Minnameier, Gerhard; Bonowski, Tim Jonas

Conference Paper

Morality and Trust in Impersonal Relationships

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Morality and Trust in Impersonal Relationships

Abstract

Experiments with the trust game reveal that people are generally more trusting and trustworthy than can be explained by profit maximizing strategies. Usual attempts to explain this social preferences. Contrary to such a view, we suggest that trust and trustworthiness depend on a specific kind of morality that functions as a social norm and therefore comes with specific moral punishments and rewards. Since these moral incentives are confounded with material incentives in the trust game, an augmented version of the game is used, in which moral communication without material consequences is possible. In comparison with the original trust game, we find higher levels of both trust and trustworthiness in the augmented game. We also find that these levels remain stable across ten rounds of randomized anonymous interaction.

1 Introduction

Trust is an important condition for human cooperation in general and business in particular. Owing to (necessarily) incomplete contracts, market transactions require trusting and trust-worthy trading agents (Arrow, 1974). Accordingly, Bruni and Sugden consider trust and trustworthiness as "market virtues" (2013, p. 155f.). There is a long list of economists pointing out the importance of trust for market transactions and, perhaps surprisingly, the commerce as an antecedent of trust and honesty. Adam Smith argued that the independence of agents brought about by markets

"increases the honesty of the people" and "is the best police for preventing crimes" (1896, p. 155). A similar argument can be found in Marshall's *Principles of Economics*, wherein he stated that

"the modern era has undoubtedly given new openings for dishonesty in trade" (1920, p. 6) but added that new and extended opportunities wound not automatically lead to more fraud and other forms of defection: "On the contrary, modern methods of trade imply habits of trustfulness on the one side and a power of resisting temptations to dishonesty on the other" (ibid.).

About 25 years ago, Berg, Dickhaut, and McCabe (1995) invented the "trust game" which is played sequentially by two players. Both players are endowed with $10. The first mover (i.e., the trustor) has to decide how much (or all, or nothing) of her $10 to transfer to the second mover (i.e., trustee). Whatever is transferred is tripled in that process, so that the second mover has at her disposal her initial endowment plus three times the amount transferred by the first mover. The second mover can send back any amount available but may also decide to keep it all. Berg, Dickhaut and McCabe found that most trustors actually send money and that also most trustees send money back. More than one third of the trustors who sent money received more in return that they had sent. Even though this supports the above-mentioned line of reasoning, it is surprising, since the interaction was anonymous (double-blind) and one-shot (i.e., non-repeated). Hence, social controls were ruled

out. Under such circumstances, classical rational choice theory predicts that trust would either not appear or be crowded out: Since the trustee could always keep the amount received, the trustor, knowing this, should not send anything. The trust game has since been played as part of a multitude of experiments, with the results consistently violating this pessimistic prediction. Experimental research revealed that trust and trustworthiness pervade social life, in particular in developed countries. In a meta-study including 162 replications of the trust game, Johnson and Mislin (2011) find that trustors on average send .5 of their endowment to the trustees (varying between .22 and .89; SD = .12), and that the trustees send back .37 of the amount available to them (varying between .11 and .81; SD = .11). They also find that trustors in Africa and South America send significantly less compared to trustors in Europe and North America (the difference being roughly equal to half of a standard deviation). Trustworthiness is particularly low in Africa, where trustees send back .319 as compared to .34 in North America, .369 in South America, and .382 in Europe) (ibid., p. 873).

These results confirm the assumption that real people are both trusting and trustworthy. They also suggest that living in a modern market economy fosters rather than hampers the development of trust and trustworthiness. Moreover, Sutter and Kocher (2007) have found that trust and trustworthiness increase almost linearly with age,[1] which reveals that learning and life experience does not take agents towards the inefficient Nash equilibrium of zero trust, but in the opposite direction.

Although the evidence is compelling, it still seems difficult to make sense of it from the point of view microeconomics. After all, the trustor faces a fundamental dilemma. If both players were strictly self-interested, the trustee would send nothing back, so that the trustor sending money would be clearly irrational. What then is it that motivates trustors to trust and trustees to behave trustworthily?

Common explanations point to social preferences like altruism and kindness (Cox, 2004; Ashraf, Bohnet, & Piankov, 2006) or inequity aversion (Fehr & Schmidt, 2006) and related biological processes (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). However, these explanations are themselves question-begging, because if true, we would still wonder why people develop and uphold these preferences, given the objective constraints (cf. Binmore, 2010b). There is also massive evidence showing that most people try to go back on their moral duties whenever possible, and anonymity or general lack of social control allows them to do so. The phenomenon is known as "moral hypocrisy" (Batson, Thompson, Seuferling, Whitney, & Strongman, 1999; Lönnqvist, Irlenbusch, & Walkowitz, 2014; Rustichini & Villeval, 2014), but also familiar from research on the dictator game and moral wriggle room (in particular Hoffman, McCabe, & Smith, 1996; Dana, Weber, & Kuang, 2007; Andreoni & Bernheim, 2009). According to Dana, Weber and Kuang,

"(r)ather than having a preference for a fair outcome, people may conform to situational pressures to give in certain contexts but may also try to exploit situational justifications for behaving selfishly" (2007, p. 69).

[1] Trust only decreases among retired people whose average age was 68 years. Since they are more trustworthy than any other group, their diminished level of trust may be due to reservations towards the young generation and a resulting anxiety of being betrayed.

Hence, the question is what stabilizes and destabilizes trust and trustworthiness and how we can foster trustful relationships within organizations and economic exchange in general. In the following second section, we will first explicate our theoretical approach, which deviates markedly from the social preference model. We explain interpersonal trust and trustworthiness in terms of a specific kind of morality, i.e., a specific moral principle that guides the interaction. However, in doing so we do not consider moral principles as (fundamental) preferences in the sense of internalized moral values, but as social rules operative at the social level. We then explain how moral rules shape and transform the situation, in particular in transforming mixed-motive games (Schelling, 1960) into coordination games. According to the theoretical approach we present, moral principles function as institutions.

Transforming games with the help of institutions is not a new idea. However, economists almost exclusively rely on material rewards and punishments to change the payoffs of games and thus change them. Conversely, we aim at rewards and punishments in moral currency, which do not change the monetary or material payoffs. One might consider this as a weak form of incentives. However, given that material incentives are often counter-productive and erode moral cultures (see Bowles & Polanía-Reyes, 2012, for an overview), such forms of incentives could be even fairly strong.

In the third section, we will introduce our experimental, the methods, sample, and result. We conducted an experiment based on the trust game, which we augmented by offering the opportunity for moral communication (called the "trust and trustworthiness game") – a communication, however, that is cheap talk with respect to material consequences. Nonetheless, we observe an increase of trust and trustworthiness in comparison with a control group playing an ordinary trust game. Additionally, there are signs that this higher level of trust and trustworthiness remains stable over the course of repeated games. The final section summarizes, concludes and points to important implications for management in general.

2 Trust in the context of game theory

As mentioned in the introduction, the relevance of trust for trade, today, is almost a truism and is echoed in the literature (see also Arrow, 1972; Fukuyama, 1996; Kosfeld et al., 2005; Algan & Cahuc, 2013; Bottazzi, Da Rin, & Hellmann, 2016). However, it remains unclear how trustful interaction is established and stabilized, in particular in non-repeated encounters. As Arrow states, "virtue may not always be its own reward, but in any case, it is not usually bought and paid for at market rates" (1972, p. 346). Some think that trust and trustworthiness, like other virtues, are acquired through cultural socialization and incorporated in people's personalities (at least among decent and well-behaved people). This is what modern virtue ethicists like Michael Sandel (2009, 2012) believe and, as mentioned in the introduction, this view is also reflected by theories on social preferences.

2.1 Problems with explanations of trust and trustworthiness

However, there seem to be serious and principled flaws in all these approaches. Binmore, for instance, holds that rather than overcoming the all too narrow framework of neoclassical economics – which behavioral economists claim they do – "theories of other-regarding utility functions might better be classified as retroclassical economics, as they revert to the dogma that people actually do have utility functions in their heads that they seek to optimize when interacting with others – an idea that was popular in Victorian times, but which was abandoned by neoclassical economists many years ago" (2010b, p. 150). Against this view, Binmore proposes social norms instead of social preferences.

In contrast to the latter, social norms are institutions (Binmore, 2010a). Therefore, they are *social* entities rather than individual traits. Why is this important? Consider the phenomenon of "moral hypocrisy" (Batson et al., 1999; Rustichini & Villeval, 2014). When people act in anonymous contexts or when their actions cannot be monitored, they are more likely to make self-interested choices.[2] This has also been revealed in seminal studies on the dictator game with a so-called "plausible deniability" condition, where a computer interferes with a certain probability and overrides the proposer's choice. It was shown that if this probability is 50% or higher, and if the computer always allocates the full endowment to the proposer, then at least two thirds of the proposers choose to keep the entire endowment. It therefore seems clear that, even if people actually have social preferences in the sense described, these preferences alone do not suffice to uphold morality in real life beyond a small group of moralists resilient to temptation.

What we observe is obviously different, but we do not (yet) know how it works. What we know from relating research is that moral appeals might function to some extent, but are also likely to wear out overtime (Ariely, 2012; Danese & Mittone, 2018). Apart from moral suasion, one could also rely on monitoring, which would be the strategy recommended by classical management theory on how to deal with moral hazards. However, in Danese and Mittone's study, monitoring has a very limited effect and only on trust, but not on trustworthiness. Outside of the laboratory, monitoring and enforcement are costly. And on top, it is known that monitoring can even reduce compliance (Falk & Kosfeld, 2006), and that material incentives can crowd out moral motivation (see Bowles & Polanía-Reyes, 2012, for an overview). Hence, if we do not want to throw out the baby with the bath water, we will have to find out, how morality really works and how this can be reinforced and employed in management approaches.

2.2 Morality and game theory

Folk wisdom takes morality in terms of zero-sum games, wherein moral agents make sacrifices for the sake of others. However, this is an oversimplification. The assumption of social preferences on the part of the former in

[2] Batson holds that moral hypocrites try to avoid bearing the costs of their morality, if they can manage to appear moral to others and to themselves. This implies that they might not be selfish, inasmuch as they deceive themselves. However, Lönnqvist et al. (2014) showed that moral hypocrites basically deceive others rather than themselves and therefore would have to be classified as selfish on this account.

order to rationalize their giving behavior is yet another oversimplification which regularly fails when anonymity is introduced (see e. g. Hoffman et al., 1996; Dana et al., 2007; Andreoni & Bernheim, 2009). Therefore, it seems more suitable to understand morality in terms of positive-sum games, and in particular in terms of mixed-motive games like the prisoners' dilemma (see Bicchieri, 2006).

There has been a (Humean) tradition in economics that stresses that morality has to pay off somehow for the individual, if agents are to follow it (Gauthier, 1986; Binmore, 1994, 1998, 2005; Sugden, 1986/2005, 2018; Bicchieri, 2006). The idea that moral principles play the role of institutions is most salient in Binmore's and Bicchieri's work, but Robert Sugden had used the Hawk-Dove game even earlier to show how a simple property rule can be introduced as an institution to transform the game, which is a mixed-motive game, into "Hawk-Dove-Property", which is a coordination game (1986/2005, pp. 61–74; see also Gintis 2014, pp. 145–146).

In Hawk-Dove, players compete over a resource of value v and can either fight, i.e., play "hawk" (H), or relinquish the resource if a becomes imminent, i.e., play "dove" (D). H may look favorable, but it entails the risk of being injured. Let the value v of the resource be $v = 20$ and the cost c of being injured $c = 40$. Let us further assume that both have an equal chance of winning the fight. The resulting payoffs are shown in the right-hand panel of Figure 1.

Unlike the prisoners' dilemma, Hawk-Dove does not have a stable Nash equilibrium in pure strategies. (H, D) and (D, H) are pure strategy Nash equilibria, but they are asymmetric and therefore not stable. However, there is also a symmetric equilibrium in mixed strategies, in which each player chooses H and D with probability $p = .5$. In this case, the expected payoff to each player is 5. Similar to the prisoners' dilemma, the players end up with expected earnings of (5, 5) where (10, 10) are possible, at least in principle.

	D	H
D	$v/2$, $2/v$	v, 0
H	0, v	$(v-c)/2$, $(v-c)/2$

a

	D	H
D	10, 10	20, 0
H	0, 200	−10, −10

b

Figure 1: The hawk-and-dove game with (a) payofs in general form and (b) payoffs if $v = 20$, $c = 40$.

A simple property rule P stating that the one who reaches the resource first is entitled to use it, allows the agents to solve the problem. It simply means that the players choose H when first and D when second. This allows them to avoid conflict and reap an expected payoff of (10, 10) (see also Gintis, 2014, pp. 145–146). Figure 2 shows the result. For reasons of simplicity, only the row-player's payoffs are displayed.

If the column-player chooses P, the row-player's best response is also P. Apart from the fact that the augmented Hawk-Dove-Property game is clearly a coordination game, we can also see that it implies a punishment mechanism, which consists in defending one's "property", when being the owner and the other player attacking. For an illustration, let us assume the column-player has been brought up to be always nice to others and never engage in fighting and so exclusively plays D. The behavior would make them vulnerable to exploitation and it would be an invitation for the row-player to exclusively choose H. Their behavior would lead to an unraveling of the property rule.

	P	D	H
P	$v/2$	$3v/4$	$(v-c)/4$
D	$v/4$	$v/2$	0
H	$3v/4 - c/4$	v	$(v-c)/2$

a

	P	D	H
P	10	15	−5
D	5	10	0
H	5	20	−10

b

Figure 2: The Hawk-Dove-Property game with the payoffs for the row-player: (a) in general form and (b) if $v = 20$, $c = 40$.

Rules like the property rule are social norms that evolutionarily stable, meaning they are resistant to subversion (Yee, 2003, p.185–186). Now, if we consider property as a *moral* rule, which it certainly is (whether we understand it as genuine right or as a contingent agreement), we can observe several important consequences. (1) Moral rules function as solution concepts for mixed-motive games (in the sense of turning them into coordination games). (2) Since the base game is not a coordination game, such rules have to come with sanctions. These do not have to be enforced by additional institutions, in the Hawk-Dove-Property game they are enforced within the game itself which makes P evolutionarily stable. (3) Not using the sanctioning mechanism, i.e., choosing D invariably in all encounters, would in fact not indicate a nice and good-natured person but it would corrupt morality (in the sense of an invitation for the other to choose H and thus destroy coordination).

2.3 Moral rules and moral currencies

Although norms like the property rule solve the cooperation problem, the above analysis they entail some caveats. If we have it that players "fight for their right" when the property rule is violated, they will have to resort to material incentives to implement a punishment mechanism. However, above we stated that we want to discuss specifically *moral* rewards and punishments, and these were described as non-material. Accordingly, what we are looking for, must be of a different kind.

In fact, we suggest that ordinary moral rules come with other kinds of incentives (although material incentives might be allowed as a last resort). Furthermore, specific moral rules appear as solution concepts for specific mixed-motive games. As Curry, Whitehouse, and Mullins (2019) have shown, this allows us to understand how basic moral concepts work in virtually all cultures around the world.

The prisoners' dilemma is a case in point. It constitutes a social dilemma, because the Pareto-superior strategy combination cannot be implemented, even though it would make both parties better off than they are in the Nash equilibrium. This could be changed, if the prisoners were allowed to talk to each other before making their choices and follow up with an enforcement mechanism. This modification not only introduces an essential precondition, but for the game to change substantially the players also need a specific understanding of the situation. Let us say, they are not close friends or relatives, but mutually disinterested. In this case, it is crucial that they understand the situation such that it is allows them to make a mutual promise, i.e., strike a contract. By agreeing on some course of action (here: to cooperate), they implement a moral rule, which obliges each agent to fulfill the contract. Doing so is not only rewarded by the benefits of cooperation, but also by being respected for one's honesty and reliability. Conversely, breaching the contract would be despicable.

Hence, being respected or held in contempt are the positive and negative sanctions that apply in this situation.[3] Overall, we consider three distinctive types of moral sanctioning mechanisms: *affection*, *respect*, and *repute*:

i. *Affection* refers to the interaction among individuals who have an interest in each other in the sense that they want to be loved or at least liked by the other. Hence, signs of affection (or of the withdrawal of affection) will reward (or punish) the agents, if they want to be loved or liked by the other. However, if the agent is disinterested towards the other person, those sanctions will not work.

ii. *Respect* applies to these disinterested relationships. In everyday encounters do not seek affection in the sense of *i.*, but would like to be respected and treated with respect. This is, e. g., why people tip waiters in faraway countries whom they will probably never see again (see Bicchieri, 2006, p. 43). In these relationships we do not expect favors but do expect fairness.

[3] The enforcement mechanism may also be intrinsic to the individual, e.g., in the form of self-concept maintenance, as we will discuss below.

iii. *Repute*, finally, is a moral currency that does not apply to (direct) inter-individual interaction, but to one's standing in a certain social frame of reference (which may be a working team, a company, a neighborhood, or another social entity to which one belongs). Specifically, a good or bad reputation does not depend on direct interaction (as it is true for respect), but the dissemination of information across the group (through gossip and other means of sharing information).

The order of these sanctioning mechanisms and the underlying moralities can be understood as concentric circles: The first – which is also called care ethics – applies to the inner circle of personal relationships among relatives and friends. The second goes beyond and applies to disinterested relationships with other individuals (where true conflicts of interest emerge). The third goes beyond the former two by applying to level of social entities like groups, teams, tribes and so on.

These moralities are expounded and discussed in more detail elsewhere (AUTHOR). For the present purpose it may suffice to concentrate on the interaction of mutually disinterested individuals, i.e., the second form of morality, which we think applies to trust in terms of the trust game.

2.4 The morality of interpersonal trust

The second kind of morality is based on *respect*. The trustor's trust indicates respect, just as the trustee's trustworthiness does. Although both benefit from cooperation, there is also a conflict of interest inherent in the trust game, especially if it is played one-shot. Why should the trustee return a positive amount, if the interaction is not repeated? And knowing this, why should the trustor send a positive amount in the first place? Provided there are no feelings of (mutual) affection involved, this constitutes a true conflict of interest. Hence, respect is needed in the sense that the agents take account of each other's interests, even though they do not benefit from doing so (as they would if they felt affection).

Basic respect is implied in accepting competition or, generally speaking, where it is accepted that everyone pursues their own interests. If agents' interests are in conflict, they might solve this conflict by striking a deal. In an ordinary commercial transaction, both parties agree on a certain price for a certain product. They show respect by fulfilling their respective duties out of fairness and respect, not as favors. The trust game, however, does not model such a deal, since players do not have a chance to agree on one. Accordingly, cooperation in the trust game cannot depend on a deal between the two players. Rather, the players have to strike deal with themselves, i.e., in the sense of the so-called *Golden Rule* ("Do unto others what you would have others do unto you"). Trustors should trust, because they would want to be trusted, if they were the trustees. Accordingly, trustees should be trustworthy (and send a fair amount back), because this is what they would want, if they were the trustors' shoes. However, for the Golden Rule to function in the trust game it is necessary that both players understand it and take it as a rule of the game.

Of course, we can also trust a shop or a company because they have a good reputation. And we may be convinced of their reliability because they would have a reputation to lose. In such cases, "trust" refers to a rational expectation, just as we might trust the weather forecast or any other kind of reliable prediction. However, a trustor's trust in the trust game does not imply the rational expectation that the trustee will be trustworthy and return a decent amount of money.

Dunning, Anderson, Schlösser, Ehlebracht, and Fetchenhauer (2014) have made this very point. Their claim – in the subtitle – is that trust is "More a Matter of Respect Than Expectation of Reward" (p. 122). On their account, respect relates to specific duties the trustors and trustees have in the trust game. The trustor has the moral duty to trust, because it would be disrespectful to presume that the trustee is not trustworthy. Conversely, the trustee has the moral duty to return a fair amount, because it would be disrespectful to abuse the other's trust.

Across a variety of treatments, Dunning et al. were able to show that trust depends on rational expectations (which would require knowledge about the trustee's reputation). Trustors consistently sent more money than they would, if their choice were merely based on their expected return. And they took a higher risk – as measured by their expectation about the trustee's behavior – than they would in an ordinary lottery. Moreover, when asked what they "wanted" to do as well as what they "should" do, the amount that trustors stated they *should* transfer always exceeded the amount that they actually *want* to transfer (Dunning et al., 2014, pp. 125–127). Hence, above and beyond their rational expectations, trustors' levels of trust are driven by duty rather than by desire. "People trust not because it is what they want to do but because they feel it is an obligation of their current social role" (ibid., p.124).

Dunning et al. define trust "as a psychological state comprising the intention to accept vulnerability based on the chance of reward from positive intentions or behavior of another" (p.123). Since this definition does not require the expectation of a monetary benefit, Dunning et al. see the main driver for trust in "a norm to respect another person's character and so trust others to ensure their social behavior aligns with that norm" (ibid.). However, they also think of this norm as a moral norm, as opposed to a social norm. And they explain that while social norms implied sanctioning mechanisms, moral norms did not (p.125). In other words, their view of moral norms is that of internalized values – as explained above – rather than of institutions.

We believe this is a mistake because, like most behavioral economists, Dunning et al. identify sanctions with material incentives. Against this view, we suggest that by showing trust (in terms of sending money to the trustee), trustors convey a signal of *showing respect* and that this signal itself functions as a positive incentive in this very sense. On our view, it is an incentive in moral currency. Therefore, moral norms can be social norms. Of course, moral norms require that the underlying moral content is shared by the agents, i.e., that everyone knows about their moral duties, but this only emphasizes the institutional role of morality in the game-theoretic frame of reference.

Dunning et al. show just this in their further studies, wherein they established that trustors' trusting behavior crucially depends on whether they can signal that they trust in the other person. For instance, their willingness to send money declines, if the other players make no morally motivated choice, but mere flips a coin to determine whether a fixed amount is sent back or not (ibid. pp.132–134). Being able to actually show respect is the main driver of cooperation above and beyond rational expectations of financial returns.

It should be highlighted that those studies investigated trustors' behavior, not that of trustees. This is important because it seems more commonsensical that trustees feel obliged to return a decent amount and commit themselves to the Golden Rule, than for trustors. However, the general result is that the same applies to trustors, since they obviously feel obliged to treat the other player as a respectful and trustworthy person (without knowing, whether she really is). Especially those trustors, who consider themselves respectful and trustworthy, would be obliged to treat trustees as if they were the same kind of person. Hence, the Golden Rule seems to be a suitable moral principle to govern the interaction for both sides, trustor and trustee (see also Costa-Gomes, Ju, & Li, 2019).

2.5 Establishing stable trust and trustworthiness

Since signaling respect is so important for players in the trust game, it may be asked *why* it is so important. One reason could be that they care about their self-image, especially if they have a pronounced moral identity (Blasi, 1983, 1993; Aquino & Reed, 2002; Reed, Aquino, & Levy, 2007; Mazar, Amir, & Ariely, 2008). This is what Dunning et al.(2014) seem to suggest, and we would not deny this. However, we think it is even more important to oblige the other player and to establish trust and trustworthiness as rules of the game. If our above analysis of moral principles as solution concepts for mixed-motive games is correct, there has to be a mechanism that overcomes the social dilemma (in which no money is transferred and no value added).

On this account, the Golden Rule appears as a suitable concept and respect as the moral currency for suitable moral incentives. Note that if this is the case, neither rewards nor punishments would have to come in monetary or material units. Does it suffice that both players have internalized the Golden Rule (or a similar moral rule like treating others with respect), and thus understand what the rule requires, to overcome the social dilemma and secure coordination? This is what we want to find out.

We expect that allowing for moral communication in the sense of sending simple feedback to indicate whether the other's behavior is deemed respectful and the amount transferred appropriate (or not) has a positive effect on cooperation in the trust game. Such communication allows the players to reward and punish each other. However, it would only work, if both understood the rules of the moral game. Otherwise the content of communication would be cheap talk as there are no material costs or consequences associated. Therefore, we would expect higher levels of cooperation, if moral and material incentives could be disentangled, so that players could issue specifically moral rewards and punishments.

Hypothesis 1. *Trustors with access to moral communication are more trusting as measured by their confidence to be treated fair and amounts sent to trustees.*

Hypothesis 2. *The availability of moral communication makes trustees more trustworthy as measured by their tendency to behave reciprocally when trust is extended.*

The litmus test is whether a certain level of cooperation, especially one that exceeds the players' initial rational expectations about returns from interaction, is upheld in a succession of one-shot games. In such a setting, players learn what to expect from other players and can adapt their beliefs, in following games. Moral communication helps in establishing shared behavioral expectations, even in impersonal and random interactions. We expect that without the possibility of giving feedback about the respectfulness of the other player, cooperation is vulnerable and will be reduced or even break down completely in the long run, because the only way to give negative feedback to non-cooperators is to reduce one's payments. Therefore, the crucial test for our theory is whether access to moral communication can stabilize cooperation over a series of one-shot trust games, where it diminishes or is crowded out completely, if moral communication is not possible.

Hypothesis 3. *Trustors' trust and trustees' trustworthiness remain stable under the condition of moral communication but trust deteriorates without it.*

3 Experimental Design and Data

To test our hypotheses, we conducted an experiment based on a repeated trust game (Berg et al., 1995). The experiment was conducted in August 2019 and March 2020 at a large German university. Subjects were recruited using an online recruitment system (Greiner, 2015) and the experiment was implemented using oTree (Chen, Schonger, & Wickens, 2016). Subjects were led into a room and seated in front of computers. All desks were equipped with visual separators so neighboring screens were not visible. An experimenter read out a standardized introduction and asked the subjects to carefully read a hardcopy of the instructions located on their desk. This printout explained the rules of the game, informed them about their role as trustor or trustee (labeled Participant A and Participant B), and that they would remain in this role for the remainder of the experiment. During the experiment, all payoffs were labeled as *points*. To ensure that subjects understood the rules, the hardcopy included a number of questions describing potential decision scenarios, asking the participants to determine the resulting payoffs. The experiment started once all subjects had correctly answered all questions.

In each session, players played ten trust games. For every game, they were paired with a random player. They received no information about the other player's identity. At the beginning of every round, both trustors and trustees received 40 points of experimental currency. Trustors were asked to send either 0, 10, 20, 30 or 40 points to their respective trustee. The amount of points sent to the trustee was tripled before it was received

by them. Trustees could then decide to send any positive integer amount of the points at their disposal back to the trustor. Pre-tests showed that some participants struggled with reliably calculating the resulting payoffs, so for some treatments we included a table with payoffs on the decision screen of trustees that updated on input. As this may be visual cue for equal outcomes, we did not include them for all sessions.
Games with the results display are coded as "calculator" treatments in the results section.

In addition to these decisions, subjects were asked to answer a number of questions before and after their decisions. Before deciding how much to send, trustors were asked how likely they thought it was that trustees would send a "fair amount" back. No additional information about fairness was primed and answers were selected on five-step linear scale ranging from 0% to 100%. Before trustees were informed about the amount their trustor had send them, they were asked to state how much they would hypothetically send back for each of the five decisions available to the trustor. It was clearly stated that these answers would be non-binding. Both the confidence stated by the trustor and the hypotheticals stated by trustees remained private and were not shared with other participants.

To investigate our main research questions, in half of our sessions, we offered an opportunity to communicate morally relevant information to the other player. In these sessions, subjects were asked to rate the other player's decision. To do so, they had to respond to the following two statements: "How would you describe Participant A (B)?", answered on a five-step Likert scale running from "Fair" to "Unfair", and "How would you describe the amount of points you received from Participant A (B)?", rated on a five-step scale from "Too little." to "Too much.". Responses from the other player were displayed alongside the payoff information on the results screen for every trust game. Results from sessions where players rated each other are coded as "moral communication" in the results section.

After the ten trust games, players were informed about their cumulative payoff and were asked to answer a number of additional questions about their demographics, to control for potential differences in trust associated with gender and age. To control for the effects of self-image concerns, subjects were asked to rate items from internalization scale of moral identity centrality (Aquino & Reed, 2002; Reed et al., 2007). Individuals high in moral identity centrality would be expected to show high levels of trustworthiness, even when their behavior remains unobserved. They also answered questions about their strategic approach, a 3-minute version of a German MiniQ (Baudson & Preckel, 2016) to measure verbal reasoning aptitude and a stepwise risk lottery to elicit risk preferences. The latter two measures did not enter the analysis reported in this paper.

At the end of each session, payoffs were calculated using a conversion rate of 0.03€ or 0.05€ per point and participants were asked into a separate room to receive their payoffs in cash. A total of eight sessions with between 16 and 24 participants were conducted. Of the 171 subjects, 87 stated they were male, 82 stated they were female and 2 gave their gender as Other. The average subject age was 24.36 (SD = 4.7).[4] Participants on

[4] Data from an 75-year-old participant were dropped from analysis, as it falls outside the working age demographic our study targets.

average scored 4.25 (SD = 0.56) on the moral identity internalization scale and earned an average of 27.48€ (SD = 10.30).

The experimental design allows us to test for the effects of moral communication by allowing for moral evaluation of the other's choice. Since this moral evaluation is carried out ex post, and since players are randomly matched in each round of game, it has no strategic value for self-interested agents. Additionally, there is no material cost associated with moral sanctioning and random, anonymous interactions offer no reputational incentives for cooperation.

4 Results

This section summarizes our main results. Confidence intervals at the 95% level are reported in square brackets and were generated through 5000 sample bootstrapping. Data were analyzed using R 4.0.3 (R Core Team, 2013). Linear Mixed-Effects Models were fitted using the lme4 package for R (Bates, Mächler, Bolker, & Walker, 2015).

4.1 Moral communication makes trustors more trusting

Trustors stated that they expected to receive back a fair amount of points with an average probability of 46.72% [42.12%, 51.33%]. A two-sample bootstrap sample comparison shows that in sessions with the possibility for moral communication, probabilities stated by trustors were on average 10.96% [2.10%,19.48%] higher than in the control group. Across all sessions, trustors on average sent more than half of their 40 points to trustees (M = 26.94[25.98,27.85]). When moral communication was available, trustors sent more points (M = 30.07[28.89,31.14]) than in control treatments (M = 23.95[22.57,25.39]).

This means that across all games, availability of moral communication was associated with trustors sending an additional 6.11[4.26,7.93] points to trustees, supporting *Hypothesis 1*. As amounts sent are not normally distributed, differences in means must be interpreted with caution.

A Wilcoxon text shows the difference to be statistically significant (p = .03, effect size r = .23). Figure 3 illustrates the behavior of trustors across games as within subject session averages. Interestingly, the two distributions do not only differ with respect to their mean but are also differently shaped. The control group has many trustors sending few points to their trustees, whereas a significant share of trustors in the treatment group average larger amounts of points sent. Notably, both groups include individuals that often send all or almost all of their points.

Result 1: *On average, trustors with access to moral communication are more confident to receive a fair amount from trustees and send more points to their trustees.*

4.2 Moral communication increases trustee trustworthiness

Pillutla, Malhotra, and Murnigham (2003) identify two standards of reciprocity from trustees in the trust game: At a minimal level, trustees may feel obligated to return at least as much as trustors sent, so the latter do not end up with less than the 40 point endowment. A higher standard of reciprocity requires trustees to equalize the outcomes of trustor and trustee. In the following analysis, we will refer to the former standard as *minimal reciprocity* and to the latter as

equal outcome.

We hypothesized that moral communication would be associated with more trustworthy behavior on the part of trustees. As measures of how trustworthy and equitable trustees behaved, we calculated two values for each game based on the *amount sent* by trustors and the *amount sent back* by trustees. *Payoff ratio* measures how equal the outcomes of trustor and trustee were as a ratio of their payoffs:

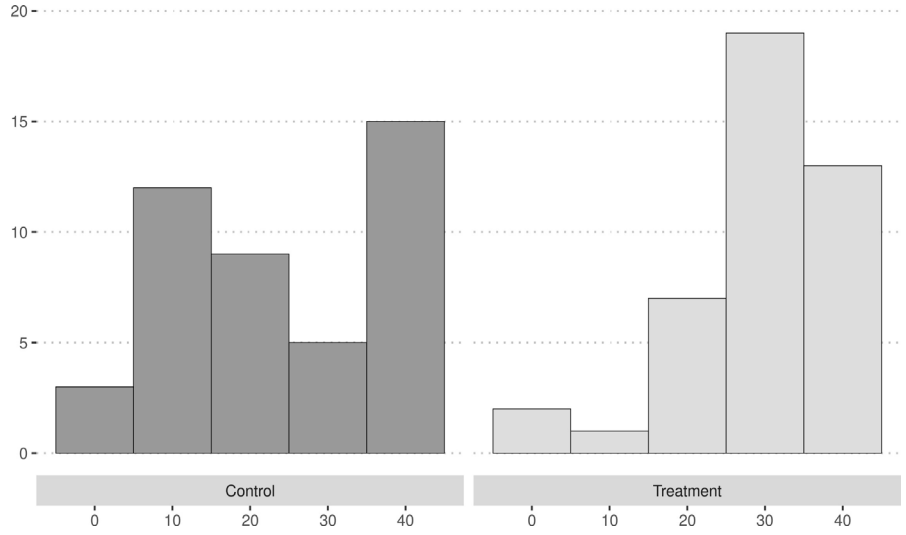

Figure 3: Histogram illustrating the within subject average amounts sent by trustors across all ten games (n = 86). In both groups, a number of trustors send all or close to all their points in all games. In the treatment group, trustors are willing to send a much higher share of their points, whereas trustors in the control group often sent only 10 or fewer points on average.

$$\text{payoff ratio} = \frac{40 \text{ - amount sent} + \text{amount sent back}}{40 + (3 * \text{amount sent}) \text{ - amount sent back}}$$

.

Payoff ratio of 0 indicates that the trustee was sent 40 points and kept all points for themselves, whereas a value of 1 indicates equal payoffs for trustor and trustee. Payoff ratios between 0 and 1 represent a result that favors the trustee. In cases where the payoff ratio > 1, the trustor gained more points than the trustee. Such cases were rare but were observed in a total 12 games. Based on this ratio, we classified trustee decisions as constituting *minimal reciprocity*, *equal outcome*, or neither. Our analysis of trustee behavior focuses on games in which trustors sent at least 10 points. As we noted above, trustors in the control group were more likely to

send no points at all, doing so in 71 games vs. 30 games in the treatment group. This makes the exclusion criterion somewhat biased in favor of excluding games from the control group. We feel that inclusion is justified as our study focuses on trustworthiness which can only be observed when some trust has been extended. In addition, the quality of payoff ratio as an estimator of reciprocity suffers if games in which neither trustor nor trustee sent any points are included with a payoff ratio of 1. This would result in an underestimation of the effect of moral communication on reciprocity.

Across all games and treatments, games in which a trustor sent at least 10 points resulted in a payoff ratio of .72 [.66, .77]. Payoff ratios in the treatment group were $0.13[0.02, 0.25]$ higher than in the control group, indicating that outcomes were more equal. The finding is supported by a highly significant Kolmogorov-Smirnov test for difference in distribution ($D = 0.22$, $p < 0.001$).

To estimate how moral communication affected the likelihood of reciprocal behavior, we estimated two generalized linear mixed models using logit regression on dummies for "equal payoff" and "at least minimal reciprocity". To account for individual tendencies to act according to specific reciprocity standards, both models include random intercepts at the participant level. Results are reported in Table 1. Both types of reciprocity were less likely in later rounds, with odds ratios of $.91[0.84,0.99]$ for equal payoff and $.82[0.73,0.91]$ for minimal reciprocity. For both estimated coefficients, confidence intervals do not include 1. In the treatment condition, trustees were five times more likely to decide for equal payoffs than in the control group (OR = $5.02[1.20,20.91]$), supporting *Hypothesis 2*. Differences in moral self centrality were associated with a high likelihood of deciding for an equal outcome (OR = $10.81[2.64,44.28]$). The visual display of decision consequences had statistically significant impact on the likelihood of equal payoff (OR = $5.27[1.03,27.06]$) but did not affect likelihood of minimal reciprocity. Neither moral communication nor differences in moral self centrality were associated with statistically significant differences in the likelihood of minimal reciprocity.

Confidence intervals for both the treatment effect and moral self centrality are wide. A high intraclass correlation and conditional R^2 indicate that participants stuck to a particular standard of reciprocity. Overall, results indicate that trustees who received moral feedback on their decision were more likely to engage in reciprocity and decide for equal outcomes. The opportunity for moral communication did not make minimal reciprocity more likely. In addition, trustee reciprocity declined over the course of experiment.

Result 2: *When moral communication is available, trustees send higher amounts back to trustors and are more likely to exhibit reciprocity.*

Table 1: Moral communication and reciprocity.

	Consequent			
	(1) Equal payoff		(2) Minimal reciprocity	
Fixed Effects	*OR*	CI_{95}	*OR*	CI_{95}
(Intercept)	0.00**	[0.00, 0.01]	0.38	[0.00, 2977.40]
Round Number	0.91*	[0.84, 0.99]	0.82***	[0.73, 0.91]
Moral Commu.	5.02*	[1.20, 20.91]	3.89	[0.41, 36.81]
Calculator	5.27*	[1.03, 27.06]	0.87	[0.07, 10.62]
Moral Identity	10.81**	[2.64, 44.28]	4.78	[0.63, 63.56]
Random Effects				
σ^2	3.29		3.29	
τ_{00}	8.25		21.79	
ICC	0.71		0.87	
marg. R2 / cond. R2	.19 / .77		.06 / .88	

Note Dependent variables are dummies for (1) equal outcome and (2) trustor ending up with at least their initial endowment. n = 85 with 753 observations. *OR* columns report the odds ratios associated with changes in the independent variable. Numbers in brackets report non-bootstrap confidence intervals of odds ratios at the 95%-level. σ^2 gives the mean random effects variance, τ_{00} gives the between-subject variance, and ICC reports the intraclass correlation. Marginal R^2 estimates the variance explained by fixed effects, conditional R^2 estimates the variance explained by fixed *and* random effects. * $p < 0.05$; *** $p < 0.001$.

4.3 Stabilizing effects of moral communication

We hypothesized that moral communication helps to stabilize cooperative behavior across the rounds of the game. As we report above, trustees in the treatment condition were more likely to decide for equal outcomes but were less likely to do so in later rounds. To test whether moral communication had a stabilizing effect on trustor willingness to extend trust, two linear mixed models were fitted to test for a time trend effect on the amounts sent by trustors, which we expected to drop in the control group (see, e.g., Engle-Warnick and Slonim (2004)) but either increase or remain stable across rounds when moral communication was available. Both models included random intercepts for subjects to control for individual differences in behavior and confidence. Figure 4 shows the concept of our model and Table 2 reports the results of our analysis.

In each round, we asked trustors how confident they were to receive a fair amount back. This confidence can be interpreted as the trust they place in trustees. The first model, testing for effects on confidence, shows that how much a trustor received back in the previous round had a meaningful effect on the amount they were willing to send. Stated confidence to receive a fair amount of trustors who received an "equal payoff" in the

previous rounds was $8.91\%[4.96\%,12.86\%]$ higher than those who received no points in the previous round.[5] A one point difference in the moral identity centrality score was associated with a difference of $10.36\%[2.47\%,18.25\%]$ in confidence.

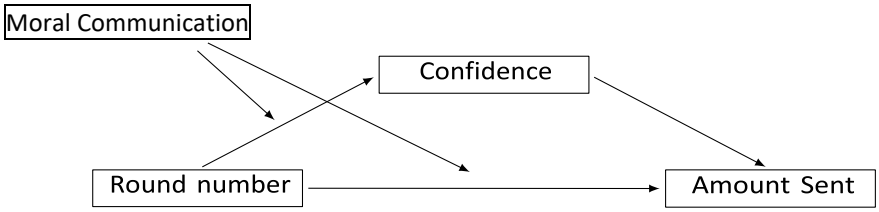

Figure 4: Conceptual diagram of the fitted model. The model tests whether there is a linear time trend in the amount trustors sent and whether that trend was mediated by their confidence. A treatment dummy is included as a moderator of both the direct and the indirect effect to test for the stabilizing effect of moral communication.

The coefficients of round number, moral communication and their interaction indicate that the treatment condition did not have an overall positive effect on trustors confidence but that it negated a negative trend present in confidence over time. While the coefficient of moral communication is not statistically significant, both round number and the interaction of round number and treatment condition had statistically significant and opposing effects. As the coefficients are of similar absolute value, the negative time trend is significantly reduced when moral communication is available. The relatively stable confidence levels in the treatment group and the dropping levels of confidence in the control group are visualized in Figure 5. This interpretation is supported by bootstrapped regression coefficients of the conditional effect: With every additional round, confidence changes by -2.01 $[-2.69,-1.285]$ in the control group but the effect is negated when moral communication is available ($-.09[-.94,0.78]$). The stable level of trustor confidence supports *Hypothesis 3*.

The main variable of interest is not stated trust but trust as revealed in the amounts sent. Trustors who were confident to receive a fair amount back were willing to send more points to the trustee. As discussed above, trustor confidence levels dropped over time in the control group but remained stable in the treatment group, indicating that confidence acted as a conditional mediator of the time trend. The effect of payoff ratio in the previous round was fully mediated by confidence and had no direct effect on amounts sent. Moral identity centrality had a significant effect with a difference of one point in the score being associated with trustors sending $6.78[3.44,-10.12]$ additional points.

[5] Note that rounds in which trustors themselves sent no points in the previous round are included in this analysis.

Table 2: Moral communication as a moderator of trustor behavior.

	Consequent			
	(3) Confidence		(4) Trustor Sent	
Fixed Effects	*B*	*CI*$_{95}$	*B*	*CI*$_{95}$
(Intercept)	1.93	[–32.36, 36.23]	–10.62	[–25.00, 3.77]
Round Number	–2.01***	[–2.69, –1.33]	–0.22	[–0.48, 0.04]
Moral Communicat.	–1.33	[–12.03, 9.36]	–1.33	[–5.71, 3.05]
Round * Moral Com.	1.92*	[0.95, 2.89]	0.79***	[0.42, 1.16]
Moral Identity	10.36*	[2.47, 18.25]	6.78***	[3.44, 10.12]
Payoff ratio t-1	8.91***	[4.96, 12.86]		
Confidence			0.18***	[0.16, 0.21]
Random Effects				
σ^2	313.72		44.57	
τ_{00}	408.79		74.29	
ICC	0.57		0.63	
marg. R^2 / cond. R^2	.12 / .62		.34 / .75	

Note: Dependent variables are dummies for (1) equal outcome and (2) trustor ending up with at least their initial endowment. n = 86 with 773 observations. One observation with undefined payoff ratio was dropped from analysis. *OR* columns report the odds ratios associated with changes in the independent variable. Numbers in brackets report non-bootstrap confidence intervals of odds ratios at the 95%-level. σ^2 gives the mean random effects variance, τ_{00} gives the between-subject variance, and ICC reports the intraclass correlation. Marginal R^2 estimates the variance explained by fixed effects, conditional R^2 estimates the variance explained by fixed *and* random effects. * $p < 0.05$; *** $p < 0.001$.

If moral communication served as a stabilizing factor of trustors' willingness to send as *Hypothesis 3* states, the model would identify a positive coefficient on the interaction between round number and moral communication. Our model shows no significant direct effect of the round or the treatment independently. However, the interaction between the two is highly statistically significant, indicating a positive trend between rounds of the amounts sent by trustors in the treatment group but not of those in the control group. This trend is visualized in Figure 6. Using bootstrapping, we find estimated change between rounds in amounts sent by trustors as mediated by their confidence and moderated by the availability of moral communication to be $-.58[-.92, -.26]$ points per round in the control group and $.56[.23, .87]$ points per round in the treatment group. We can therefore accept *Hypothesis 3* that moral communication helps to stabilize trust and reciprocal behavior. The treatment effect is of meaningful size: In round 10, instead of cooperation unraveling, an average trustor in the treatment group sent 55% more point to their trustee than an average trustor in the control group. As trusting is efficient and moral communication is also associated with higher trustworthiness, average trustor earnings in the treatment group were $106.63[63.9, 148.6]$ points higher.

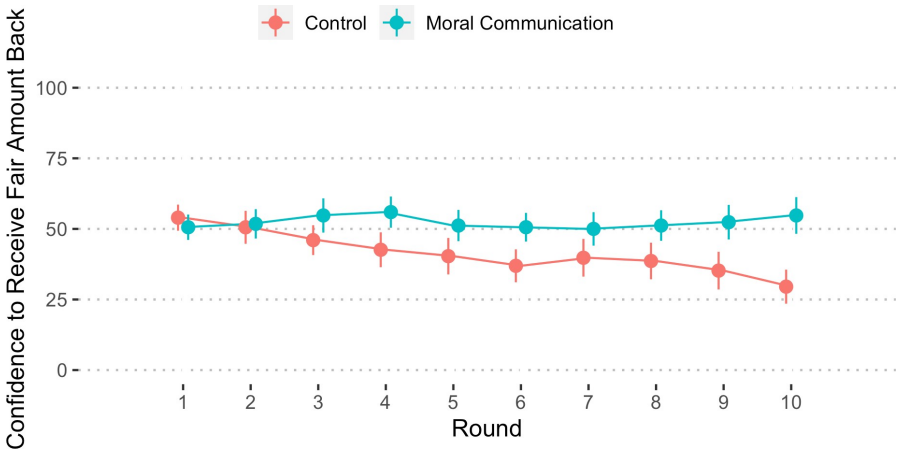

Figure 5: Stated confidence of trustors to receive a fair amount back from the trustee by round, split by treatment. Bars indicate 95% confidence intervals of the means.

Result 3: *The effects of moral communication are established over time. When moral communication is available, trust remains stable and overall efficiency in the trust game increases.*

5 Conclusion and Discussion

In this paper, we investigate how morality and specifically moral communication may contribute trust in impersonal interactions. We do so by incorporating a mechanism for moral feedback in a trust game. Trust and trustworthiness above and beyond levels that maximize individual payoffs are as common as they are irreconcilable with classical game theory based on self-interested profit maximizers that might only be influences by monetary (or, more broadly, material) incentives. Subjects maximizing material payoff would be expected to send no points to other player. As in previous studies, such behavior is exceedingly rare in our results. The typical reaction of behavioral economists over the last 25 years has been to ascribe such behavior to social preferences. However, research has shown that individuals leap at every opportunity not to act in accordance

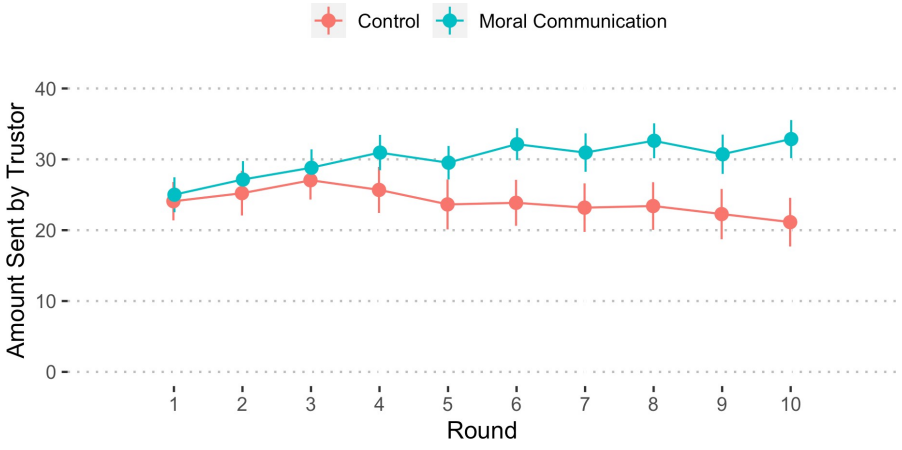

Figure 6: Amounts sent by trustors across rounds, split by treatment. Bars indicate 95% confidence intervals of means.

with these supposed preferences which calls into question whether they are preferences after all.

Another, perhaps more promising approach is an explanation in terms of social norms, in particular in the sense that moral rules function as institutions incorporating specifically moral rewards and punishments. On this account, trust and trustworthiness can be established, if agents understand the underlying morality of their action and can use communication to sanction violation of these norms. In our version of the trust game, both players could provide standardized feedback, saying to what extent the other player was fair or unfair and whether the amount sent or sent back was appropriate (versus too high or too low). As feedback was revealed ex post, it was of no strategic value for influencing the other's choice. It merely allowed an exchange in "moral currency" that was not confounded with the underlying payoff consequences. That is, agents could convey signals of respect (or disrespect) through this channel, which did not affect the other's payoff but would still punish or reward them on the premise that they would like to be respected and to respect others (which is the specific moral content).

Our results show significant effects of moral communication on both willingness to trust and trustworthiness. Availability of moral feedback about decisions resulted in higher levels of points sent to trustees and in rates of points sent back. We also found moral communication to stabilize trust and thereby contribute to overall higher payoffs in the treatment group for both trustees and trustors. In the light of economic institutions that are heavily reliant on trust and the high cost of monitoring and enforcement, our results highlight the importance of morality for individuals, organizations and societies. The ability to extend trust in the context of incomplete contracts is a notable feature of many interactions such as the purchase of goods, hiring of employees and investment of capital. In all of these cases, it is the combination of trust and trustworthiness that leads to increased welfare of the parties involved. Our sanctioning mechanism made the moral judgments of decisions salient by offering a way to communicate them. The positive effects of this kind of feedback on behavior highlight the importance of communication in economic relationships that goes beyond material exchange.

Interpretation of our results should keep in mind the small sample size it is based on. Some models resulted in low percentages of explained variance and significant intercepts warrant an analysis of additional data we gathered, such as the content of moral communication, the influence of moral strategies during the game and reasoning ability. Interestingly, we found no significant effects of age and gender on either trust or trustworthiness. As for age, the lack of effect may be cause by low variance in the student population that our subject pool draws from and reflect similar lack of differences between students, professionals and retired persons observed by Sutter and Kocher (2007, p. 373), who also find no effects of gender.

One should keep in mind that as interaction were anonymous and randomized, there is no systematic difference between earlier rounds and the last round other than past experience. We show that in a traditional trust game, levels of trust deteriorate over time. Whereas cooperation declines in the control group, it remains

stable in the experimental group. Hence, even though moral signals in the treatment group are cheap talk in terms of classical game theory, they seem to be worthwhile in reality, i.e., where moral agents interact.

In terms of morality, we explain this result in terms of the Golden Rule as the guiding moral principle. Agents are thought to frame the game in terms of the Golden Rule and expect both others and themselves to abide by it. Compliance is rewarded by reinforcing moral signals of respect, violations are punished by signaling the other's behavior was disrespectful. And it seems to work.

Conversely, self-signaling in settings where communication with the other is precluded, fails to work. This has been shown by many experiments on so-called "moral hypocrisy" and on dictator games including a "plausible deniability" condition. In these situations, self-interested choices prevail, which supports our general hypothesis that moral rules have to be understood as social norms rather than social preferences.

The time trends we identify have at least two important consequences for future research. First, as the treatment effect was only observable after repeated play, one-shot games may not be suitable to identify the effects of interventions when interactions are likely not one-shot outside of the laboratory. Second, it would be interesting to study whether trust completely deteriorates in the control group and whether it is stable in designs with more rounds when moral communication is available. Previous work on the prisoners' dilemma by Mao, Dworkin, Suri, and Watts (2017) may serve as a guide for such studies. Future work should also vary the communication channels and contents used for moral communication. The design of our experiment forced participants to use a narrow set of messages that may not be realistic for real world interactions. Increasing frequency of interactions with strangers through communication channels that offer little possibility for communicating paraverbal and relational information reduces the possibility of sending morally relevant information. Offering possibilities for interaction through various media channels would help to study the effects of these communication media and extend the external validity of our findings. Future research should also focus on generating results from games other than the trust game which would help to further establish the role of morality in impersonal interaction.

References

Algan, Y., & Cahuc, P. (2013). Trust and growth. *Annu. Rev. Econ.*, *5*(1), 521–549.

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), 1607–1636.

Aquino, K., & Reed, I., Americus. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. doi: 10.1037/0022-3514.83.6.1423

Ariely, D. (2012). The (honest) truth about dishonesty : how we lie to everyone - especially ourselves.

Arrow, K. J. (1972). Gifts and exchanges. *Philosophy & Public Affairs*, 343–362.

Arrow, K. J. (1974). *The limits of organization*. New York: Norton.

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental economics*, *9*(3), 193–208.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: appearing moral to oneself without being so. *Journal of personality and social psychology*, *77*(3), 525.

Baudson, T. G., & Preckel, F. (2016). mini-q: Intelligenzscreening in drei minuten. *Diagnostica*.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122–142.

Bicchieri, C. (2006). *The grammar of society : the nature and dynamics of social norms.* Cambridge: Cambridge University Press.

Binmore, K. (1994). *Game theory and the social contract : 1. playing fair.* Cambridge: The M.I.T. Press.

Binmore, K. (1998). *Game theory and the social contract : 2. just playing.* Cambridge.

Binmore, K. (2005). *Natural justice.* Oxford: Oxford University Press.

Binmore, K. (2010a). Game theory and institutions. *Journal of Comparative Economics*, *38*(3), 245–252.

Binmore, K. (2010b). Social norms or social preferences? *Mind & Society*, *9*(2), 139–157.

Blasi, A. (1983). Moral cognition and moral action: A theoretical perspective. *Developmental Review*, *3*(2), 178–210. doi: 10.1016/0273-2297(83)90029-1

Blasi, A. (1993). The Development of identity: Some implications for moral functioning. In G. G. Noam & T. E. Wren (Eds.), *The moral self* (pp. 99–122). Cambridge, Mass.: MIT Press.

Bottazzi, L., Da Rin, M., & Hellmann, T. (2016). The importance of trust for investment: Evidence from venture capital. *The Review of Financial Studies*, *29*(9), 2283–2318.

Bowles, S., & Polanía-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, *50*(2), 368–425.

Bruni, L., & Sugden, R. (2013). Reclaiming virtue ethics for economics. *Journal of Economic Perspectives*, *27*(4), 141–64.

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Costa-Gomes, M. A., Ju, Y., & Li, J. (2019). Role-reversal consistency: An experimental study of the golden rule. *Economic Inquiry*, *57*(1), 685–704.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and economic behavior*, *46*(2), 260–281.

Curry, O., Whitehouse, H., & Mullins, D. (2019). Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1).

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80.

Danese, G., & Mittone, L. (2018). Trust and trustworthiness in organizations: The role of monitoring and moral suasion. *Managerial and Decision Economics*, *39*(1), 46–55.

Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). Trust at zero acquaintance: More a matter of respect than expectation of reward. *Journal of Personality and Social Psychology*, *107*(1), 122.

Engle-Warnick, J., & Slonim, R. L. (2004). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization*, *55*(4), 553–573.

Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, *96*(5), 1611–1630.

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism– experimental evidence and new theories. In S. Kolm & J. Ythier (Eds.), *Handbook of the economics of giving, altruism and reciprocity* (Vol. 1, pp. 615–691). Elsevier.

Fukuyama, F. (1996). *Trust : the social virtues and the creation of prosperity* (1. paperback ed. ed.). New York, NY: Free Press Paperbacks.

Gauthier, D. P. (1986). *Morals by agreement.* Oxford: Clarendon Press.

Gintis, H. (2014). *The bounds of reason : game theory and the unification of the behavioral sciences* (Rev. ed. ed.). Princeton, N.J.: Princeton University Press.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, *1*(1), 114–125.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *The American economic review*, *86*(3), 653–660.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of economic psychology*, *32*(5), 865–889.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673–676.

Lönnqvist, J.-E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: impression management or self-deception? *Journal of Experimental Social Psychology*, *55*, 53–62.

Mao, A., Dworkin, L., Suri, S., & Watts, D. J. (2017). Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoner's dilemma. *Nature communications*, *8*(1), 1–10.

Marshall, A. (1920). *Principles of economics*. London: Macmillan.

Mazar, N., Amir, O., & Ariely, D. (2008, dec). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633–644. Retrieved from http://journals.sagepub.com/doi/10.1509/jmkr.45.6.633 doi:10.1509/jmkr.45.6.633

Pillutla, M. M., Malhotra, D., & Murnigham, J. K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology*, *39*(5), 448–455. doi: 10.1016/S0022-1031(03)00015-5

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Reed, A., Aquino, K., & Levy, E. (2007). Moral identity and judgments of charitable behaviors. *Journal of Marketing*, *71*(1), 178–193. doi: 10.1509/jmkg.71.1.178

Rustichini, A., & Villeval, M. C. (2014). Moral hypocrisy, power and social preferences. *Journal of Economic Behavior & Organization*, *107*, 10–24.

Sandel, M. J. (2009). *Justice: What's the right thing to do?* Farrar, Straus and Giroux.

Sandel, M. J. (2012). *What money can't buy : the moral limits of markets*. New York: Farrar, Straus and Giroux.

Schelling, T. C. (1960). *The strategy of conflict.* Cambridge, Mass.: Harvard University Press.

Smith, A. (1896). *Lectures on justice, police, revenue and arms: delivered in the university of glasgow, reported by a student in 1763 ; and edited with an introduction and notes by edwin cannan.* Oxford: Claredon Press.

Sugden, R. (1986/2005). *The economics of rights, co-operation and welfare.* London: Palgrave Macmillan UK.

Sugden, R. (2018). *The community of advantage : a behavioural economist's defence of the market* (First edition ed.). Oxford: Oxford University Press.

Sutter, M., & Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic behavior*, *59*(2), 364–382.

Yee, K. K. (2003). Ownership and trade from evolutionary games. *International Review of Law and Economics*, *23*(2), 183–197. doi: 10.1016/S0144-8188(03)00026-7