after outcome (C,C) are also more likely to cooperate. Since we have seen before that messages are in this case mostly positive, there is a positive relationship between approving the co–player's action after this outcome and the probability that the subject cooperates. Similarly, there is also a positive relationship between the likelihood of sending a message after outcome (C,D) and the probability that the subject cooperates. However, here messages are predominantly negative and, therefore, we conclude that subjects who disapprove their co–player after outcome (C,D) are more likely to cooperate. Third, sending a message after outcome (D,D) is not correlated to the probability to cooperate. Finally, the dummy variable associated with outcome (D,C) had to be eliminated from the regression since it is co–linear with $X_{(C,D)}$. Using it as the only regressor does not show any correlation with the probability to cooperate, its parameter estimate is 0.3028 and the standard deviation equals 0.5317. Hence, the interpretation of this final regression fully coincides with the insights from Mann–Whitney U tests provided in the main text.

Appendix B: Instructions of the Feedback Treatment

Welcome

Dear participant, thank you for taking part in this experiment. It will last about 60 minutes. If you read the following instructions carefully, you can – depending on your decisions – earn some money. The entire of money which you earn with your decisions will be paid to you in cash at the end of the experiment. These instructions are solely for your private information.

We will not speak of Euros during the experiment, but rather of ECU (Experimental Currency Units). Your whole income will first be calculated in ECU. At the end of the experiment, the total amount you have earned will be converted to Euro at the following rate:

20 ECU = 1 Euro.

In order to ensure that the experiment takes place in an optimal setting, we would like to ask you to abide by the following rules. If you do not obey them, we will have to exclude you from this experiment and you will not receive any compensation.

- Do not communicate with your fellow students. If you have any doubts, raise your hand and one of the experimenters will clarify them privately.
- do not forget to switch off your mobile phone!
- you may take notes on this instruction sheet if you wish.
- when the experiment finishes, remain seated till we pay you off.

The Experiment

In the experiment you will participate in six different scenarios, and you will be paid for your decisions in one scenario, randomly chosen at the end of the experiment. More precisely, the participant playing at the computer number 9 will roll a die and we will pay you the equivalent in Euros of your ECU earning in the scenario corresponding to the number that turns up.

In what follows we will explain to you only the first scenario. Once you made your decision in this first scenario, we will introduce the five remaining scenarios. Note well that each scenario is independent of the others; that is, your payoff in any scenario does not depend on decisions taken in other scenarios.

Scenario 1

In this scenario, you have been randomly and anonymously matched with another participant and both of you have to choose independently between alternative X and alternative Y. Depending on your choices, you will get the following ECU payoff:

- if you both choose X, both of you get 180 ECU.
- if you choose X and the other participant chooses Y, you get 80 ECU and the other participant gets 260 ECU.
- if you choose Y and the other participant chooses X, you get 260 ECU and the other participant gets 80 ECU.
- if you both choose Y, both of you get 100 ECU.

The matrix below summarizes this.

		The other player	
		X	Y
You	X	(180,180)	(80,260)
	Y	(260,80)	(100,100)

Decisions at this scenario will be private; that is, you will never be informed about the decisions of any other participant in this scenario, and no other participant will know your decision in this scenario. Apart of choosing between X and Y, in this scenario both of you have the possibility to send one message to the other participant with your opinion about her/his choice. Sending a message costs 10 ECU. Since decisions are private, however, you will not know whether the other participant chose X or Y. For this reason, we will ask you whether you want to send a message for any possible contingency. More precisely, the procedure will consist of three steps:

1. You will be asked the following question: "Suppose both of you chose X. Do you want to send a message paying a cost of 10 ECU?" You can choose either *Yes* or *No*.

2. Independently of your answer to the previous question, you are then asked the following question: "Suppose you decided to send a message to the other participant in case both of you chose X. Which of the following three messages do you send?"

   - Your choice was good.
   - Your choice was neither good nor bad.
   - Your choice was bad.

3. The same two previous steps are then repeated for any of the other three possible combinations of choices: (X, Y), (Y, X), and (Y, Y).

Observe again that decisions are always private; that is, none of you will know the choices of the other participant when going to the next scenario (including messages). The actual decisions in one scenario will be anonymously revealed to both participants only if, at the end of the experiment, it turns out that this scenario is randomly chosen for payment. If the die selects scenario 1, moreover, each participant will receive the message selected by the other participant at that scenario (step 2) only if the other participant previously chose *Yes* in step 1. Finally, observe that once the first scenario has finished, you will be matched to a different participant for the remaining five scenarios.

Control Questions

Please answer the following control questions. Once you have written down all your answers, please raise your hand so that one of the experimenters can check them.

1. How many different scenarios are there?
2. If you choose X and the other participant chooses X, what will be your payoff?
3. If you choose X and the other participant chooses$Y$, what will be your payoff?
4. Are you always matched with the same participant?
5. How will your final payoff (in Euro) be determined?