Tolerant moral judgment drives evolution of collective action

Arunas Radzvilavicius

Charles Perkins Centre, University of Sydney, NSW 2006

Institute of Mathematics, University of Bergen

January 29, 2021

Abstract

In public goods games, the benefit of collective action is shared among all participants, and this creates strong incentives to defect. Theoretical studies and economic experiments predict that without enforcement mechanisms, cooperation in public goods games should collapse. But human societies have repeatedly resolved collective action dilemmas through social norms and institutions. Humans condition their social behavior on the moral reputations of other individuals, and the reputations themselves reflect their past behavior. Here I show how Indirect Reciprocity mechanisms based on group reputations and group-level norms can evolve to promote collective action in public goods games. Individual reputations reflect moral judgments of social behavior within groups, according to the prevailing social norm. Only three norms previously studied as part of Indirect Reciprocity in pairwise games can sustain public goods investments, and their performance depends on how tolerant individuals are to occasional antisocial behavior within groups. When members of the society have predominantly tolerant moral views towards groups, only the norm that abstains from judgment in morally ambiguous interactions (known as “Staying”) can sustain collective action.

1 Introduction

Some of the most consequential interactions in human societies resemble public goods games, where the aggregate benefit of collective action is distributed equally among all participants. If there are no special mechanisms enforcing cooperation in this context, standard economic and evolutionary theories predict that rational players will choose to defect and cooperation will collapse [1]. And yet, empirical observations overwhelmingly show that human societies have repeatedly overcome collective action problems to benefit from individually costly behaviors, including voting, paying taxes, managing natural resources and volunteering [2, 3]. In many cases, communities manage to resolve public goods dilemmas without external top-down enforcement through social norms and self-organized institutions [2].

These observations raise the question of which mechanisms human societies use to overcome the individual temptation to exploit the common-pool resource, and how those mechanisms themselves come to be. While theory predicts that cooperation can be supported with repeated interactions over rigid social networks [4], or when individuals get punished for defection directly [5, 6], these mechanisms are less likely to apply to modern societies where most social interactions are with strangers, or when defection can itself be used as a form of punishment.

In two-person donation games, the evolutionary game-theory framework of indirect reciprocity explains how moral norms and individual reputations together drive the evolution of prosocial behavior [7, 8, 9, 10, 11, 12]. In these models, a donor’s action (to cooperate or not) depends on the recipient’s moral reputation (“good” or “bad”). Individuals’ actions are constantly monitored, and their reputations are updated according to the set of rules known as social norms [13, 9]. A simple social norm called *Stern Judging*, for instance, assigns a good reputation to players who refuse to cooperate with individuals of bad moral standing and to those who selectively cooperate with other good players [14, 7]. Another norm called *Shunning* does not tolerate any interactions with bad individuals, while *Standing* rewards both those who punish defectors and those who cooperate unconditionally.

According to the large theoretical literature on indirect reciprocity in two-player games, high levels of cooperation evolve under social norms like Stern Judging as individuals adopt conditional behavioral strategies and cooperate predominantly with players of good moral standing. Another important condition in these models is that objective reputations must either be provided by a public institutional observer, or individuals themselves have to be empathetic, allowing them to overcome disagreements over relative moral standings of other players [15]. Critically, theory predicts that individuals within societies will often benefit from electing a local institution of moral judgment [16].

Here I propose that a similar mechanism based on moral judgement and reputations can drive the emergence of collective action in public goods games. Indeed, in economic experiments observability alone can increase players’ investment in public goods, and this suggests that players care about how their actions are perceived by other individuals and that they may be adjusting their behavior to improve moral reputations in the eyes of observers [17, 18, 19]. In parallel, empirical observations and experimental studies have shown that humans often behave as conditional cooperators, wherein they cooperate in groups only when they expect others to do the same [20, 21]. It is not entirely clear how humans form their expectations about whether other members of the group will cooperate or not. For example, it has been suggested that in repeated interactions in large groups simple retaliation against the group would fail to sustain cooperation, because it punishes both defectors and cooperators [22, 23].

In the following sections I suggest that humans condition their social behavior based on perceived moral standing of the group as a whole, instead of monitoring individual behaviors. Group-level reputations themselves depend on players’ past behavior within other groups, given the reputations of those groups and the prevailing social norm. Both individual, and group-level reputations are provided by an institutional observer, freeing individuals from cognitively costly task of tracking reputations and behaviors privately. First, I introduce the group-level social norms and investigate their effects on the evolutionary dynamics of players’ behavioral strategies. Then, I systematically study cooperation rates in societies following each of the five social norms as strategies are allowed to evolve in a stochastic model, and

Group Reputation	G	G	B	B
Individual Action	C	D	C	D
Stern Judging	G	B	B	G
Simple Standing	G	B	G	G
Staying	G	B	-	-
Shunning	G	B	B	B
Scoring	G	B	G	B

Table 1: Moral reputations (good, “G” or bad, “B”) assigned to individuals based on their actions (cooperate, “C” or defect “D”) under five social norms. Under Staying, “-” indicates that the given observation is not taken into account in updating individual reputations. The reputation of a group is good only if the fraction of individually good players within the groups exceeds a threshold h/S .

determine which social norms can sustain collective action.

2 Results

Consider a society where individuals participate in public goods games within groups of S randomly chosen players. Players adopt one of three behavioral strategies: always cooperate (ALLC, x), always defect (ALLD, y), or discriminate based on group-level reputation (DISC, z). Cooperators always contribute to the public good, defectors never contribute, and discriminators only cooperate if the society considers their group morally good. The benefit of collective action – the sum of all S individual contributions multiplied by R – is then distributed equally among all S participants. Players occasionally copy the behavioral strategy of their peers, based on its mean payoff over multiple games.

Following previous work on indirect reciprocity in two-player games [24, 15, 16], here I consider five second-order social norms that have been shown to sustain high levels of cooperation in donation games [9, 13, 25]. These moral norms determine how the reputation of an individual should be updated based on her action (cooperate or defect) and the reputation of the group (Table 1): *Stern Judging* (SJ), *Simple Standing* (SS), *Staying* (ST), *Shunning* (SH) and *Scoring* (SC). I further consider a simple threshold rule of determining group-level reputations: the society deems a group of size S good if at least h players within the group have good individual reputation. The threshold h reflects the level of moral tolerance towards groups: when $h = 0$ the society is tolerant and the DISC players contribute to the public good unconditionally, while with $h = S$ individuals follow strict moral rules and DISC strategists cooperate only in groups that are free of bad players. All reputations within the society are provided by an objective institution, and so there are no disagreements over individual or group-level moral standings.

Replicator dynamics of strategy evolution

To analyze how social norms and group-level moral tolerance h affect cooperation in public goods games, I used the equilibrium reputation frequencies and standard replicator dynamics

Figure 1: Replicator dynamics of behavioral strategy evolution in societies following Stern Judging, Simple Standing, Staying, Shunning and Scoring norms, and with different levels of moral group-level tolerance h . Triangles describe the frequencies of three behavioral strategies: unconditional cooperators (ALLC), unconditional defectors (ALLD), and discriminators (DISC). DISC strategists contribute to the public good in groups of good moral standing, given their tolerance threshold h . Circles indicate strategic equilibria and the shaded regions represent basins of attraction towards a cooperative equilibrium in the strategy space. Dashed black lines represent neutral strategic evolution where all players have bad reputations. The group size is set to $S = 5$ and public goods multiplication factor is $R = 3$.

equations [26, 27] as detailed in the Methods section. Together, these equations describe how strategy frequencies (ALLC, ALLD, DISC) change over time in infinite populations as individual strategies reproduce based on their expected payoffs over multiple games. I follow previous work [24, 15] assuming that reputation dynamics in the society reach an equilibrium before strategy updating takes place, and the dynamics of moral evaluation can be separated from the evolutionary dynamics of behavioral strategies.

How do social norms affect strategy evolution? The numerical analysis of strategy replicator dynamics revealed that in societies adhering to the norms that punish defection against bad groups – Shunning and Scoring – strategic evolution always leads to a mix of defectors and discriminators, but this dimorphic population state also causes all players to acquire bad moral standing. As a result, there are no payoff differences between DISC and ALLD strategists, and cooperation collapses (Figure 1).

Only three second order social norms have strategic equilibria that supports non-zero investments in public good: Stern Judging, Simple Standing and Staying (Figure 1). Under Stern Judging, the population evolves either to a non-cooperative ALLD equilibrium, or a cooperative mix of ALLD and DISC strategies, depending on the initial condition, with similar dynamics observed under Simple Standing. Under both norms the basin of attraction towards the cooperative equilibrium (shaded regions in Figure 1) is always larger when players have strict tolerance thresholds h . The cooperative stable state disappears completely as players become increasingly tolerant (i.e., with lower thresholds h).

Staying norm has been previously proposed as a way to avoid making controversial judgments in ambiguous social interactions with players of bad moral character [24]. In two-player donation games, Staying has been shown to support surprisingly high levels of cooperation, often performing better than the other four leading norms [24]. In the context of public goods games, I find that under the Staying norm of moral judgment, defection (ALLD) will be the only possible equilibrium if players have strict moral thresholds h . Strikingly, with more tolerant views towards groups, a mix of discriminators and ALLD strategists becomes stable, and strategic evolution almost always leads to this cooperative state (Figure 1).

Social norms that promote collective cooperative action

Next, I asked how the prevailing social norm and its moral tolerance thresholds affect cooperation rates in finite populations. In contrast to the replicator dynamics in infinite populations, in societies of finite size strategy frequencies do not evolve towards a fixed point but rather fluctuate due to demographic stochasticity and random strategy exploration. In this context, I studied the long-term evolution of behavioral strategies in Monte Carlo simulations, where strategies spread through social contagion (see Methods). After many rounds of the public goods game, a randomly selected player copies the strategy of another individual with the probability $1/(1 + \exp(-w [P_1 - P_2]))$, where w is selection strength and P_1 and P_2 are the payoffs accumulated by the two players [28]. In addition, there is a small chance u (exploration rate) that a player will choose a behavioral strategy at random [11].

When random strategy exploration rate u is low, cooperation rates stay close to zero under both Scoring and Shunning, confirming the insights from the replicator dynamics analysis (Figure 2). Only when players are allowed to explore the strategy space more frequently, the cooperation rates reach higher, albeit still very low, values (Figure 2b). As predicted

by the replicator dynamics analysis, Stern Judging and Simple Standing both sustain high cooperation rates when players follow strict group-level moral rules, and cooperation collapses as these views become more tolerant (lower h). Frequent random exploration of the strategy space also leads to less cooperation, as players are more likely to escape the cooperative basin of attraction in the strategy space.

Figure 2: Mean cooperation rates in stochastic dynamics of strategy evolution in public goods games under five social norms. Stern Judging (SJ) and Simple Standing (SS) norms promote collective action only with strict group-level moral thresholds h , while Shunning (SH) and Scoring (SC) produce very low levels of cooperation. Staying (ST) norm supports the highest rate of cooperation in societies with tolerant moral views towards groups. Here the population size is $N = 400$, $R = 3$, and u is random strategy exploration rate.

We have seen that Stern Judging and Standing norms require strict group-level moral thresholds h . In contrast, the Staying norm of moral assessment supports high levels of public good investments when societies have tolerant views towards groups, with $h < S$ (Figure 2). This result is also less sensitive to increasing strategy exploration rates, and increasing group size S , once again confirming the predictions of strategy dynamics in societies of infinite size.

Why do some social norms fail to sustain cooperative behavior, others require strict group-

level moral judgments, while under Staying stable collective action evolves even when societies have tolerant moral views towards groups? Shunning and Scoring norms inevitably lead to equilibria in which morally bad individuals dominate the society: Scoring rewards cooperation with bad groups, which is prone to exploitation, and both norms consider defection against bad groups morally unacceptable, and this puts DISC strategists at disadvantage. Stern Judging and Simple Standing reward players who refuse to cooperate in groups of bad moral standing. This means that unconditional defectors can earn good reputation if they end up playing within groups composed mostly of defectors, and this can benefit them in subsequent games. Only strict morality thresholds h can prevent ALLD strategists from exploiting the public good in this setting. Staying eliminates the shortcomings of these four norms: it does not reward nor punish defection against bad groups, selecting against ALLD and facilitating the evolution of DISC with relaxed moral thresholds h .

3 Discussion

Human cooperation is unique among animals in that we often cooperate with complete strangers, even when there are strong incentives to behave selfishly. One explanation provided by evolutionary theory and sociology is that human societies follow complex culturally transmitted rules of what types of behavior are morally acceptable. Societies constantly perform moral judgments of observed actions and individuals then condition their behavior on perceived moral reputations of other people, often choosing to cooperate only with peers of good moral standing [29].

While the framework of indirect reciprocity has been developed for two player donation games, observations suggest that similar considerations should apply to public goods games: the level of prosocial behavior depends on whether or not interactions in the public goods game can be observed by others, and in economic experiments players within groups often behave as conditional cooperators, based on their expectations of how other players will behave [20]. It is possible that players form these expectations by considering coarse-grained representations of group moral reputations, either inferred or provided by a formal public institution reinforcing conditional cooperation [2], and there is some empirical support that such group-level reputations may be effective in promoting prosocial behavior [30].

The rules defining which types of behavior are considered morally acceptable are specified by social norms, and the question of which norms are the most efficient at promoting prosocial behavior is the subject of ongoing research [13, 9]. When moral reputations are provided by an institutional observer trusted by all members of the society and there are no disagreements over the reputations of others, Stern Judging norm is believed to sustain the most cooperation, closely followed by Simple Standing [14, 11]. But when members of the society make their own individualistic moral judgements, and reputations depend on the observer, cooperation tends to collapse and strict norms like Stern Judging become ineffective [15, 8]. Societies can eliminate these disagreements and restore high levels of prosociality by either delegating moral judgments to an institutional observer [16, 31], or by adopting empathetic strategies that perform moral evaluations from perspectives of other players [15].

Here I showed how indirect reciprocity and adherence to social norms can drive collective action in public goods games. While there are other mechanisms promoting cooperation in

this setting, studying the impact of reputations and social norms is critical to understanding cooperation in modern highly connected societies where dilemmas of common-pool resource sharing often have to be resolved within groups of strangers.

In promoting public-goods cooperation, social norms prescribing individual reputations interact with group-level moral tolerance that determines the moral standing of the group as a whole, and, ultimately, leads to conditional cooperation of DISC strategists. Two generous norms that discourage retaliation against groups and individuals of bad moral standing – Image Scoring and Shunning – fail to sustain collective action and lead to the collapse of cooperation in public good games, even though they have been shown to promote moderate rates of cooperative behavior in pairwise games [11]. Stern Judging and Simple Standing (both encouraging defection against “bad”) sustain high levels of cooperative behavior in public goods games, but only when individuals have strict moral views towards groups, requiring all or nearly all players within the group to carry positive reputations.

These four social norms differ in how they evaluate cooperation and defection against groups of bad moral standing, which itself is a subject of controversy [24]. Should societies treat cooperation with bad individuals as a moral good, or should they encourage their punishment? My key finding in this work is that the social norm that abstains from moral judgement in these ambiguous interactions – the so-called Staying norm – is superior in driving collective action, and it sustains high levels of cooperation in societies that have tolerant views towards occasional antisocial behavior within groups. Staying has been introduced by Santos et al. [24] in two-person giving games as a form of “selective inattention” which allows observers to abstain from making firm judgments in morally questionable social interactions, where it also outperformed Stern Judging and other simple norms.

While demonstrating that tolerant moral judgment can drive evolution of collective action in public goods games, this study raises a number of questions. Whereas I studied societies that universally trust institutional moral judgments or where reputations are homogeneous due to gossip, in more realistic cases opinions on individual and group-level moral standings may differ among observers. This can happen, for instance, because individuals use different social norms, make errors, or simply observe different interactions. Another question is how norms themselves can evolve when individuals have different beliefs on what constitutes a morally good behavior, and whether these moral beliefs can also spread through social contagion. And finally, while experiments suggest that humans adopt conditional behavioral strategies in public goods games, it is not entirely clear what is the underlying mechanism of their discriminatory behavior, and whether it is based on moral reputations of individuals, groups, or some other variables. These outstanding questions about the origin, spread and following of moral codes provide a roadmap for future research.

4 Methods and Materials

Payoffs of individual strategies

Consider a society where individuals participate in public goods games within groups of S randomly chosen players. Players adopt one of three behavioral strategies: always cooperate (ALLC, x), always defect (ALLD, y) and discriminate based on reputation (DISC, z). Cooperators always contribute to the public good, defectors never contribute, and discriminators

only cooperate if they perceive the group as morally good. Using standard replicator dynamics, the frequencies of the three strategies $s \in \{x, y, z\}$ change according to their payoffs relative to the mean population fitness:

$$\dot{s}_i = s_i(P_i - P) \quad (1)$$

$$P = \sum_{i \in \{x, y, z\}} s_i P_i. \quad (2)$$

The benefit of collective action (sum of all contributions multiplied by R) is distributed equally among all S participants. The average payoffs of the three strategies are then:

$$P_x = \sum_{n_x, n_z=0}^{S-1} p(n_x, n_z) \left[\frac{R}{S} (n_x + 1 + n_z p_G(n_x + 1, n_z)) - 1 \right]; \quad (3)$$

$$P_y = \sum_{n_x, n_z=0}^{S-1} p(n_x, n_z) \frac{R}{S} (n_x + n_z p_G(n_x, n_z)); \quad (4)$$

$$P_z = \sum_{n_x, n_z=0}^{S-1} p(n_x, n_z) \left[\frac{R}{S} (n_x + (n_z + 1) p_G(n_x, n_z + 1)) - p_G(n_x, n_z + 1) \right]. \quad (5)$$

Here $p(n_x, n_z)$ is the probability that a focal individual will find herself in a group with n_x cooperators and n_z discriminators, which depends on the frequencies of the three strategies within the general population x , y , and z :

$$p(n_x, n_z) = \frac{(S-1)!}{n_x! n_z! (S-1-n_x-n_z)!} x^{n_x} z^{n_z} (1-x-z)^{S-1-n_x-n_z}. \quad (6)$$

Here $p_G(n_x, n_z)$ is the probability that a group with n_x cooperators and n_z discriminators is seen as morally good. I assume that a group of size S is seen as good if at least h individuals within the group have good reputation. $p_G(n_x, n_z)$ depends on the frequencies of good players among the three strategy types in the general population f_x , f_y and f_z :

$$p_G(n_x, n_z) = \sum_{k=h}^S \sum_{i,j}^{n_x, n_z} \binom{n_x}{i} f_x^i (1-f_x)^{n_x-i} \binom{n_z}{j} f_z^j (1-f_z)^{n_z-j} \\ \times \binom{S-n_x-n_z}{k-i-j} f_y^{k-i-j} (1-f_y)^{S-n_x-n_z-k+i+j}. \quad (7)$$

In the special case of strict moral views where the group is good only if all of its members are morally good ($h = S$), $p_G(n_x, n_z) = f_x^{n_x} f_y^{S-n_x-n_z} f_z^{n_z}$.

Reputation dynamics

In societies following the Stern Judging norm, a good reputation is assigned to an individual only if she cooperates within the group that is itself seen as morally good, or defects in a

group seen as bad. At equilibrium, the frequencies of good individuals among the ALLC (f_x), ALLD (f_y) and DISC (f_z) strategists in the general population are then

$$f_x = \sum_{n_x, n_z=0}^S p(n_x, n_z) p_G(n_x, n_z); \quad (8)$$

$$f_y = \sum_{n_x, n_z=0}^S p(n_x, n_z) (1 - p_G(n_x, n_z)); \quad (9)$$

$$f_z = 1. \quad (10)$$

In societies following the Simple Standing norm, only defection against good groups will result in bad individual reputation, with all other interactions considered morally good:

$$f_x = 1; \quad (11)$$

$$f_y = \sum_{n_x, n_z=0}^S p(n_x, n_z) (1 - p_G(n_x, n_z)); \quad (12)$$

$$f_z = 1. \quad (13)$$

Because moral judgement of interactions within bad groups is a controversial issue (both cooperation and defection against bad players can be considered a moral good), in the context of two-player games Sasaki et al. [24] have recently proposed a new norm – Staying – that does not update individual reputations following their interactions with morally bad players (or, in our case, bad groups). Equilibrium frequencies of good individuals among the three types of strategies can then be expressed as

$$f_x = \sum_{n_x, n_z=0}^S p(n_x, n_z) [p_G(n_x, n_z) + f_x (1 - p_G(n_x, n_z))]; \quad (14)$$

$$f_y = \sum_{n_x, n_z=0}^S p(n_x, n_z) [f_y (1 - p_G(n_x, n_z))]; \quad (15)$$

$$f_z = \sum_{n_x, n_z=0}^S p(n_x, n_z) [p_G(n_x, n_z) + f_z (1 - p_G(n_x, n_z))]. \quad (16)$$

Under the Shunning norm, good individual reputation can only be assigned as a result of cooperation within morally good groups:

$$f_x = \sum_{n_x, n_z=0}^S p(n_x, n_z) p_G(n_x, n_z); \quad (17)$$

$$f_y = 0; \quad (18)$$

$$f_z = \sum_{n_x, n_z=0}^S p(n_x, n_z) p_G(n_x, n_z). \quad (19)$$

And finally, under Scoring, all cooperative acts result in good donor reputation, regardless of the recipient group’s moral standing:

$$f_x = 1; \tag{20}$$

$$f_y = 0; \tag{21}$$

$$f_z = \sum_{n_x, n_z=0}^S p(n_x, n_z) p_G(n_x, n_z). \tag{22}$$

To determine the expected change in strategy frequencies x , y and z , I first solved for f_x , f_y and f_z by combining (7) with the reputation dynamics equations for a given norm. I used Matlab’s Symbolic Math Toolbox, for each norm manually verifying that the numerical solution lied in the range of $[0, 1]$ and was unique. Then, I used these equilibrium frequencies of good players within the three strategic subpopulations to calculate the expected payoffs (3-5), and the change in strategy frequencies (1).

Stochastic simulations

To measure mean cooperation rates in public goods games under five social norms, I performed a series of Monte Carlo simulations in populations of $N = 400$ individuals. Each individual is characterized by its own strategy (ALLC, ALLD or DISC) and reputation in the eyes of the public institution. Players then participate in a series of L public goods games, according to the rules described above, each time randomly forming a group of S individuals. Reputations are updated after each round according to the prevailing social norm, taking into account the group-level reputation with the tolerance threshold h .

After L rounds, two players are chosen for strategy updating. A player copies the strategy of another individual with the probability $1/(1 + \exp(-w [P_1 - P_2]))$, where w is selection strength and P_1 and P_2 are the payoffs accumulated by the two players [28]. In addition, after the strategy updating step, each individual may choose to adopt a new strategy chosen at random with a probability u (exploration rate). The simulation is initialized with a random mix of strategies and randomly assigned reputations. Mean cooperation rates are calculated after 100,000 strategy updating steps over the last 10,000 rounds.

Code availability

Matlab scripts used to solve the replicator dynamics equations and to perform Monte Carlo simulations are publicly available at <https://osf.io/m9t5p/files/>.

References

- [1] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.
- [2] Elinor Ostrom. Collective action and the evolution of social norms. *Journal of economic perspectives*, 14(3):137–158, 2000.
- [3] Thomas Dietz, Elinor Ostrom, and Paul C Stern. The struggle to govern the commons. *Science*, 302(5652):1907–1912, 2003.

- [4] Francisco C Santos, Marta D Santos, and Jorge M Pacheco. Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201):213–216, 2008.
- [5] Robert Boyd and Peter J Richerson. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, 13(3):171–195, 1992.
- [6] Herbert Gintis. Strong reciprocity and human sociality. *Journal of theoretical biology*, 206(2):169–179, 2000.
- [7] Michihiro Kandori. Social norms and community enforcement. *The Review of Economic Studies*, 59(1):63–80, 1992.
- [8] Christian Hilbe, Laura Schmid, Josef Tkadlec, Krishnendu Chatterjee, and Martin A Nowak. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the national academy of sciences*, 115(48):12241–12246, 2018.
- [9] Hisashi Ohtsuki and Yoh Iwasa. How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of theoretical biology*, 231(1):107–120, 2004.
- [10] Isamu Okada, Tatsuya Sasaki, and Yutaka Nakai. Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. *Scientific reports*, 7(1):1–11, 2017.
- [11] Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Scientific reports*, 6(1):1–9, 2016.
- [12] Martin A Nowak and Karl Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577, 1998.
- [13] Hisashi Ohtsuki and Yoh Iwasa. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of theoretical biology*, 239(4):435–444, 2006.
- [14] Jorge M Pacheco, Francisco C Santos, and Fabio AC C Chalub. Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput Biol*, 2(12):e178, 2006.
- [15] Arunas L Radzvilavicius, Alexander J Stewart, and Joshua B Plotkin. Evolution of empathetic moral evaluation. *Elife*, 8:e44269, 2019.
- [16] Arunas Radzvilavicius, Taylor Kessinger, and Joshua B Plotkin. How to design institutions that foster cooperation. *SocArXiv*. October, 10, 2019.
- [17] Alex Bradley, Claire Lawrence, and Eamonn Ferguson. Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences*, 285(1875):20180116, 2018.
- [18] Emel Filiz-Ozbay and Erkut Y Ozbay. Effect of an audience in public goods provision. *Experimental Economics*, 17(2):200–214, 2014.

- [19] Anya Savikhin Samek and Roman M Sheremeta. Recognizing contributors: an experiment on public goods. *Experimental Economics*, 17(4):673–690, 2014.
- [20] Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, 71(3):397–404, 2001.
- [21] Martin G Kocher, Todd Cherry, Stephan Kroll, Robert J Netzer, and Matthias Sutter. Conditional cooperation on three continents. *Economics letters*, 101(3):175–178, 2008.
- [22] Robert Boyd and Peter J Richerson. The evolution of reciprocity in sizable groups. *Journal of theoretical Biology*, 132(3):337–356, 1988.
- [23] Shinsuke Suzuki and Eizo Akiyama. Reputation and the evolution of cooperation in sizable groups. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570):1373–1377, 2005.
- [24] Tatsuya Sasaki, Isamu Okada, and Yutaka Nakai. The evolution of conditional moral assessment in indirect reciprocity. *Scientific reports*, 7(1):1–8, 2017.
- [25] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695):242–245, 2018.
- [26] Peter D Taylor and Leo B Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978.
- [27] Martin A Nowak and Karl Sigmund. Evolutionary dynamics of biological games. *science*, 303(5659):793–799, 2004.
- [28] Arne Traulsen, Jorge M Pacheco, and Martin A Nowak. Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of theoretical biology*, 246(3):522–529, 2007.
- [29] Martin A Nowak and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298, 2005.
- [30] Craig McIntosh, Elisabeth Sadoulet, Steven Buck, and Tomas Rosada. Reputation in a public goods game: Taking the design of credit bureaus to the lab. *Journal of Economic Behavior & Organization*, 95:270–285, 2013.
- [31] Robert Boyd and Sarah Mathew. Arbitration supports reciprocity when there are frequent perception errors. *Nature Human Behaviour*, pages 1–8, 2021.