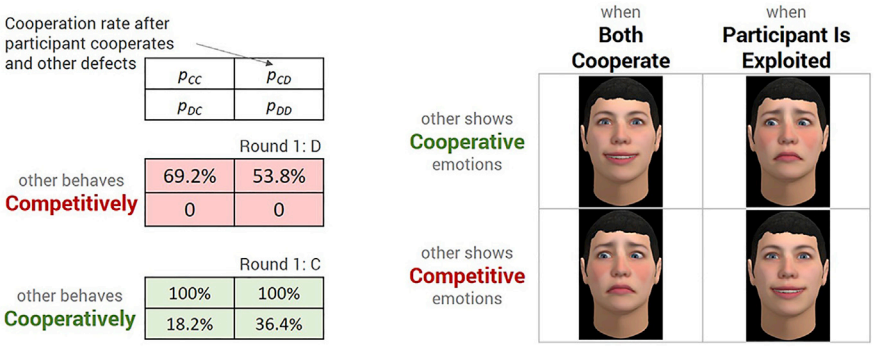
iScience

CellPress
OPEN ACCESS

Article

# Emotion expressions shape human social norms and reputations




| Initial Reputation | Strategy | Emotion | Final Reputation | Reputation Change |
|---|---|---|---|---|
| Negative | Competitive | Competitive | Negative | = |
| | | Neutral | Negative | = |
| | | Cooperative | Negative | = |
| | Cooperative | Competitive | Neutral | ↑ |
| | | Neutral | Neutral | ↑ |
| | | Cooperative | Positive | ↑ |
| Unknown | Competitive | Competitive | Negative | ↓ |
| | | Neutral | Negative | ↓ |
| | | Cooperative | Neutral | ↓ |
| | Cooperative | Competitive | Positive | ↑ |
| | | Neutral | Positive | ↑ |
| | | Cooperative | Positive | ↑ |
| Positive | Competitive | Competitive | Negative | ↓ |
| | | Neutral | Neutral | ↓ |
| | | Cooperative | Positive | ↓ |
| | Cooperative | Competitive | Positive | = |
| | | Neutral | Positive | = |
| | | Cooperative | Positive | = |

Celso M. de Melo, Kazunori Terada, Francisco C. Santos

celso.miguel.de.melo@gmail.com

## HIGHLIGHTS

Emotion expressions' role in shaping reputation in society has been neglected

In our experiment, others had positive/negative reputation, strategy, and emotion

Results showed that emotion shaped how participants assigned reputations to others

This exposes a new class of emotion-based social norms missing in the literature

de Melo et al., iScience *24*, 102141
February 19, 2021
https://doi.org/10.1016/j.isci.2021.102141

iScience

CellPress
OPEN ACCESS

Article

# Emotion expressions shape human social norms and reputations

Celso M. de Melo,[1,4,*] Kazunori Terada,[2] and Francisco C. Santos[3]

## SUMMARY

**The emergence of pro-social behaviors remains a key open challenge across disciplines. In this context, there is growing evidence that expressing emotions may foster human cooperation. However, it remains unclear how emotions shape individual choices and interact with other cooperation mechanisms. Here, we provide a comprehensive experimental analysis of the interplay of emotion expressions with two important mechanisms: direct and indirect reciprocity. We show that cooperation in an iterated prisoner's dilemma emerges from the combination of the opponent's initial reputation, past behaviors, and emotion expressions. Moreover, all factors influenced the social norm adopted when assessing the action of others — i.e., how their counterparts' reputations are updated – thus, reflecting longer-term consequences. We expose a new class of emotion-based social norms, where emotions are used to forgive those that defect but also punish those that cooperate. These findings emphasize the importance of emotion expressions in fostering, directly and indirectly, cooperation in society.**

## INTRODUCTION

The evolution of cooperation among humans often relies on reputations assessing strangers' past behavior (Alexander, 1987; Nowak and Sigmund, 1998, 2005; Ohtsuki and Iwasa, 2004; Milinski, 2016). Prior research identified social norms for updating these reputations, a set of rules defining which actions are perceived as good or bad actions (Bicchieri, 2005; Brandt and Sigmund, 2004, 2005; 2005; Hitoshi et al., 2020; Leimar and Hammerstein, 2001; Ohtsuki and Iwasa, 2004, 2006, 2007; Kandori, 1992; Milinski et al., 2001; Nowak and Sigmund, 1998, 2005; Pacheco et al., 2006; Panchanathan and Boyd, 2003; Santos et al., 2018; Sasaki and Okada, 2017). This research has, however, neglected the effects of nonverbal signals on these social norms. Building on a growing literature showing effects of emotion expressions on decision-making (de Melo and Terada, 2019, 2020; de Melo et al., 2014; Frank, 2004; Keltner and Lerner, 2010; Lerner et al., 2015; Scherer and Moors, 2019; van Kleef et al., 2010), here we show that others' emotion expressions during repeated interaction play a critical role in determining how others' reputations are updated. We present an experiment where participants ($n$ = 711) engaged in the iterated prisoner's dilemma with counterparts that had positive, unknown, or negative reputation, acted cooperatively or competitively (Hilbe et al., 2013; Press and Dyson, 2012; Stewart and Plotkin, 2013), and showed positive, neutral, or negative emotion expressions (de Melo and Terada, 2019; de Melo et al., 2014). The experimental results revealed clear effects of initial reputation, behavior, and emotion expression on cooperation and final reputation. These emotion-based social norms emphasize the insufficiency of others' reputations and actions in explaining how reputation is updated and show that nonverbal communication shapes how reputation is built in society.

There is increasing evidence that emotion expressions can influence human decision-making (de Melo and Terada, 2019, 2020; de Melo et al., 2014; Frank, 2004; Keltner and Lerner, 2010; Lerner et al., 2015; Scherer and Moors, 2019; van Kleef et al., 2010). Recent studies show that emotion displays can enhance or hinder cooperation, according to the contextual meaning of the expressions (de Melo and Terada, 2019, 2020; de Melo et al., 2014; van Kleef et al., 2010). This experimental evidence aligns with general arguments that emotion expressions serve important social functions, including communicating one's mental states to others (Keltner and Haidt, 1999; Keltner and Lerner, 2010; Morris and Keltner, 2000). Moreover, it is, in general, accepted that emotions are elicited from, conscious or nonconscious, appraisal of ongoing events with respect to the individual's goals (Frijda, 1986; Scherer and Moors, 2019). Thus, different emotions can be experienced as a result of different appraisal patterns which, then, result in concomitant

[1]CCDC U.S. Army Research Laboratory, Playa Vista, CA 90094, USA

[2]Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan

[3]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal

[4]Lead Contact

*Correspondence: celso.miguel.de.melo@gmail.com

https://doi.org/10.1016/j.isci.2021.102141

physiological experiences, action tendencies, and expressions. Expressed emotions, therefore, reflect differentiated information about how others are appraising the ongoing interaction with respect to their goals (de Melo et al., 2014; Hareli and Hess, 2010; van Kleef et al., 2010). Accordingly, experimental researchers have shown that people are able to make inferences about others' goals from their emotion expressions (Hareli and Hess, 2010; Parkinson and Simons, 2009), including in decision-making tasks (de Melo and Terada, 2019, 2020; de Melo et al., 2014; van Kleef et al., 2004, 2010). However, whereas this work, including our own, has focused on the consequences of others' emotion expressions on the receiver's immediate responses, the impact of these nonverbal cues on the expresser's reputation – thus reflecting longer-term consequences – has been left unexplored, let alone its potential role in the evolution of human cooperation.

Reciprocity, in several of its flavors, remains as one of the most fundamental cooperation principles discovered to date (Axelrod, 1984; Rand and Nowak, 2013; Sigmund, 2016; Trivers, 1971). In direct reciprocity, cooperation emerges from long-term interactions where players have the chance to return a favor at a later stage or retaliate against wrongdoers (Trivers, 1971). Famous strategies such as Tit-for-Tat, or the recently discovered zero-determinant strategies (Hilbe et al., 2013; Press and Dyson, 2012; Stewart and Plotkin, 2013), provide insightful examples of the advantages and complexities of conditional behaviors based on past interactions. Cooperation may also emerge in the absence of preceding experiences with an opponent. Humans exhibit an astounding ability to judge strangers' behavior towards others (Alexander, 1987; Bicchieri, 2005; Hitoshi et al., 2020; Leimar and Hammerstein, 2001; Ohtsuki and Iwasa, 2004; Kandori, 1992; Milinski, 2016; Nowak and Sigmund, 1998, 2005; Panchanathan and Boyd, 2003), communicate those judgments through reputation mechanisms such as gossip (Dunbar and Dunbar, 1998; Giardini and Conte, 2012; Giardini and Vilone, 2016; Sommerfeld et al., 2007), and behave with strangers based on those reputations (Milinski, 2016; Gross and De Dreu, 2019). This is indirect reciprocity at work. Cooperation with strangers can, thus, emerge in one-shot interactions if information about strangers' past behavior is available and reliable (Hilbe et al., 2018; Santos et al., 2018b; Radzvilavicius et al., 2019; Szabolcs et al., 2016; Uchida, 2010), even though the payoff-maximizing move would be to defect. When combined with particular social norms (Ohtsuki and Iwasa, 2006), defining how individuals should behave and how reputations should be updated, indirect reciprocity can effectively promote cooperation in a community.

Some experimental work indicates that these two forms of reciprocity can interact in important ways, with the effects of direct interaction often overriding the effects of reputation in time (Molleman et al., 2013). This interplay, furthermore, may conflict with humans' cognitive capacity to reason on multiple (and, possibly, conflicting) sources of information through complex heuristics (Feldman, 2000; Melamed et al., 2020; Radzvilavicius et al., 2019; Santos et al., 2018, 2018b). Also, actions and assessment of others' actions are often made in absence of complete information about the decision maker or recipient's reputations and motivation for past actions (Feldman, 2000; Radzvilavicius et al., 2019; Uchida, 2010). Here we resort to a novel experimental setup showing that nonverbal emotion signals may provide an escape hatch to this complexity. Despite the complexity of including an additional layer of information, we show that human cooperation and social norms emerge through simple rules that depend on opponent's emotion profiles, reputations, and past actions. Interestingly, even after several rounds of first-hand interactions, emotions and reputations continue to influence participants' choices. We show how expressions of emotion reveal a nuanced view of the decision maker's mind, and, in some cases, help forgive those that defect or punish those that cooperate.

We present an experiment where participants face an iterated prisoner's dilemma with 20 rounds. In each round, the two players could either cooperate or defect, receiving a return given by the payoff matrix in Figure 1A. The task payoff had financial consequences for participants, as each point would increase their chances of winning a $30 lottery. Participants were given the opportunity to learn about their counterpart's reputation prior to the task, which was either negative, unknown, or positive. This reputation, according to the instructions, had presumably been calculated based on responses to questionnaires administered earlier (see Figure 1B and further details in the Supplemental information). Participants, however, were informed that their own (initial) reputation would not be shared with the counterpart – i.e., from the perspective of the counterpart, the participant's reputation was always unknown. Thus, even though prior work suggests that own reputation can influence behavior and social norms (Nowak and Sigmund, 1998, 2005; Ohtsuki and Iwasa, 2004, 2006), to keep the experimental design simple, here we simply controlled

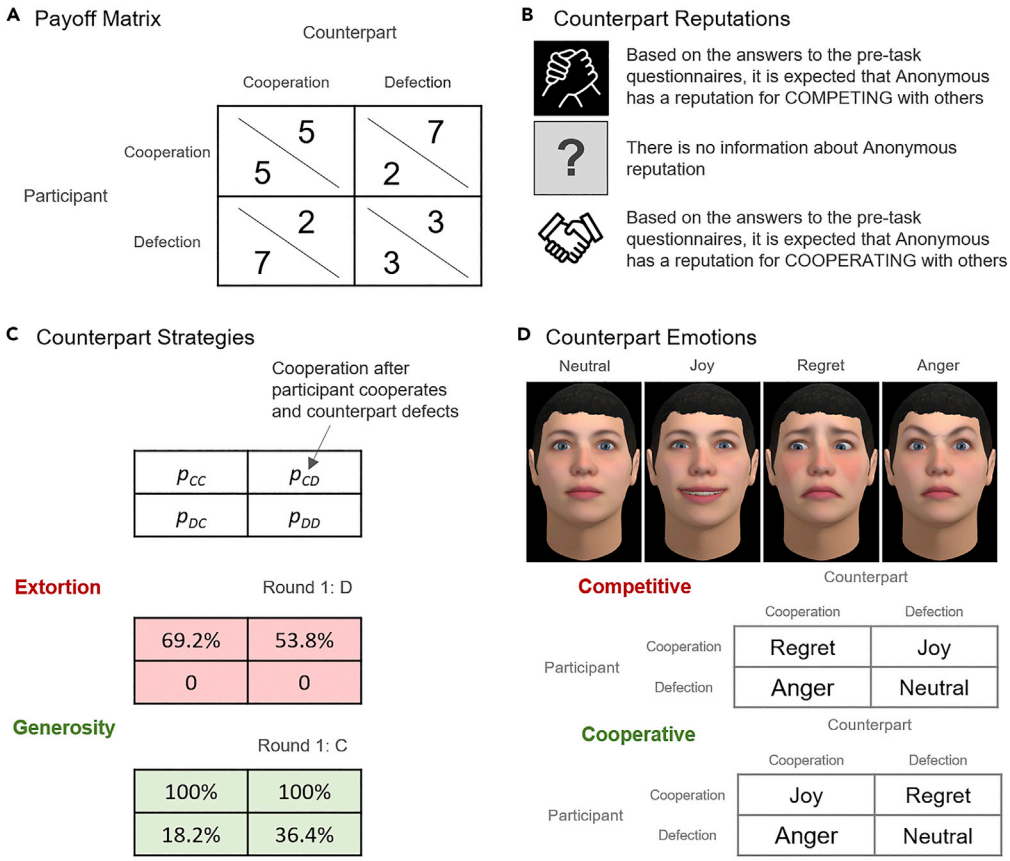



**Figure 1. Experimental design for the iterated prisoner's dilemma task**

(A) The Prisoner's dilemma payoff matrix adopted in the experiments.

(B) The reputation manipulation: participants were instructed that the counterpart's reputation was negative, unknown, or positive based on pre-task questionnaires.

(C) The counterpart strategies: each strategy was defined by the cooperation probability following each possible outcome of the prisoner's dilemma (Hilbe et al., 2013).

(D) The cooperative and competitive emotion expression patterns: the facial expressions were shown after each outcome, according to the corresponding pattern.

for this factor and left the study of higher order social norms (Santos et al., 2018) for future work. Furthermore, to ensure experimental control in the administration of the reputation, strategy, and emotion expression manipulations, participants always engaged with computer scripts (de Melo and Terada, 2019). All experimental procedures were approved by the Gifu University IRB, and participants were fully debriefed at the end.

Participants engaged with counterparts that either acted cooperatively or competitively. To implement this behavior, we looked at recent work on zero-determinant strategies, which include strategies that unilaterally enforce a linear relationship between the players' payoffs (Press and Dyson, 2012). On the competitive end, extortion strategies ensure that the counterpart cannot earn more than the extortionist by exploiting often while cooperating just enough to keep the counterpart cooperating (Figure 1C, top) (Hilbe et al., 2013). On the cooperative end, generous strategies reward cooperation while only punishing defection mildly (Figure 1C, bottom) (Stewart and Plotkin, 2013). See Supplemental information for details and proof that the proposed strategies and payoff matrix meet the requirements for zero-determinant strategies. Prior work (de Melo and Terada, 2020), indicates that emotion expressions moderated the effect of these strategies on cooperation; however, this work did not explore the impact of these factors on social norms. The experimental design presented here, thus, allows for the systematic study of the effects of direct reciprocity – through the counterpart's strategy – crossed with the effects of indirect reciprocity – through the counterpart's reputation.

Participants engaged in the prisoner's dilemma with counterparts that showed cooperative, neutral, or competitive emotion expressions. To accomplish this, players were represented by virtual faces, which is a methodology that has been shown in the past to have high ecological validity while allowing for high experimental control over the emotion expressions (Blascovich et al., 2002; de Melo and Terada, 2019; de Melo et al., 2014). The counterparts' virtual face was young and Caucasian and was kept constant across conditions to control for biases related to physical characteristics of the face (de Melo and Terada, 2019; Todorov et al., 2008). The face showed typical, validated, expressions (de Melo and Terada, 2019; de Melo et al., 2014) for joy, regret, and anger (Figure 1D). After the outcome of the round was revealed, the facial expression was animated in real-time and shown to the participant (see the Video S1 showing the software developed for this experiment). Building on prior work that certain patterns of emotions can promote or hinder cooperation (de Melo and Terada, 2019, 2020; de Melo et al., 2014), we implemented emotion patterns compatible with cooperative, neutral, and competitive intentions (Figure 1D): cooperative – joy following mutual cooperation, regret after exploiting the participant, anger after being exploited, and neutral otherwise; neutral – no emotion was shown; and, competitive – regret following mutual cooperation (given that it missed the opportunity to exploit the participant), joy after exploiting the participant, anger after being exploited and, neutral otherwise.

## RESULTS

The experiment, thus, followed a 3 × 2 × 3 between-participants factorial design: reputation (negative vs. unknown vs. positive) × strategy (extortion vs. generosity) × emotion (competitive vs. neutral vs. cooperative). We first looked at cooperation rate, averaged across the 20 rounds. For an analysis of round effects, please see the Supplemental information. In the Supplemental information we also provide an extended analysis confirming that, in the first round, cooperation rate was only influenced by reputation. We ran an analysis of variance (ANOVA) on cooperation rate (Figure 2A), which showed a main effect of reputation ($F(2, 693) = 5.65$, $p = 0.004$, partial $\eta^2 = 0.016$) in the entire time span of the game. Indeed, post-hoc tests with a Bonferroni correction revealed that participants cooperated less with counterparts with a negative reputation than unknown ($p = 0.020$) or positive ($p = 0.007$) reputations. This result, thus, emphasizes that the effect of indirect reciprocity still existed, despite 20 rounds of direct interaction with the counterpart (Melamed et al., 2020). Moreover, we note that participants appeared to cooperate with those with unknown reputation similarly to those with a positive reputation. There was a main effect of strategy ($F(1, 693) = 155.51$, $p < 0.001$, partial $\eta^2 = 0.183$), with participants cooperating more with generous than extortion strategies. This effect emphasizes the strength of direct reciprocity on overall cooperation (de Melo and Terada, 2020; Molleman et al., 2013). There was also a main effect of emotion ($F(2, 693) = 5.35$, $p = 0.005$, partial $\eta^2 = 0.015$), with participants cooperating more with players showing cooperative emotions than neutral ($p = 0.019$) or competitive ($p = 0.011$) emotions. This effect is in line with earlier research on cooperation and emotion expressions (de Melo and Terada, 2019, 2020; de Melo et al., 2014), despite the existence of direct and third-party information (reputations) available on the counterpart.

To understand the combined effects of actions, current reputation and emotion on the social norm that reckons the next reputation of an individual, we asked participants to rate, on a 100-point Likert scale (−50, *likely to compete*, to 50, *likely to cooperate*), the counterpart's reputation at the start and end of the task (see Supplemental information for details). The analysis focused on perceived reputation at the end, and reputation change (final reputation minus initial reputation) – see Supplemental information for an extended analysis confirming that perceived reputation at the start was only influenced by counterpart's reputation. We first ran an ANOVA on final reputation. This analysis revealed a main effect of (initial) reputation (Figure 2B, $F(2, 693) = 24.69$, $p < 0.001$, partial $\eta^2 = 0.067$). Post-hoc tests revealed that final reputation was higher for those with a positive reputation than unknown reputation ($p = 0.004$) and higher for those with unknown reputation than negative reputation ($p < 0.001$). This result reveals that the current reputation of an individual was taken into consideration when being evaluated by others, suggesting the use of a high-order social norm (Nowak and Sigmund, 1998; Ohtsuki and Iwasa, 2006) where the actions of players do not suffice to assess which actions are deemed good or bad. There was also an effect of the strategy on the final reputation ($F(1, 693) = 157.09$, $p < 0.001$, partial $\eta^2 = 0.185$): Counterparts with generous strategies received a higher final reputation than those with extortion strategies. Emotions also played an important, yet subtler role ($F(2, 693) = 6.52$, $p = 0.002$, partial $\eta^2 = 0.018$): Those showing cooperative emotion received higher final reputation than those showing competitive displays ($p = 0.001$). The analysis also revealed a reputation × strategy interaction ($F(2, 693) = 4.81$, $p = 0.008$, partial $\eta^2 = 0.014$), with unknown reputation being influenced the most by strategy, when compared to negative or positive reputations.


| Reputation | Strategy | Emotion | Final Reputation | Reputation Change |
|---|---|---|---|---|
| B | B | B | B | = |
| | | N | B | = |
| | | G | B | = |
| | G | B | N | ↑ |
| | | N | N | ↑ |
| | | G | G | ↑ |
| U | B | B | B | ↓ |
| | | N | B | ↓ |
| | | G | N | ↓ |
| | G | B | G | ↑ |
| | | N | G | ↑ |
| | | G | G | ↑ |
| G | B | B | B | ↓ |
| | | N | N | ↓ |
| | | G | G | ↓ |
| | G | B | G | = |
| | | N | G | = |
| | | G | G | = |

**Figure 2. Experiment results for the iterated prisoner's dilemma task**

(A) Cooperation rate across the 20 rounds. Error bars correspond to standard errors.

(B) Counterpart reputation perception at the end of the task. Error bars correspond to standard errors.

(C) Change in counterpart reputation perception, calculated as difference between final and initial reputation. Error bars correspond to standard errors.

(D) Social norms based on reputation, strategy, and emotion. Labels for counterpart reputation, strategy, and emotion: negative emotion or reputation (*bad*, B), unknown reputation (U), neutral emotion (N), and positive emotion or reputation (*good*, G). Labels and colors for final reputation perception: negative (B, red), neutral (N, yellow), and positive (G, green). Red and green cells correspond to values that are statistically significantly different than zero (see Table S1 in Supplemental information). Labels and colors for reputation change: downwards (↓, red), neutral (=, yellow), upwards (↑, green). Red and green cells correspond to values that are statistically significantly different than zero (see Table S1 in Supplemental information).

An ANOVA on reputation change (Figure 2C) showed an effect of reputation ($F(2, 693) = 42.43$, $p < 0.001$, partial $\eta^2 = 0.109$), with participants correcting their reputation ratings, when averaging across all conditions, downwards for positive reputations and upwards for negative reputations. This effect emphasizes a synergy between direct experience and initial expectations derived from the counterpart's reputation (Melamed et al., 2020; Molleman et al., 2013). There was an effect of strategy ($F(1, 693) = 148.45$, $p < 0.001$, partial $\eta^2 = 0.176$), with reputations moving downwards for extortionists and upwards for generosity. There was also an effect of emotion ($F(2, 693) = 5.70$, $p = 0.004$, partial $\eta^2 = 0.016$), with reputations being lowered when competitive emotion was shown, and raised when cooperative emotion was expressed. There was a reputation × strategy interaction ($F(2, 693) = 3.73$, $p = 0.025$, partial $\eta^2 = 0.011$), with unknown reputations once again being impacted the most by strategy.

These effects, thus, reveal that initial reputation, strategy, and emotion expressions, all contributed to shape the counterpart's reputation. To gain further insight, we ran one-way *t* tests, for each experimental condition, to understand if final reputation and reputation change were impacted in a meaningful way – i.e., if the value was statistically significantly different than zero (see Table S1 in Supplemental information for

details). Figure 2D summarizes all effects. The first three columns correspond to the different reputations (*G* and *B* indicate good/positive and bad/negative, respectively, and *U* unknown), strategy (*G* and *B* stands for generous and extortion strategies, respectively), and emotion (*G* and *B* stands for cooperative and competitive emotions, respectively). In the "Final Reputation" column, colors encode values that were statistically significant (red for negative, green for positive) and non-significant (yellow for neutral) reputations. Similarly, the "Reputation Change" column, color-codes statistically significant (red for upwards, green for downwards) and non-significant (yellow for neutral) changes in reputation perceptions.

## DISCUSSION

The results summarized in Figure 2D reveal a surprisingly simple social norm combining the effects of reputations, actions and emotions, all providing resourceful information to the evaluator. In particular, emotions offer a confirmation mechanism, both for forgiveness and punishment. First, emotions help to forgive those that defect while having a positive (good) reputation; in other words, defectors (in this case, extortionists) with a good reputation will keep their status as long as they display a cooperative emotion. The key concept "justified defection" in social norms of indirect reciprocity – whereby one should refuse to cooperate with those having a negative reputation (Panchanathan and Boyd, 2003; Ohtsuki and Iwasa 2006; Sigmund 2016; Yamamoto et al., 2020) – can therefore be revisited through the eyes of emotional expressions. Moreover, emotions show the potential to be used as an efficient error-correction mechanism, a feature of central importance in the evolutionary dynamics of cooperation. Second, to recover from a bad reputation, being generous was not sufficient: Good intentions had to be confirmed by cooperative emotions, providing another device for correcting misevaluations, in this case, whenever an individual is rehabilitated from a "bad" to a "good" status.

Emotion theorists have long recognized that the origins of human emotions are inherently social (Frijda, 1986; Haidt, 2003; Scherer and Moors, 2019). Though emotions emerge from appraisal of events with respect to the individual's beliefs and goals, these appraisals often pertain to events and decisions impacting others. The ability to empathize with the fate of others has been argued to be at the origin of human systems of moral judgment and communication (Alexander, 1987; Bicchieri, 2005; Dunbar and Dunbar, 1998; Giardini and Conte, 2012; Giardini and Vilone, 2016; Gross and De Dreu, 2019; Hitoshi et al., 2020; Leimar and Hammerstein, 2001; Ohtsuki and Iwasa, 2004; Kandori, 1992; Milinski, 2016; Nowak and Sigmund, 1998, 2005; Panchanathan and Boyd, 2003; Radzvilavicius et al., 2019; Sommerfeld et al., 2007; Traag et al., 2011). Here we show that emotion expressions are an intrinsic component of these complex systems defining social norms prescribing reputations to others and helping sustain cooperation through direct and indirect reciprocity. This work emphasizes the insufficiency of reputation priors and actions in determining whether an individual is worthy of cooperation, now and in future interactions with other members of the community. Furthermore, research exposes the difficulty in sustaining cooperation in the presence of private, noisy, and incomplete reputation systems (Feldman, 2000). Emotion expressions fill this gap by helping disambiguate social situations and supporting real-time inferences about others' intentions and social norms. Our results clearly show that reputation and behavior cannot code, by themselves, the social norms adopted by participants in the experiment (see Figure 2D), a result of particular relevance for evolutionary biologists addressing the evolution of cooperation through indirect reciprocity, and its interplay with direct reciprocity. Only emotion expression, for instance, explains why an individual that starts with a negative reputation but behaves cooperatively and expresses cooperative emotion can end with a positive reputation. This work, thus, exposes a whole new class of emotion-based social norms that has been missing and potentially complements previous work on social norms of different orders (Alexander, 1987; Bicchieri, 2005; Feldman, 2000; Hitoshi et al., 2020; Leimar and Hammerstein, 2001; Kandori, 1992; Milinski, 2016; Nowak and Sigmund, 1998, 2005; Ohtsuki and Iwasa, 2004; Panchanathan and Boyd, 2003). At the same time, our work also confirms the impressive role of indirect reciprocity and third-party information in human cooperation. Reputations are shown to still have an impact on participants' decisions, even after a significant number of first-hand interactions, and potentially puzzling messages provided by emotions.

Finally, this work has important practical implications, in particular, for the design of autonomous machines – such as robots, autonomous vehicles, and personal assistants (de Melo et al., 2019; Stone and Lavine, 2014). As these machines become increasingly pervasive, their success hinges on humans being willing to cooperate with them (de Melo and Terada, 2019), or on how machines may trigger cooperation among humans (Crandall et al., 2018). Our results suggest that designers should not only consider reputation mechanisms

(e.g., similar to online trading systems) and appropriate behavior (e.g., generous or tit-for-tat strategies), but use nonverbal communication, such as emotion expressions, to build trust and encourage cooperation with humans. Because they can be designed from the ground up, moreover, these machines are in a unique position to shape human behavior and promote cooperation in society.

### Limitations of the study

The study presented here has some limitations that introduce opportunities for future work. The experimental design controlled for the effect of the participants' reputation on their behavior by always assigning them an unknown reputation. However, prior work suggests that own reputation can influence behavior and social norms (Nowak and Sigmund, 1998, 2005; Ohtsuki and Iwasa, 2004, 2006) and, thus, it is worth exploring if this factor interacts in important ways with other's emotion expressions, reputation, and strategy. In this work, we also controlled for possible biases introduced by the counterpart's physical characteristics by keeping the virtual face constant across conditions but, prior research suggests that people can form important judgments from these characteristics (de Melo and Terada, 2019; Todorov et al., 2008) and, so, future work should also account for the effect of this factor on direct and indirect reciprocity. Our sample was collected from a single online pool (Mechanical Turk) and, therefore, we cannot exclude the possibility of a shared sense of social group membership between participants and their counterparts (e.g., participants tended to be as cooperative with counterparts with unknown and positive reputations). However, participants may behave differently and follow different social norms with out-group members. It is, thus, important to study the role of emotion expressions in reputation building with different samples, including involving participants from different social groups. Finally, individual factors can influence people's propensity for cooperation (Balliet et al., 2009; Eckel and Grossman, 1999) and should also be explored, in future work, in conjunction with the factors studied here.

### Resource availability

#### *Lead contact*

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Celso M. de Melo (celso.miguel.de.melo@gmail.com).

#### *Materials availability*

This study did not generate new unique reagents.

#### *Data and code availability*

The published article includes all experimental data collected and analyzed during this study with the supplemental materials. The code supporting the current study has not been deposited in a public repository because it includes proprietary and licensed software but some materials are available from the corresponding author on request.

## ETHICS DECLARATION

All experimental methods were approved by the Medical Review Board of Gifu University Graduate School of Medicine (IRB ID#2018-159). As recommended by the IRB, written informed consent was provided by choosing one of two options in the online form: 1) "I am indicating that I have read the information in the instructions for participating in this research and have had a chance to ask any questions I have about the study. I consent to participate in this research.", or 2) "I do not consent to participate in this research." All participants gave informed consent and, at the end, were debriefed about the experimental procedures. All the experiment protocols involving human-subjects was in accordance to guidelines of the Declaration of Helsinki.

## METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2021.102141.

## ACKNOWLEDGMENTS

This research was supported by JSPS KAKENHI Grant Number JP16KK0004, and the US Army. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred. F.S. acknowledges the support from FCT-Portugal (grants UIDB/50021/2020, PTDC/MAT-APL/6804/2020, and PTDC/CCI-INF/7366/2020).

## AUTHOR CONTRIBUTIONS

C.M., K.T., and F.S. designed the experiment, analyzed the data, and prepared this manuscript. C.M. implemented the experimental software. K.T. ran the experiment and collected the human-subjects data.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 22, 2020
Revised: January 5, 2021
Accepted: January 28, 2021
Published: February 19, 2021

## REFERENCES

Alexander, R. (1987). The Biology of Moral Systems (Transaction Publishers).

Axelrod, R. (1984). The Evolution of Cooperation (Basic Books).

Balliet, D., Parks, C., and Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: a meta-analysis. Group Process. Intergroup Relat. *12*, 533–547.

Bicchieri, C. (2005). The Grammar of Society: The Nature and Dynamics of Social Norms (Cambridge University Press).

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., and Bailenson, J. (2002). Immersive virtual environment technology as a methodological tool for social psychology. Psychol. Inq. *13*, 103–124.

Brandt, H., and Sigmund, K. (2004). The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theor. Biol. *231*, 475–486.

Brandt, H., and Sigmund, K. (2005). Indirect reciprocity, image scoring, and moral hazard. Proc. Nat. Acad. Sci. U.S.A. *102*, 2666–2670.

Crandall, J., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J., and Cebrian, M. (2018). Cooperating with machines. Nat. Comm *9*, 1–12.

de Melo, C., and Terada, K. (2019). Cooperation with autonomous machines through culture and emotion. PLoS One. https://doi.org/10.1371/journal.pone.0224758.

de Melo, C., and Terada, K. (2020). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. Sci. Rep. *10*, 14959. https://www.nature.com/articles/s41598-020-71919-6.

de Melo, C., Marsella, S., and Gratch, J. (2019). Human cooperation when acting through autonomous machines. Proc. Nat. Acad. Sci. U.S.A. *116*, 3482–3487.

de Melo, C., Carnevale, P., Read, S., and Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. J. Pers. Soc. Psychol. *106*, 73–88.

Dunbar, R., and Dunbar, R. (1998). Grooming, Gossip, and the Evolution of Language (Harvard University Press).

Eckel, C., and Grossman, P. (1999). Differences in the economic decisions of men and women: experimental evidence. In Handbook of Experimental Results C, V. Smith Plott, ed. (Elsevier), pp. 509–519. https://www.sciencedirect.com/science/article/pii/S1574072207000571.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. Nature *407*, 630–633.

Frank, R. (2004). Introducing moral emotions into models of rational choice. In Feelings and Emotions, A. Manstead, N. Frijda, and A. Fischer, eds. (Cambridge University Press), pp. 422–440.

Frijda, N. (1986). The Emotions (Cambridge University Press).

Giardini, F., and Conte, R. (2012). Gossip for social control in natural and artificial societies. Simulation *88*, 18–32.

Giardini, F., and Vilone, D. (2016). Evolution of gossip-based indirect reciprocity on a bipartite network. Sci. Rep. *6*, 37931.

Gross, J., and De Dreu, C. (2019). The rise and fall of cooperation through reputation and group polarization. Nat. Comm *10*, 776. https://www.nature.com/articles/s41467-019-08727-8.

Haidt, J. (2003). The moral emotions. In Handbook of Affective Sciences, R. Davidson and K. Scherer, eds. (Oxford University Press), pp. 852–870.

Hareli, S., and Hess, U. (2010). What emotional reactions can tell us about the nature of others: an appraisal perspective on person perception. Cogn. Emot. *24*, 128–140.

Hilbe, C., Nowak, M., and Sigmund, K. (2013). Evolution of extortion in iterated prisoner's dilemma games. Proc. Natl. Acad. Sci. U S A *110*, 6913–6918.

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., and Nowak, M. (2018). Indirect reciprocity with private, noisy, and incomplete information. Proc. Natl. Acad. Sci. U S A *115*, 12241.

Hitoshi, Y., Suzuki, T., and Umetani, R. (2020). Justified defection is neither justified nor unjustified in indirect reciprocity. PLoS One *15*, e0235137.

Kandori, M. (1992). Social norms and community enforcement. Rev. Econ. Stud. *59*, 63–80.

Keltner, D., and Haidt, J. (1999). Social functions of emotions at four levels of analysis. Cogn. Emot. *13*, 505–521.

Keltner, D., and Lerner, J. (2010). Emotion. In The Handbook of Social Psychology, D. Gilbert, S. Fiske, and G. Lindzey, eds. (John Wiley & Sons), pp. 317–352. https://psycnet.apa.org/record/2010-03505-009.

Leimar, O., and Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. Proc. Roy. Soc. Lond. B *268*, 745–753.

Lerner, J., Li, Y., Valdesolo, P., and Kassam, K. (2015). Emotion and decision making. Annu. Rev. Psychol. *66*, 799–823.

Melamed, D., Simpson, B., and Abernathy, J. (2020). The robustness of reciprocity: experimental evidence that each form of reciprocity is robust to the presence of other forms of reciprocity. Sci. Adv. *6*, https://doi.org/10.1126/sciadv.aba0504.

Milinski, M. (2016). Reputation, a universal currency for human social interactions. Philos. Trans. R. Soc. B: Biol. Sci. *371*, 20150100.

Milinski, M., Semmann, D., Bakker, T., and Krambeck, H.-J. (2001). Cooperation through indirect reciprocity: image scoring or standing strategy? Proc. R. Soc. B: Biol. Sci. *268*, 2495–2501.

Molleman, L., van den Broek, E., and Egas, M. (2013). Personal experience and reputation interact in human decisions to help reciprocally. Proc. R. Soc. B Biol. Sci. *280*, 20123044.

Morris, M., and Keltner, D. (2000). How emotions work: an analysis of the social functions of emotional expression in negotiations. Res. Organ. Behav. *22*, 1–50.

Nowak, M., and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. Nature *393*, 573–577.

Nowak, M., and Sigmund, K. (2005). Evolution of indirect reciprocity. Nature *437*, 1291–1298.

Ohtsuki, H., and Iwasa, Y. (2004). How should we define goodness? —reputation dynamics in indirect reciprocity. J. Theor. Biol. *231*, 107–120.

Ohtsuki, H., and Iwasa, Y. (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. J. Theor. Biol. *239*, 435–444.

Ohtsuki, H., and Iwasa, Y. (2007). Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. J. Theor. Biol. *244*, 518–531.

Pacheco, J., Santos, F., and Chalub, F. (2006). Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. PLOS Comput. Biol. *2*, e178.

Press, W., and Dyson, F. (2012). Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. Proc. Natl. Acad. Sci. U.S.A. *109*, 10409–10413.

Panchanathan, K., and Boyd, R. (2003). A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J. Theor. Biol. *224*, 115–126.

Parkinson, B., and Simons, G. (2009). Affecting others: social appraisal and emotion contagion in everyday decision making. Pers. Soc. Psychol. Bull. *35*, 1071–1084.

Radzvilavicius, A., Stewart, A., and Plotkin, J. (2019). Evolution of empathetic moral evaluation. eLife *8*, e44269.

Rand, D., and Nowak, M. (2013). Human cooperation. Trends Cogn. Sci. *17*, 413–425.

Santos, F.P., Santos, F.C., and Pacheco, J. (2018). Social norm complexity and past reputations in the evolution of cooperation. Nature *555*, 242–245.

Santos, F., Pacheco, J., and Santos, F. (2018b). Social norms of cooperation with costly reputation building. In Thirty-Second AAAI Conference on Artificial Intelligence, pp. 4727–4734.

Sasaki, T., and Okada, Y. (2017). The evolution of conditional moral assessment in indirect reciprocity. Sci. Rep. *7*, 41870.

Scherer, K., and Moors, A. (2019). The emotion process: Event appraisal and component differentiation. Annu. Rev. Psychol. *70*, 719–745.

Sigmund, K. (2016). The Calculus of Selfishness (Princeton University Press).

Sommerfeld, R.D., Krambeck, H.-J., Semmann, D., and Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. Proc. Natl. Acad. Sci. USA *104*, 17435–17440.

Stewart, A., and Plotkin, J. (2013). From extortion to generosity, evolution in the Iterated Prisoner's Dilemma. Proc. Natl. Acad. Sci. U.S.A. *110*, 15348–15353.

Stone, R., and Lavine, M. (2014). The social life of robots. Science *346*, 178–179.

Szabolcs, S., Ferenc, S., and István, S. (2016). Deception undermines the stability of cooperation in games of indirect reciprocity. PLOS ONE *11*, e0147623.

Todorov, A., Baron, S., and Oosterhof, N. (2008). Evaluating face trustworthiness: a model based approach. Soc. Cog. Aff. Neuro. *3*, 119–127.

Traag, V., Van Dooren, P., and Nesterov, Y. (2011). Indirect reciprocity through gossiping can lead to cooperative clusters. IEEE, 154–161.

Trivers, R. (1971). The evolution of reciprocal altruism. Q. Rev. Biol. *46*, 35–57.

Uchida, S. (2010). Effect of private information on indirect reciprocity. Phys. Rev. E *82*, 036111.

van Kleef, G., De Dreu, C., and Manstead, A. (2004). The interpersonal effects of emotions in negotiations: a motivated information processing approach. J. Pers. Soc. Psychol. *87*, 510–528.

van Kleef, G., De Dreu, C., and Manstead, A. (2010). An interpersonal approach to emotion in social decision making: the emotions as social information model. Adv. Exp. Soc. Psychol. *42*, 45–96.

Yamamoto, H., Suzuki, T., and Umetani, R. (2020). Justified defection is neither justified nor unjustified in indirect reciprocity. PLoS One. https://doi.org/10.1371/journal.pone.0235137.

**iScience, Volume *24***

## Supplemental Information

# Emotion expressions shape human social norms and reputations

**Celso M. de Melo, Kazunori Terada, and Francisco C. Santos**

# Supporting Information for

## Emotion expressions shape human social norms and reputations

Celso M. de Melo, Kazunori Terada, Francisco C. Santos.

Correspondence to: celso.miguel.de.melo@gmail.com

**This PDF file includes:**



**Other Supplementary Materials for this manuscript include the following:**

## Transparent Methods

**Participant Sample.** Participants for the experiment were recruited from Amazon Mechanical Turk. All participants were from the United States and had an approval rate, based on prior work in this pool, of at least 95%. We excluded participants from prior emotion expression studies[1, 2]. To estimate the sample size for each experiment, we followed the power calculations proposed by Jacob Cohen and implemented in G*Power[3]. We estimated sample size for a 3 × 2 × 3 mixed factorial design: reputation (negative vs. unknown vs. positive) × strategy (extortion vs. generosity) × emotion (competitive vs. neutral vs. cooperative). For a small effect size (Cohen's $f = 0.15$), $\alpha = .05$, and statistical power of 0.90, the recommended total sample size was 690 participants, which rounds up to 702 participants to keep the distribution even across cells. When recruiting from this pool, it is common for some participants to fail to successfully complete the task or otherwise make data entry errors. To account for that, we increased the target sample size experiment to 720 participants. In practice, we ended up with a sample of 711 participants with the following demographics: 62.2% males; age distribution (18 to 21 years, 2.7%; 22 to 34 years, 48.5%; 35 to 44 years, 25.2%; 45 to 54 years, 13.4%; 55 to 64 years, 7.9%; over 64 years, 2.4%).

**Full Anonymity.** All experiments were fully anonymous for participants. To accomplish this, counterparts had anonymous names, and we never collected any information that could identify participants. To ensure understanding, participants were instructed and quizzed on these anonymity conditions prior to starting the task. To preserve anonymity with respect to experimenters, we relied on the anonymity system available through Mechanical Turk. When interacting with participants, researchers are unable to identify the participants, unless we explicitly ask for information that can identify them (e.g., name, email, or photo), which we did not. Note, however, that even though it is not possible to identify the participants, the system supports: (1) rewarding participants, which we only used to pay the lottery winner; and, (2) block participants from participating in (our) future studies, which we never used.

**Financial Incentives.** Participants were paid $2.50 for participating in the experiments, which is a typical amount for this online pool. Moreover, they could earn more money according to their performance in the task. Each point earned in the task was automatically converted to a ticket for a lottery worth $30.00.

**Pre- and post-task questionnaires and additional task measures.** Prior to receiving the instructions for the iterated prisoner's dilemma task, participants were asked to answer demographics questions – gender and age – and the 6-item slider social value orientation (SVO) scale. The SVO scale is used to measure an individual's propensity for cooperation. The administration of this scale, thus, supported the cover story that "based on the answers to pre-task questionnaires" the counterpart's reputation was negative, unknown, or positive. Participants, then, received detailed instructions for the prisoner's dilemma task, including a quiz and tutorial. Prior to starting the task, to measure initial perception of the counterpart's reputation, we asked on a 100-point Likert scale (-50, *likely to compete*, to 50, *likely to*

---

[1] de Melo C., Terada K. 2019 Cooperation with autonomous machines through culture and emotion. *PLOS ONE*, https://doi.org/10.1371/journal.pone.0224758.

[2] de Melo C., Terada K. 2020 The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Sci. Rep.* ***10***.

[3] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html (Last accessed: May-31, 2020)

*cooperate*): "What is *Anonymous* reputation?". After completing the task, to measure final perception of the counterpart's reputation, we asked: "Now that the task is over, what is *Anonymous* reputation?". In addition to collecting the participants' decisions in the prisoner's dilemma for each round, based on prior work[1, 2, 4], we collected three additional measures: (a) participants' decision time; (b) participants' self-reported emotion – from neutral, joy, sadness, anger, and regret – after each round; (c) participants' expectations of cooperation for the next round. However, for the work presented in this paper, these measures were not used.

**Supplementary Text**

Zero-determinant requirements for extortion and generosity strategies

Zero-Determinant (ZD) strategies are memory-one strategies in which the decision for the current round only depends on the outcome of the previous round and they enforce a linear relationship between the players' payoffs in the prisoner's dilemma (Press & Dyson, 2012). ZD strategies are written as a 5-tuple $(p_0, p_R, p_S, p_T, p_P)$, where $p_0$ is the player's probability of cooperation in the first round ($m = 1$), $p_i$ is the probability of cooperation in round $m \geq 2$ given the payoff $i \in \{R, S, T, P\}$ in the previous round. Payoff $R$ and $S$ are given to both players when both player cooperate and defect, respectively. If one player cooperates and the other defects, $T$ is given to the defector and $S$ is given to the cooperator. The relation $T > R > P > S$ is typically assumed to hold. According to Hilbe et al.[5], the probabilities of cooperation are defined as follows:

$$p_R = 1 - \phi(1 - s)(R - l) \quad (1)$$
$$p_S = 1 - \phi[(1 - s)(S - l) + T - S] \quad (2)$$
$$p_T = \phi[(1 - s)(l - T) + T - S] \quad (3)$$
$$p_P = \phi(1 - s)(l - P) \quad (4)$$

, where $l$, s, and $\phi$ are constants.

While ZD strategies are able to enforce a linear relationship between average payoff $\pi$ of the ZD strategist and the expected payoff $\tilde{\pi}$ of the counterpart when the game is repeatedly and infinitely played, Hilbe et al. (*15*) showed that when the game is played $M$ rounds, the relationship between $\pi$ and $\tilde{\pi}$ follows these inequalities:

$$-\frac{p_0}{\phi M} \leq (1 - s)l + s\pi - \tilde{\pi} \leq \frac{1 - p_0}{\phi M} \quad (5)$$

.

We used the payoff values $T = 7$, $R = 5$, $P = 3$, $S = 2$, and a total number of rounds $M = 20$. The following are the values in our experiment for the constants in Equations (1)-(4), and the relation between $\pi$ and $\tilde{\pi}$ predicted by the inequalities in (5):

Extortion

$l = P, s = {}^{1}/_{3}, \phi = {}^{3}/_{13}$

[4] de Melo C., Carnevale P., Read S., Gratch J. 2014 Reading people's minds from emotion expressions in interdependent decision making. *J. Pers. Soc. Psychol.* ***106***, 73-88.

[5] Hilbe C., Nowak M., Sigmund K. 2013 Evolution of extortion in Iterated Prisoner's Dilemma games. *Proc. Natl. Acad. Sci. U.S.A.* ***110***, 6913-6918.

$$p_0 = 0.000, p_R = 0.692, p_S = 0.000, p_T = 0.538, p_P = 0.000$$
$$\frac{1}{3}\cdot\pi + \frac{2}{3}\cdot 3 - \frac{13}{60} \leq \tilde{\pi} \leq \frac{1}{2}\cdot\pi + \frac{2}{3}\cdot 3$$

Generosity

$l = R, s = {}^{1}/_{3}, \phi = {}^{3}/_{11}$

$$p_0 = 1.000, p_R = 1.000, p_S = 0.182, p_T = 1.000, p_P = 0.364$$
$$\frac{1}{3}\cdot\pi + \frac{2}{3}\cdot 5 \leq \tilde{\pi} \leq \frac{1}{2}\cdot\pi + \frac{2}{3}\cdot 5 + \frac{11}{60}$$

We conducted computer simulations to confirm that the strategies used in our experiment met the zero-determinant requirements. Fig. S1 (left panel) shows that average payoffs $\pi$ and $\tilde{\pi}$ are distributed within the range of the linear relationship given by the inequities in (5). Figure S1 (right panel) shows the comparison of experimental results to theoretical predictions, confirming that the relationship between the payoffs of the ZD strategist and the participants fits the linear relationship prediction.

Cooperation in first round

We ran an analysis of variance (ANOVA) on cooperation in the first round. This analysis confirmed an effect of reputation ($F(2, 693) = 6.23$, $p = 0.002$, partial $\eta^2 = 0.018$), with participants cooperating less with counterparts with a negative reputation than unknown ($p = 0.008$) or positive ($p = 0.007$) reputations. The results also suggest that participants appear to cooperate with those with unknown reputation similarly to those with a positive reputation, a comforting and timely message. As expected, however, there was no effect of strategy ($F(1, 693) = 0.235$, $p = 0.628$) and emotion ($F(2, 693) = 2.42$, $p = 0.090$). There were also no statistically significant interactions.

Round effects for cooperation rate

To understand if there were any round effects, we ran a round × reputation × strategy × emotion mixed ANOVA. The results, shown in Fig. S2, confirmed main effects for reputation ($F(2, 693) = 5.65$, $p = 0.004$, partial $\eta^2 = 0.016$), strategy ($F(1, 693) = 155.51$, $p < 0.001$, partial $\eta^2 = 0.183$), and emotion ($F(2, 693) = 5.35$, $p = 0.005$, partial $\eta^2 = 0.015$), as detailed in the main text. They also showed a main effect of round, $F(19, 13167) = 12.52$, $p < 0.001$, partial $\eta^2 = 0.018$, with cooperation tending to decrease as the game progressed – this was particularly evident in the last round, which is in line with prior work suggesting participants defect in the last round since there is no further opportunity for retribution[6]. There were no round × reputation ($F(38, 13167) = 0.69$, $p = 0.925$) and round × emotion ($F(38, 13167) = 1.08$, $p = 0.342$) interactions; however, there was a round × strategy ($F(19, 693) = 6.71$, $p < 0.001$, partial $\eta^2 = 0.010$) interaction, with cooperation tending to decrease in time for extortion but not generosity.

Reputation perceptions before the first round

To understand initial perceptions of reputation, we ran an ANOVA on the pre-task question on counterpart reputation perceptions. This analysis revealed a main effect of reputation ($F(2, 693) = 202.98$, $p < 0.001$, partial $\eta^2 = 0.369$), with negative reputations being rated lower than

[6] Kollock P. 1998 Social dilemmas: The anatomy of cooperation. *Annu. Rev. Sociol.* ***24***, 183-214.

unknown reputations ($p < 0.001$), and unknown reputations being rated lower than positive reputations ($p < 0.001$). As expected, there was no effect of strategy ($F(1, 693) = 0.001$, $p = 0.980$) and emotion ($F(2, 693) = 0.48$, $p = 0.622$). There were also no statistically significant interactions.

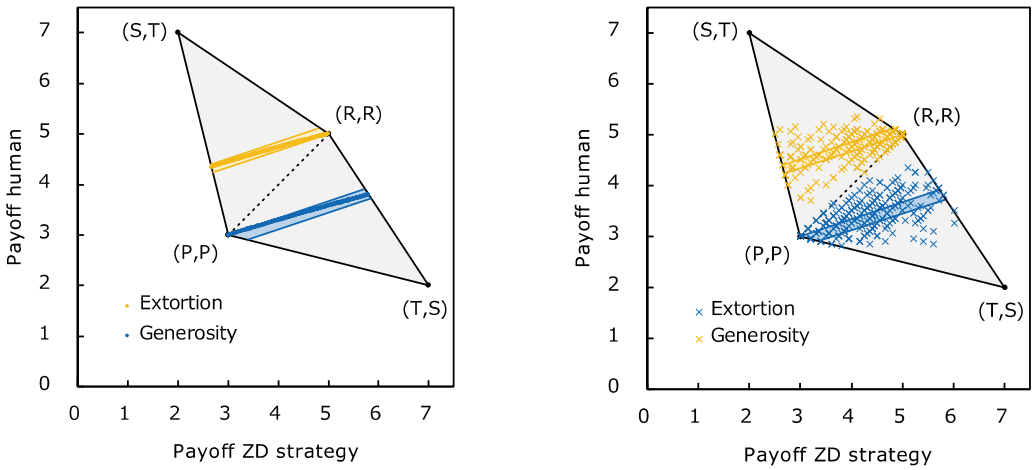



*Figure S1.* Experimental results, theoretical predictions, and simulations for extortion and generosity strategies, Related to Figure 1. A gray-shaded rectangular area surrounded by a solid line indicates the space of possible payoffs for the two players of the prisoner's dilemma. X-axis and y-axis indicate payoffs of ZD strategy and counterpart, respectively. The color-shaded areas between two straight colored solid lines indicate expected payoff ranges according to the inequalities in Eq. (5) — i.e., the theoretical predictions. The left panel shows a comparison of simulated payoffs to the theoretical prediction. Each dot between two color solid lines indicates the average payoff obtained from $10^3$ simulated prisoner's dilemma interactions for a fixed cooperation rate (randomly chosen from 0 to 1). The right panel shows a comparison of experimental results to theoretical predictions.

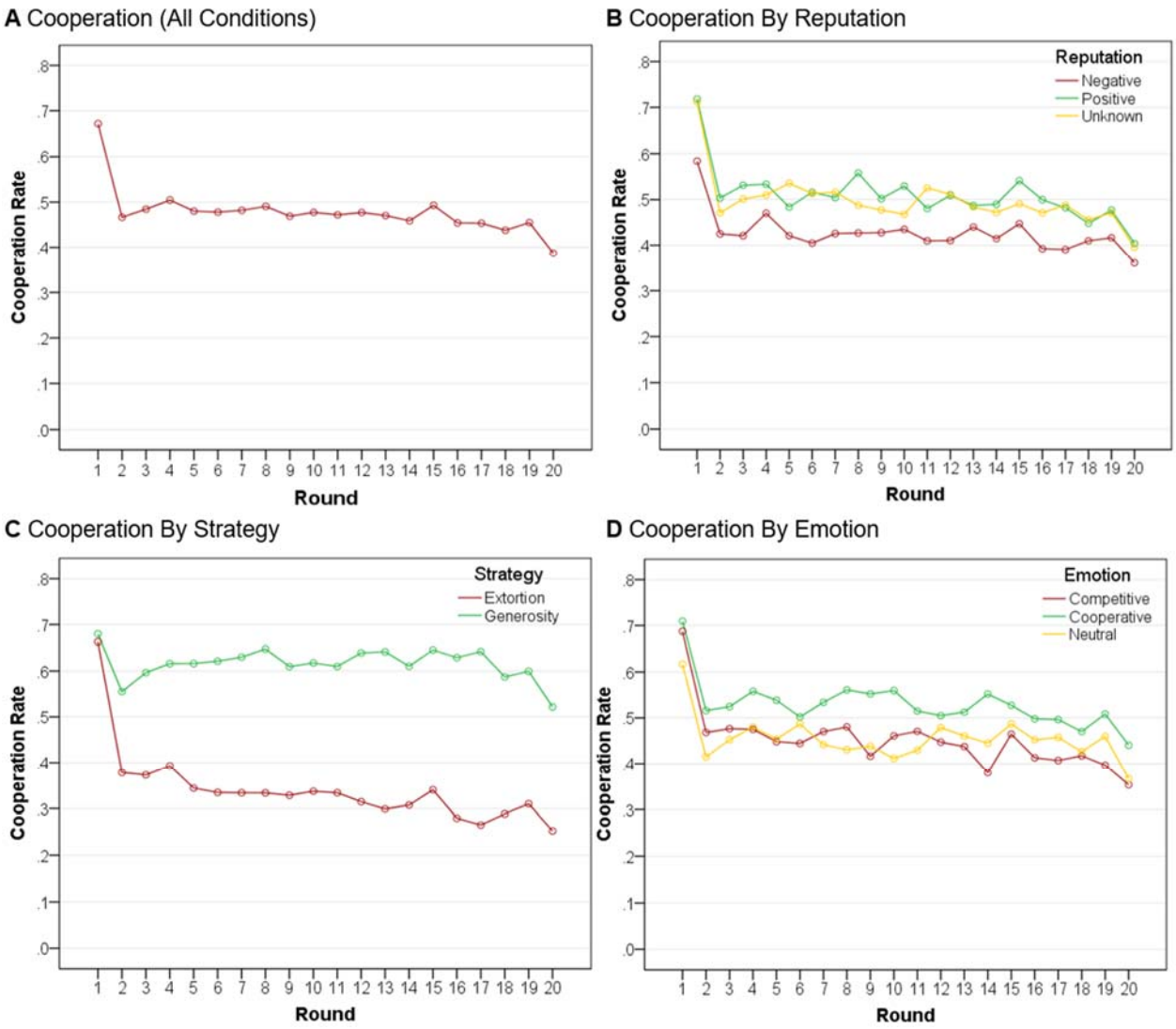



*Figure S2.* Cooperation per round, Related to Figure 2: (**A**) Collapsed across all conditions, (**B**) Collapsed by reputation, (**C**) Collapsed by strategy, (**D**) Collapsed by emotion expression.

Table S1.

Final reputation and reputation changes for each combination of counterpart's reputation, strategy, and emotion, Related to Figure 2. * Final reputation perception is statistically significantly different than zero. † Reputation change is statistically significantly different than zero.

| Reputation | Strategy | Emotion | Final Reputation | | Reputation Change | |
|---|---|---|---|---|---|---|
| | | | *Mean* | *SD* | *Mean* | *SD* |
| Negative | Extortion | Competitive | -17.77 * | 5.58 | -5.33 | 5.75 |
| | | Neutral | -16.10 * | 4.88 | 1.47 | 5.02 |
| | | Cooperative | -13.51 * | 5.31 | -2.49 | 5.47 |
| | Generosity | Competitive | 2.37 | 5.65 | 17.76 † | 5.82 |
| | | Neutral | 4.23 | 5.25 | 30.23 † | 5.41 |
| | | Cooperative | 16.19 * | 5.80 | 26.75 † | 5.98 |
| Unknown | Extortion | Competitive | -17.89 * | 5.14 | -25.35 † | 5.29 |
| | | Neutral | -16.59 * | 5.58 | -27.49 † | 5.75 |
| | | Cooperative | -5.15 | 4.83 | -11.19 † | 4.98 |
| | Generosity | Competitive | 24.85 * | 5.51 | 18.03 † | 5.67 |
| | | Neutral | 29.63 * | 6.16 | 19.56 † | 6.34 |
| | | Cooperative | 34.51 * | 5.44 | 28.73 † | 5.60 |
| Positive | Extortion | Competitive | -12.53 * | 5.08 | -52.98 † | 5.23 |
| | | Neutral | 5.50 | 5.97 | -27.76 † | 6.15 |
| | | Cooperative | 13.81 * | 6.26 | -21.77 † | 6.45 |
| | Generosity | Competitive | 33.68 * | 5.97 | -3.76 | 6.15 |
| | | Neutral | 36.53 * | 6.36 | -5.60 | 6.55 |
| | | Cooperative | 36.29 * | 5.97 | -5.59 | 6.15 |