The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# Social Norms of Cooperation with Costly Reputation Building

**Fernando P. Santos,[1,4] Jorge M. Pacheco,[2,3,4] Francisco C. Santos[1,4]**
[1] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal
[2] Centro de Biologia Molecular e Ambiental, Universidade do Minho, 4710-057 Braga, Portugal
[3] Departamento de Matemática e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal
[4] ATP-group, P-2744-016 Porto Salvo, Portugal

## Abstract

Social norms regulate actions in artificial societies, steering collective behavior towards desirable states. In real societies, social norms can solve cooperation dilemmas, constituting a key ingredient in systems of indirect reciprocity: reputations of agents are assigned following social norms that identify their actions as good or bad. This, in turn, implies that agents can discriminate between the different actions of others and that the behaviors of each agent are known to the population at large. This is only possible if the agents report their interactions. Reporting constitutes, this way, a fundamental ingredient of indirect reciprocity, as in its absence cooperation in a multiagent system may collapse. Yet, in most studies to date, reporting is assumed to be cost-free, which collides with many life situations, where reporting can easily incur a cost (costly reputation building). Here we develop a new model of indirect reciprocity that allows reputation building to be costly. We show that only two norms can sustain cooperation under costly reputation building, a feature that requires agents to be able to anticipate the reporting intentions of their opponents, depending sensitively on both the cost of reporting and the accuracy level of reporting anticipation.

## Introduction

Social norms are a cornerstone of human societies, being a fundamental mechanism to solve coordination (Young 2015), cooperation (Fehr and Fischbacher 2004) and collective action problems (Nyborg et al. 2016). In general, social norms are public and establish an expected pattern of behavior. When violated, they may lead to responses that range from gossip to open censure, ostracism, or dishonor for the transgressor (Bicchieri 2005). Examples range from bargaining norms – determining the behavior of buyers and sellers – to ancient practices such as foot binding in China or dueling in Europe (Young 2015). In artificial intelligence (AI), social norms have gathered special attention, constituting an appealing tool that can be used to efficiently steer behaviors towards desirable states (Wooldridge 2009).

Research on social norms in artificial societies is usually divided in a top-down approach – in which norms are designed offline and imposed in agents by a central authority – or a bottom-up approach – in which norms are studied as phenomena emerging after the interaction of a large number of agents (Wooldridge 2009). In the former case, norms directly constrain agents' possible actions, constituting what were named *social laws* (Shoham and Tennenholtz 1995). In the second case, norms are conceived as globally adopted behaviors, emerging due to local interactions of agents, and constituting the so-called *social conventions* (Shoham and Tennenholtz 1997). Often, however, norms that prevail in a society only enforce behaviors indirectly, functioning as a top-down mechanism that influences the bottom-up adherence (or not) to certain behaviors. This is particularly evident when systems of reputations are used to enforce social norms (Castelfranchi, Conte, and Paolucci 1998): acting in a certain way may provide a reputation uplift/downgrade whose tangible effect emerges in the future, as a form of reciprocation. In fact, enforcing behaviors indirectly, through norms and reputations, underlies Indirect Reciprocity (IR), known as a fundamental mechanism for the evolution of cooperation among humans (Nowak and Sigmund 2005).

Besides originating in the scope of evolutionary biology and economics, IR is particularly relevant for AI: first, it allows studying incipient moral values and ethical principles in computational environments, providing clues for the formalization of artificial ethics and morality (Greene et al. 2016); second, IR has been claimed as a cornerstone behind the evolution of human language and intelligence (Nowak and Sigmund 2005); third, reputation mechanisms play a central role in multiagent systems (Pinyol and Sabater-Mir 2013) with special applications in present-day online platforms that sustain high rates of cooperation between its users (Ho et al. 2012).

Despite promising, IR shares a fundamental challenge with other reputation systems working as enforcement mechanisms of social norms: they require observability (Haynes et al. 2017), *i.e.*, agents able to observe the behaviors of their peers. With rare exceptions (Suzuki and Kimura 2013; Sasaki, Okada, and Nakai 2016), previous models typically assume that observability is an exogenous factor – for instance, an adjustable parameter that controls the observability level in the system (Ohtsuki, Iwasa, and Nowak 2015). In reality, however, accessing the information about a private interaction depends on the decision of the agents involved that may share (or not) its outcome. As an example, in *e-commerce* or *p2p* platforms, private in-

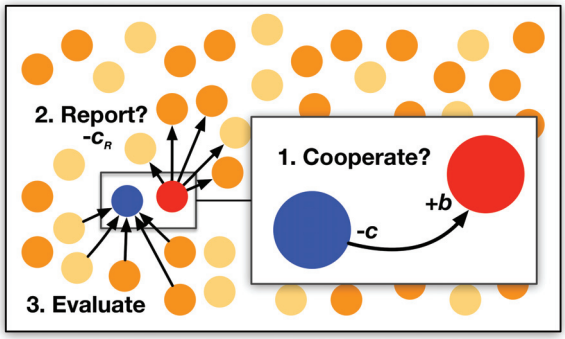


Figure 1: IR with costly reputation building: Pairs of agents interact following a random structure. In each interaction, agents play the donation game - a Donor may Cooperate (paying cost $c$ to provide benefit $b$, $b > c > 0$) or not with a Recipient. Secondly, the Recipient decides to Report (paying $c_R > 0$) (or not) the outcome of the interaction. Thirdly, society uses any reported information to attribute a new public reputation to the Donor. When both agents act as Donor and Recipient simultaneously, this interaction leads to the well-known Prisoner's Dilemma.

teractions take place and the individuals need to be incentivized to rate their opponents, that is, to provide information about their actions. This naturally involves time and effort. When the process of information sharing is costly, reporting is hardly fulfilled by rational agents, such that the system of IR – and the own sustainability of cooperation – may collapse. Information reporting thus constitutes a second-order free-rider problem: everyone would benefit from an IR system yet, its maintenance is costly, and it is tempting to avoid the burden (Rand and Nowak 2013; Suzuki and Kimura 2013).

Here we address the problem of costly reputation building in multiagent systems and the sustainability of cooperation. We pose three main questions:

1. Will cooperation emerge if reputation building is costly?
2. Which social norms excel in promoting cooperation, if reputation building is costly?
3. Which factors preclude cooperation in this scenario?

To answer these questions, we develop a model based on evolutionary game theory (EGT) (Sigmund 2010) in which agents play with each other the donation game described in Fig. 1 and revise their behaviors through social learning. Moreover, agents have a reputation ($Good$, $G$ or $Bad$, $B$). After each interaction, the reputation of an agent X who decided $C$ or $D$ against another agent Y may change depending on the social norm at work, that is, the rule that attributes a new reputation to X (G/B), given the actions and characteristics of X and Y. The reputation update will only occur provided that Y reported the outcome of the interaction (at a cost of reporting $c_R > 0$). This prototypical interaction, depicted in Fig. 1, allows us to study the conditions under which the willingness a) to report the outcome of an interaction and b) to discriminate between reputations and cooperate, are able co-evolve. Additionally, we confer to agents the possibility to anticipate the intention of their opponents to report (or not) information about the interaction. Whereas anticipative decision-making often requires complex architectures to enable the prediction of future outcomes (Domingos, Burguillo, and Lenaerts 2017), in the context of, *e.g.*, *e-commerce* platforms and, in general, artificial agent societies with reputation systems, this can be achieved by making publicly available (and highlighting) the previous reviews of agents.

Within this model, we show that, under costly reputation building ($c_R > 0$), 1) the mechanism of *reporting anticipation* suffices to sustain cooperation under IR; 2) only two social norms, stern judging (*an agent is G if cooperates with G and defect with B; all else is B*) and simple standing (*an agent is G if cooperates; defecting with a B opponent is justified*) are able to promote cooperation and 3) the cooperation that emerges relies on both the low cost of reporting ($c_R$) and the capacity to anticipate errors ($\tau$).

## Related Work

Social norms play a pivotal role in steering collective behavior in multiagent societies, thus constituting a prolific research field within artificial intelligence (Dignum 1999; Wooldridge 2009; Savarimuthu, Arulanandam, and Purvis 2011; Haynes et al. 2017). In the context of multigent systems (MAS), social norms were divided in two main classes (Villatoro, Sen, and Sabater-Mir 2010): *conventional* – those used to establish a convention, typically solving coordination dilemmas (Shoham and Tennenholtz 1997; Sen and Airiau 2007; Morales et al. 2013); and *essential* – those that seek to solve cooperation dilemmas and collective action problems (Griffiths and Luck 2010; Ho et al. 2012; Peleteiro, Burguillo, and Chong 2014; Santos, Santos, and Pacheco 2016). As evidenced below, here we shall focus on *essential* norms that dictate the expected behavior of agents in a cooperation dilemma, by attributing them reputations.

Works on social norms are traditionally divided in those focusing on 1) a top-down approach – also denoted legalistic (Villatoro, Sen, and Sabater-Mir 2010) or prescriptive (Savarimuthu, Arulanandam, and Purvis 2011) – in which norms are designed off-line and imposed by a central authority (Shoham and Tennenholtz 1995), or 2) a bottom-up approach – also called interactionist (Villatoro, Sen, and Sabater-Mir 2010) – in which norms are studied as emergent phenomena. As an example of top-down approach, the seminal work of Shoham and Tennenholtz focused on ensuring cooperative behavior in a multiagent system through the implementation of social laws, *i.e.*, rules that explicitly constrain the behavior of agents (Shoham and Tennenholtz 1995). Shaping the environment in which agents interact in a top-down fashion, such that some behaviors are directly restricted, originates the so-called Electronic Institutions (García-Camino, Noriega, and Rodríguez-Aguilar 2005). Regarding bottom-up works, Shoham and Tennenholtz originally combined ideas from economics and AI in order to analyze the local strategy update rules that would lead to success in solving a coordination dilemma (Shoham and Tennenholtz 1997). In particular, the authors found that

*highest cumulative reward* (HCR) stands as a remarkably successful rule, guaranteeing the eventual emergence of coordination into the cooperative equilibria. More recently, Sen and Airiau showed that a mechanism of social learning – in which strategies are updated following individual information gathered after the interaction with many opponents – also excels in allowing the evolution of stable social norms, i.e, reaching a state in which (almost) all agents play the same payoff-maximizing policy (Sen and Airiau 2007). This work was later extended to consider complex networks of interaction (Airiau, Sen, and Villatoro 2014). Often, however, evaluating the effect of a social norm involves a combination of both approaches: even if norms apply to a given multiagent system in a top-down fashion, their effectiveness can only be computed in terms of a self-organizing, decentralized and bottom-up process.

Here we follow this view. We evaluate the benefits that a norm provides at the system level (Haynes et al. 2017), in terms of achieved cooperation, depending on a self-organized process in which agents learn to use a given policy. We consider reputations as the prevailing norm enforcement mechanism (Savarimuthu, Arulanandam, and Purvis 2011). In this context, we shall highlight the work of Liu et al., that used EGT in order to evaluate the robustness of incentive mechanisms based on reputations (Liu et al. 2016). Also Ho et al. studied social norms and reputations as a mechanism to incentivize cooperation in the context of crowdsourcing markets (Ho et al. 2012).

In a broad and multidisciplinary scope, the relationship between norms, reputations and cooperation has been mathematically linked in models of Indirect Reciprocity (IR) (Nowak and Sigmund 2005). IR was pointed as the most cognitively demanding mechanism of cooperation discovered so far (Nowak and Sigmund 2005). The relation between cooperation and IR has been addressed within the multiagent systems community. Peleteiro, Burguillo, and Chong showed that cooperation under IR is boosted by coalitions and agents able to change their neighbors. In that work, the reputation of agents increases anytime they cooperate (Peleteiro, Burguillo, and Chong 2014). Often, however, the update of reputations also depends on the agents against whom actions are directed to. The notion of social norm is central at this point, as the reputation changes depend on the adopted social norms that define what actions (and in which contexts) are reckoned as Good or Bad.

In the context of IR, Ohtsuki and Iwasa, in their seminal work, extensively studied the potential of social norms in leading agents to adopt cooperative strategies. In that work, a social norm prescribes a new reputation given the action and reputation of a Donor and the reputation of the Recipient. The authors found that, remarkably, only 8 norms (out of 2080) were able to guarantee the stability of cooperative strategies. While Ohtsuki and Iwasa assume an infinite population, here we assume that the population contains a finite number of agents (Santos, Santos, and Pacheco 2016), whereas social norms rely on the action of a Donor and the reputation of the Recipient to define a new reputation to the Donor.

Finally, it is worth pointing out that EGT was recently employed in the multiagent systems community to evaluate the stability of normative systems (Morales et al. 2017), *i.e.*, systems in which norms are used to directly regulate agents' actions, and also to study how social norms emerge and change over time (De, Nau, and Gelfand 2017).

# Model

## Background

We consider a finite population of $Z$ agents who have the option to help (that is, to Cooperate, $C$) or not (to Defect, $D$) another agent. Random pairs of agents are chosen and play the donation game, one acting as Donor and the other as Recipient. After playing the donation game, the Recipient may decide to publicly share the outcome of this interaction (that is, to Report, $R$) or not (to remain Silent, $S$). To report incurs a cost $c_R$ to the Recipient. Clearly, the advantages of cooperating can only exist indirectly, to the extent that a cooperative act leads to a reputation that allows future benefits. Reporting contributes to build these reputations. We assume that the reporting intentions of an agent can be anticipated by others, although we allow for errors of anticipation. As a result, an agent is regarded by opponents as belonging to one of four classes – 1) with $G$ (Good) reputation and willing to Report the outcome of an encounter (G/R); 2) with $G$ reputation yet unwilling to Report (Silent, G/S); 3) with $B$ (Bad) reputation and willing to Report (B/R) and 4) B/S.

The strategy of each agent constitutes a policy that dictates the probability of cooperating/reporting when interacting with different opponents, both in the role of Donor ($C$ or $D$) and Recipient ($R$ or $S$). This way, a strategy is fully defined as $\mathbf{p} = (\mathbf{p_d}, p_r)$, where $p_r$ defines the behavior as Recipient (probability of $R$) and $\mathbf{p_d} = (p_{GR}, p_{GS}, p_{BR}, p_{BS})$ is the counterpart that translates the behavior as Donor (*i.e.*, probability of $C$), given the four possible classes of opponent. We consider pure strategies (probabilities are either 0 or 1) with a small perturbation $\epsilon$, often called execution error. This error simulates the inability of individuals to act in the way that their strategy dictates. It is common practice to consider errors in the form of failed intended cooperation, due, for instance, the lack of resources to or opportunity donate (Santos, Santos, and Pacheco 2016). Concerning the Recipient, execution errors imply either an unintended report or an unintended silence. The effective strategy will thus read $\mathbf{p}_\epsilon = \mathbf{p} \circ (1-\epsilon, 1-\epsilon, 1-\epsilon, 1-\epsilon, 1-2\epsilon) + (0, 0, 0, 0, \epsilon)$ (where $\circ$ is the element-wise product). Understanding the effect of a given social norm will involve analysing the evolutionary dynamics between all these $2^5 = 32$ strategies. For reasons that will be clear below, it suffices to analyse the interplay between pairs of strategies (Santos, Santos, and Pacheco 2016).

## Probability of cooperation and payoff calculation

Assuming that two action rules ($\mathbf{p}$ and $\mathbf{p}'$) may exist simultaneously in the population – $k$ agents use $\mathbf{p}$ and $Z-k$ agents use $\mathbf{p}'$ – and given that $h$ and $h'$ of the individuals that use, respectively, action rules $\mathbf{p}$ and $\mathbf{p}'$, have a $G$ reputation, the probability that an individual $\mathbf{p}$ is regarded by others as one of the four different classes mentioned – 1) G/R; 2) G/S; 3)

B/R; 4) B/S – is given by,

$$\mathbf{s}^{\mathbf{p}}_{\chi,\tau}(k,h,h') = E_{\chi,\tau}.[\frac{h}{k}p_r, \frac{h}{k}\bar{p_r}, \frac{k-h}{k}p_r, \frac{k-h}{k}\bar{p_r}]^{\mathbf{T}} \tag{1}$$

where we use $\bar{p_r} = 1 - p_r$. A stochastic error matrix ($E_{\chi,\tau}$) allows us to incorporate both private errors of assessment ($\chi$) and the error of not anticipating accurately the opponent intention to report ($\tau$, hereafter called anticipation error). Denoting $(1-\chi)$ and $(1-\tau)$ as $\bar{\chi}$ and $\bar{\tau}$, $E_{\chi,\tau}$ reads as

$$E_{\chi,\tau} = \begin{bmatrix} \bar{\chi}\bar{\tau} & \bar{\chi}\tau & \chi\bar{\tau} & \chi\tau \\ \bar{\chi}\tau & \bar{\chi}\bar{\tau} & \chi\tau & \chi\bar{\tau} \\ \chi\bar{\tau} & \chi\tau & \bar{\chi}\bar{\tau} & \bar{\chi}\tau \\ \chi\tau & \chi\bar{\tau} & \bar{\chi}\tau & \bar{\chi}\bar{\tau} \end{bmatrix} \tag{2}$$

The probability that any individual ($\mathbf{p}$) regards an opponent ($\mathbf{p}$ or $\mathbf{p}'$) as belonging to each of the four classes can be calculated as,

$$\mathbf{s}_{\chi,\tau}(k,h,h') = E_{\chi,\tau}. \begin{bmatrix} \frac{h}{Z-1}\frac{k-1}{k}p_r + \frac{h'}{Z-1}p'_r \\ \frac{h}{Z-1}\frac{k-1}{k}\bar{p_r} + \frac{h'}{Z-1}\bar{p'_r} \\ \frac{k-h}{Z-1}\frac{k-1}{k}p_r + \frac{Z-1-k-h'}{Z-1}p'_r \\ \frac{k-h}{Z-1}\frac{k-1}{k}\bar{p_r} + \frac{Z-1-k-h'}{Z-1}\bar{p'_r} \end{bmatrix} \tag{3}$$

We may thus conveniently define the probability that an individual with action rule $\mathbf{p}$ cooperates as

$$C_{\mathbf{p}}(k,h,h') = \mathbf{p_d}.\mathbf{s}_{\chi,\tau}(k,h,h') \tag{4}$$

and the probability that anyone cooperates with an individual $\mathbf{p}$ as

$$\bar{C_{\mathbf{p}}}(k,h,h') = (\frac{k-1}{Z-1}\mathbf{p}_d + \frac{Z-k}{Z-1}\mathbf{p'}_d).\mathbf{s}^{\mathbf{p}}_{\chi,\tau}(k,h,h'). \tag{5}$$

For notational simplicity, we will use $\sigma^{\mathbf{p}} \equiv \mathbf{s}^{\mathbf{p}}_{\chi,0}$ and $\sigma \equiv \mathbf{s}_{\chi,0}$ to denote the probabilities of effectively facing a reporter (and not only perceiving that, *i.e*, $\tau = 0$). The payoff of strategy $\mathbf{p}$ (when there are $k$ $\mathbf{p}$ and $Z-k$ $\mathbf{p}'$) is computed as,

$$\Pi_{\mathbf{p},\mathbf{p}'}(k,h,h') = b\bar{C_{\mathbf{p}}}(k,h,h') - cC_{\mathbf{p}}(k,h,h') - c_R p_r \tag{6}$$

where $b$ is the benefit of receiving a cooperation, $\bar{C_{\mathbf{p}}}$ is the probability that anyone cooperates with $\mathbf{p}$, $C_{\mathbf{p}}$ is the probability that $\mathbf{p}$ cooperates with anyone, $c$ is the cost of cooperating and $c_R$ is the cost of reporting.

## Dynamics of reputations

The reputation dynamics occurs under the influence of a social norm, $\mathbf{d} = (d_{GC}, d_{GD}, d_{BC}, d_{BD})$, that dictates the new reputation of a Donor depending on whether she Cooperates or Defects against a Recipient that is either Good ($G$) or Bad ($B$). Here, $d_{ij}$ is the probability that a reputation becomes $G$ after a reported interaction, in which a Donor takes action $j$ against a Recipient with reputation $i$. The *rationale* is that, when the Recipient reports publicly the action of the Donor, the society will judge that action taking the reputation of the Recipient into account as well. In general, we consider that $d_{ij} = \{0 = B, 1 = G\}$; however, we admit assignment errors, with probability $\alpha$, that reflect the likelihood that the wrong reputation is attributed to the Donor, given the possibility that i) the information reported is misinterpreted or ii) a wrong assessment of the Recipient reputation is made. The effective norm will thus read as $\mathbf{d}_\alpha = \mathbf{d}(1-2\alpha) + \alpha$. The probability of assigning a reputation $G$ to an individual using $\mathbf{p}$ is therefore,

$$g_{\mathbf{p}}(k,h,h') = \mathbf{d}.[\mathbf{p}_G\sigma_{GR} + \frac{\sigma_{GS}}{2}, \bar{\mathbf{p}}_G\sigma_{GR} + \frac{\sigma_{GS}}{2}, \mathbf{p}_G\sigma_{BR} + \frac{\sigma_{BS}}{2}, \bar{\mathbf{p}}_G\sigma_{BR} + \frac{\sigma_{BS}}{2}] \tag{7}$$

where $\mathbf{p}_G$ ($\mathbf{p}_B$), calculated using $s_{\chi,\tau}$, is the probability that $\mathbf{p}$ meets and cooperates with a $G$ ($B$) ($\bar{p}_G$ ($\bar{p}_B$) being the probability of not doing so) and $\sigma_{GR}$ is the probability of effectively facing an individual that is $G$ and that will Report. Naturally, the reputation assessment only depends on the action rule when one faces an opponent that is willing to report ($R$). Oppositely, when facing someone that prefers to be silent ($S$), others will not know the action employed. Here we assume that, with probability $0.5$ the action was Cooperation. Different assumptions concerning this last point can naturally be tested in further works. The probability of assigning $B$ reads similarly as,

$$b_{\mathbf{p}}(k,h,h') = (1-\mathbf{d}).[\mathbf{p}_G\sigma_{GR} + \frac{\sigma_{GS}}{2}, \bar{\mathbf{p}}_G\sigma_{GR} + \frac{\sigma_{GS}}{2}, \mathbf{p}_G\sigma_{BR} + \frac{\sigma_{BS}}{2}, \bar{\mathbf{p}}_G\sigma_{BR} + \frac{\sigma_{BS}}{2}] \tag{8}$$

Eqs. (7) and (8) allow us to write the probabilities of having one more $G$ with strategy $\mathbf{p}$ or one more $B$ with strategy $\mathbf{p}$ in the population through a one-step process. They are given by $G^+_{\mathbf{p}} = \frac{k-h}{Z}g_{\mathbf{p}}$ and $G^-_{\mathbf{p}} = \frac{h}{Z}b_{\mathbf{p}}$, respectively. With these ingredients, we may now define a Markov process with associated stochastic matrix $H$ (where, *e.g.*, $H_{(h,h')\to(h+1,h')} = G^+_{\mathbf{p}}$), and whose stationary distribution ($\lambda^{rep}$) is given by the normalised eigenvector associated with the eigenvalue 1 of the transposed of $H$. $\lambda^{rep}_{h,h'}$ represents the fraction of time spent, on average, having $h$ $G$-agents adopting $\mathbf{p}$ and $h'$ $G$-agents adopting $\mathbf{p}'$. With $\lambda^{rep}$ we can calculate $f_{\mathbf{p},\mathbf{p}'}(k)$, the fitness value of an agent adopting $\mathbf{p}$, when there are $k$ agents adopting $\mathbf{p}$ and $Z-k$ adopting $\mathbf{p}'$:

$$f_{\mathbf{p},\mathbf{p}'}(k) = \sum_{h,h'} \lambda^{rep}_{h,h'} \Pi_{p,p'}(k,h,h') \ . \tag{9}$$

## Dynamics of strategies

To model the dynamical behavior of agents when two strategies are present in the population, we adopt a social learning process implemented via the pairwise comparison rule [as in (Santos, Santos, and Pacheco 2016; Santos, Pacheco, and Santos 2016; De, Nau, and Gelfand 2017; Han et al. 2017)], where the probability ($p_{i,j}$) that agent $j$ imitates $i$ increases with the fitness difference $\Delta_{i,j}(k) = f_{i,j}(k) - f_{j,i}(Z-k)$ following $p_{i,j} = (1 + e^{-\Delta_{i,j}(k)})^{-1}$. Assuming that two agents are randomly sampled from the population in which $k$ agents use strategy $i$ and $Z-k$ use strategy $j$, the probability of having $\pm 1$ agent using strategy $i$ is given by

$$T^{\pm}(k) = \frac{Z-k}{Z}\frac{k}{Z-1}(1 + e^{\mp\Delta_{i,j}(k))})^{-1} \tag{10}$$

Note that $\frac{Z-k}{Z}$ ($\frac{k}{Z-1}$) represent the sampling probabilities of choosing one agent with strategy $j$ ($i$). Additionally, with probability $\mu$, a "mutation" occurs and individuals change their strategy to a random one, exploring a new behavior. Thus, the imitation process described above occurs with probability $(1-\mu)$. In the limit of *rare* mutations, *i.e.,* $\mu \to 0$ (Fudenberg and Imhof 2006; Santos, Santos, and Pacheco 2016), we are able to derive analytical insights from this model. It turns out that these insights remain valid over a much wider interval of mutation regimes (Santos, Pacheco, and Santos 2016). Moreover, while this assumption reduces the random exploration of behaviors, it does not prevent us to consider other stochastic effects, as $\chi$ or $\tau$ defined above. Under this *rare mutation* regime, any *mutant* strategy will either fixate in the population or will become extinct (Fudenberg and Imhof 2006), as the time between two mutation events will be so large that the population will always evolve to a monomorphic state (*i.e.*, all agents using the same strategy) before the next mutation occurs. Thus, the dynamics can be approximated by means of an embedded Markov Chain whose configuration states correspond to the different monomorphic states of the population. This fact allows us to conveniently use the payoff functions defined in Eq. (6) in the calculation of the transition probabilities. In this context, the time spent in polymorphic configurations is merely transient, being disregarded (Fudenberg and Imhof 2006). The transitions between states of the embedded chain are obtained through the fixation probability of every single mutant of strategy $i$ in every resident population of strategy $j$, reflecting how easy it is for a strategy originated by a rare mutation to fixate in a population. A strategy $i$ will fixate in a population composed by $Z-1$ agents using strategy $j$ with a probability given by (Sigmund 2010):

$$\rho_{i\to j} = \left(\sum_{l=0}^{Z-1}\prod_{k=1}^{l}\frac{T^-(k)}{T^+(k)}\right)^{-1} \tag{11}$$

These probabilities define the stochastic matrix $T$ ($T_{i,j} = \rho_{i\to j}$) associated with the embedded Chain described above, and whose stationary distribution $\lambda^{str}$ is, as usual, given by the normalized eigenvector associated with the eigenvalue 1 of the transposed of $T$. $\lambda^{str}_{\mathbf{p}}$ represents the fraction of time spent, on average, in a state where all agents use strategy $\mathbf{p}$. The total cooperation level ($\eta$) is given by the weighted average of the cooperation levels of the population in each monomorphic state,

$$\eta = \sum_{\mathbf{p},h} C_{\mathbf{p}}(Z,h,0)\lambda^{str}_{\mathbf{p}}\lambda^{rep}_{h,0} \tag{12}$$

## Results

Employing the framework just described, we now investigate the three main research questions. We find that the capacity of agents to anticipate the reporting intentions of their opponents is sufficient to allow cooperation to emerge in a context of costly reputation building. This, however, happens only under specific social norms. As Fig. 2 conveys, there are social norms that efficiently allow cooperation to be sustained. In particular, *stern judging* (SJ, $d = (1,0,0,1)$,

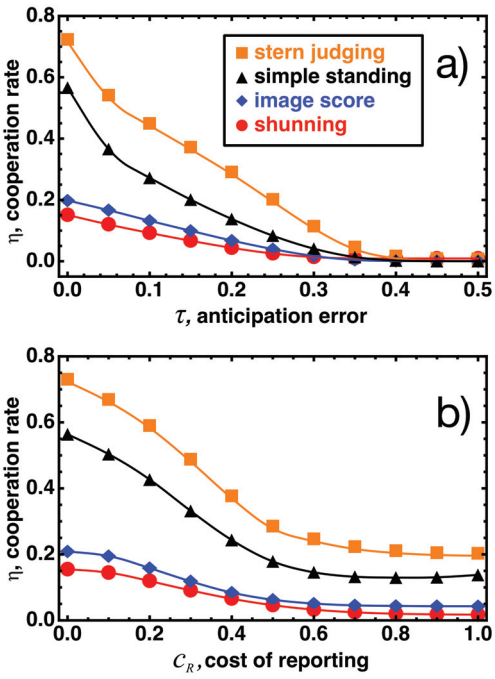


Figure 2: Cooperation emerges even if reputation building is costly. Stern judging is the social norm that allows the highest values of cooperation, followed by simple standing. $Z = 50$, $b = 5$, $c = 1$, $c_R = 0.1$, $\chi = \epsilon = \alpha = \tau = 0.01$ (when not explicitly varied).

or 9 in decimal notation: An agent is $G$ if cooperates with $G$ and defects with $B$; all else is $B$) is the one leading to the highest values of cooperation, regardless of the anticipation error $\tau$ and the reporting cost $c_R$. A more benevolent norm called *simple standing* (SS, $d = (1,0,1,1)$, or 11: An agent is $G$ if cooperates; defecting with a $B$ opponent is justified), is the one promoting the second highest levels of cooperation. *Image score* ($d = (1,0,1,0)$, or 10: Whoever Cooperates is $G$), together with *shunning* ($d = (1,0,0,0)$, or 8: Only to those that cooperate with $G$ become $G$) fail to promote levels of cooperation higher than 20%. The remaining norms – that we will refer to using their decimal identification – promote cooperation levels that match those already depicted in Fig. 2: norm 1 is quantitatively equivalent to SS due to mirror symmetry (indeed, we can switch $B$ and $G$ everywhere and the same results would ensue (Santos, Santos, and Pacheco 2016)); norms 0, 4, 8, 13, 14 and 15 lead to the same results as *shunning*; norms 2, 3, 5, 7, 10 and 12 promote the same levels of cooperation as *image score* while norm 6 promotes cooperation levels slightly above *image score*. For all 16 norms studied, it is clear that high levels of $\tau$ and $c_R$ are detrimental for cooperation.

The negative effect of $\tau$ and $c_R$ on cooperation is further evidenced in Fig. 3, where both quantities are simultaneously varied for the particular cases of SJ and SS. It becomes clear that 1) agents should report their interactions and make that intention known and 2) the means to report should be

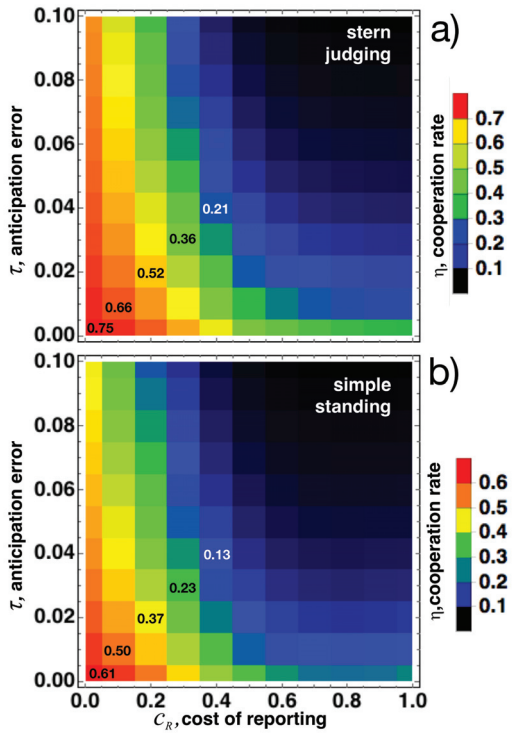


Figure 3: For **a)** stern judging and **b)** simple standing, cooperation heavily relies on the smallness of $\tau$ and $c_R$. $Z = 50$, $b = 5$, $c = 1$, $\chi = \alpha = \epsilon = 0.01$.

simplified, such that $c_R$ is effectively low. Nonetheless, under a simple norm such as SJ, cooperation exists even when the cost of reporting represents 20% of the actual cooperation cost, and anticipation errors are of the order of 10%.

Finally, we explore in the following the strategy dynamics that sustains the cooperation levels shown in Figs. 2 and 3. We consider the cases when the norms governing reputation assignment are SJ and SS. Figs. 4 and 5 depict the stationary distribution of strategies, in this case providing information on the fraction of time spent in each possible state where all agents adopt a given strategy ($\lambda_{\mathbf{p}}^{str}$, see Model section). There are 32 possible strategies: 16 in which reporting occurs (panels a, right) and 16 in which not (panels a, left). For a better interpretation, this information is represented in bars whose colours depict the average fraction of agents with a $G$ reputation (with a blend $\lambda_x^{rep} Blue + (1 - \lambda_x^{rep}) Red$, where $\lambda_x^{rep}$ gives the fraction of $G$ individuals in each state). We also depict the transitions occurring between the six most prevailing states (panels b). Transitions are calculated using Eq. (11), corresponding to the fixation probability of one mutant (with strategy associated with the endpoint on the depicted arrow) in a population previously composed by agents adopting the strategy located at the starting point of the arrow. We only represent the transitions whose value is higher than the neutral drift transition ($1/Z$), that would correspond to the fixation of a mutant strategy with the exact same fitness as the resident one.

In Fig. 4 we show that, for SJ, agents spend most of their time (and with equal proportion) either in a state where everyone uses strategy $\mathbf{p} = (DDCD|R)$ or in a state where everyone uses strategy $\mathbf{p} = (CDDD|R)$. In other words, Reporting is a stable behavior. This only occurs because those strategies that discriminate the reporting intentions of agents, besides reputations, are stable. A population adopting $\mathbf{p} = (DDDD|R)$ will naturally evolve to a state where everyone adopts $\mathbf{p} = (DDDD|S)$, as reporting, being costly, is not properly rewarded with cooperation. From $\mathbf{p} = (DDDD|S)$, the population will evolve either to $\mathbf{p} = (DDCD|R)$ or $\mathbf{p} = (CDDD|R)$. This duality occurs given the particular symmetry of SJ. In fact, this norm is able to sustain two highly cooperative (monomorphic) states, with opposite definitions of $G$ and $B$: 1) Cooperate with $G$ and remain $G$ or 2) Cooperate with $B$ and remain $B$. Once a population reaches the state $\mathbf{p} = (CDDD|R)$ (or $\mathbf{p} = (DDCD|R)$), the only favourable transition is towards $\mathbf{p} = (CCDD|R)$ (or $\mathbf{p} = (DDCC|R)$). The transition probabilities are low, however: Since everyone Reports, it becomes almost indistinguishable to Cooperate or not with those that are GR or GS (or BR or BS). In fact, the transition $(CDDD|R) \rightarrow (CCDD|R)$ is only higher than $1/Z$ due to the existence of errors. Given that Reporters may have actually remained Silent in the past (due to execution error $\epsilon$) and given that donors may wrongly anticipate the Silent intentions of their opponents (due to anticipation error $\tau$), it is marginally beneficial to cooperate also with those that are anticipated to be Silent, avoiding the surprise of refusing help to a Reporter and guaranteeing that a good reputation is always maintained.

In Fig. 5 we focus on SS. We show that, with this norm, agents spend most of their time in a state where everyone adopts $\mathbf{p} = (CDDD|R)$. Again, Reporting constitutes a stable behavior and agents learn to discriminate based on the anticipated reporting intention and reputation. As with SJ, once the population reaches $\mathbf{p} = (CDDD|R)$, the only favourable transition is towards $\mathbf{p} = (CCDD|R)$.

## Conclusion

Here we investigate whether indirect reciprocity can promote cooperation when reputation building is costly. We pose three main questions: If reputation building is costly, 1) will cooperation emerge? 2) which social norms excel in promoting cooperation? 3) which factors preclude the emergence of cooperation? To answer these questions, we developed an evolutionary game theoretical model which describes the dynamics of strategy adoption when the reputation of agents is governed by different social norms. Importantly, this new model allows us to understand which social norms promote cooperation, and why. We conclude that cooperation can emerge with indirect reciprocity, even if reputation building is costly, provided agents are able to anticipate the reporting intentions of their opponents. We also conclude that cooperation is able to emerge when the social norms SJ or SS govern reputation assignment in a population of agents. Under SJ, two highly cooperative states are remarkably prevalent: one in which agents report, cooperate with $G$ label opponents and remain $G$; other in which

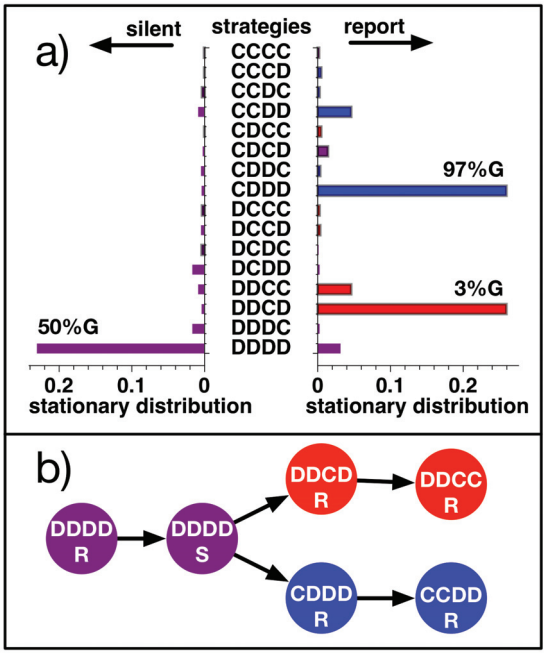


Figure 4: Stationary distribution and evolutionary dynamics under stern judging. Panel **a)** depicts the stationary distribution over each (monomorphic) state, and also the distribution of reputations per state. We use the blend $\gamma_x Blue + (1 - \gamma_x) Red$ (where $\gamma_x$ gives the fraction of $G$) to display the reputation distribution, a piece of information also indicated numerically in some cases. Panel **b)** depicts transitions between the six most prevailing states. An arrow is included whenever the transition probability between two states is higher than $1/Z$ (neutral transition probability). $Z = 50$, $b = 5$, $c = 1$, $c_R = 0.2$, $\chi = 0$, $\epsilon = \alpha = \tau = 0.01$.

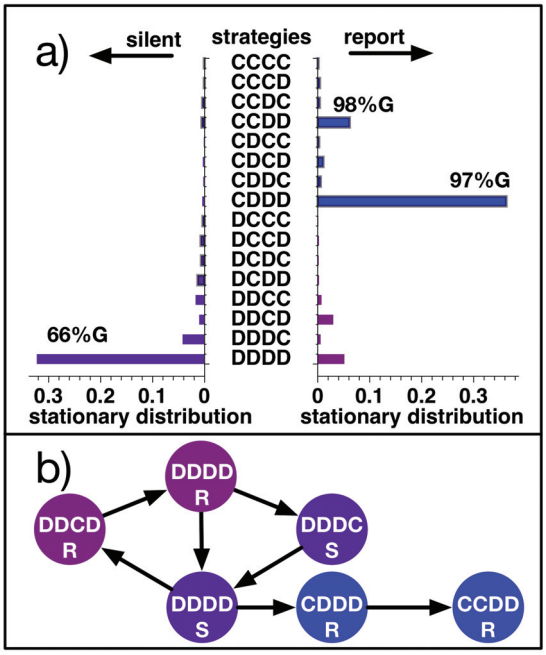


Figure 5: Stationary distribution and evolutionary dynamics under simple standing; we use the same notation of Fig. 4.

agents report, cooperate with $B$ label opponents and remain $B$; interestingly, the selected state is itself a *social convention* (Shoham and Tennenholtz 1997), that attributes a positive/negative valuation to a $G$ or $B$ label (Santos, Pacheco, and Santos 2016). We further find that cooperation under costly reputation building depends sensitively on the cost of reporting ($c_R$) and the accuracy of anticipating the reporting intentions of agents ($\tau$).

Even considering the simple case of binary reputations and discrete strategies, the results that we obtain nicely fit experimental studies, showing that cooperation in social dilemmas is conditioned on whether individuals recognize that their decisions will be known in the future (Semmann, Krambeck, and Milinski 2004). It was shown that cooperation declines when individuals believe that their actions will not be known by others, a situation that, in our case, is naturally dependent on accurately anticipating that the opponent will share the interaction outcome with others.

Our findings provide important insights regarding the design of social norms, enforced by reputation systems, in multiagent societies. In practice, we show that it is fundamental that agents are given the conditions to signal their intention of reporting an interaction. Particularly in web-platforms (Ho et al. 2012), this can be achieved by *e.g.* publicising the previous reviews or feedback provided by agents. Additionally, the cost of reporting (for instance accruing from the effort to write a review) can be alleviated by providing simple and intuitive feedback platforms or even by elucidating the indirect benefits of being a reporter.

Finally, we contribute with a novel analytical framework that allows studying the interplay between social norms and cooperation, while avoiding the burden of large-scale simulations. The framework models an environment in which interaction observability is costly and depends on the agents decisions, opening the opportunity for studying central aspects of social norms, reputation systems and cooperation. Future extensions may include, for example, the role of social norms that prevent/instigate malicious reports and lying (Savarimuthu, Arulanandam, and Purvis 2011), new incentives for honest reporting, or even scenarios where one takes into account the role of *bluffing*, when signalling the intention of reporting an action does not translate in an actual report (Santos, Pacheco, and Skyrms 2011). The presented framework can also accommodate other social dilemmas, such as those involving coordination, co-existence or public goods dilemmas.

## Acknowledgments

This research was supported by FCT-Portugal grants SFRH/BD/86465/2012, SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, UID/BIA/04050/2013, and UID/CEC/50021/2013.

## References


Airiau, S.; Sen, S.; and Villatoro, D. 2014. Emergence of conventions through social learning. *Auton Agent Multi Agent Syst* 28(5):779–804.

Bicchieri, C. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Castelfranchi, C.; Conte, R.; and Paolucci, M. 1998. Normative reputation and the costs of compliance. *J Artif Soc Soc Simulat* 1(3):3.

De, S.; Nau, D. S.; and Gelfand, M. J. 2017. Understanding norm change: An evolutionary game-theoretic approach. In *AAMAS'17*, 1433–1441. IFAAMAS.

Dignum, F. 1999. Autonomous agents with norms. *Artif Intell and Law* 7(1):69–79.

Domingos, E. F.; Burguillo, J.-C.; and Lenaerts, T. 2017. Reactive versus anticipative decision making in a novel gift-giving game. In *AAAI'17*, 4399–4405. AAAI Press.

Fehr, E., and Fischbacher, U. 2004. Social norms and human cooperation. *Trends Cogn Sci* 8(4):185–190.

Fudenberg, D., and Imhof, L. A. 2006. Imitation processes with small mutations. *J Econ Theory* 131(1):251–262.

García-Camino, A.; Noriega, P.; and Rodríguez-Aguilar, J. A. 2005. Implementing norms in electronic institutions. In *AAMAS'05*, 667–673. IFAAMAS.

Greene, J.; Rossi, F.; Tasioulas, J.; Venable, K. B.; and Williams, B. C. 2016. Embedding ethical principles in collective decision support systems. In *AAAI'16*, 4147–4151. AAAI Press.

Griffiths, N., and Luck, M. 2010. Changing neighbours: improving tag-based cooperation. In *AAMAS'10*, 249–256. IFAAMAS.

Han, T.; Pereira, L. M.; Martinez-Vaquero, L. A.; and Lenaerts, T. 2017. Centralized vs. personalized commitments and their influence on cooperation in group interactions. In *AAAI'17*, 2999–3005. AAAI Press.

Haynes, C.; Luck, M.; McBurney, P.; Mahmoud, S.; Vítek, T.; and Miles, S. 2017. Engineering the emergence of norms: a review. *Knowl Eng Rev* 32.

Ho, C.-J.; Zhang, Y.; Vaughan, J.; and Van Der Schaar, M. 2012. Towards social norm design for crowdsourcing markets. In *AAAI'12 Workshops, Human Computation*. AAAI Press.

Liu, Y.; Zhang, J.; An, B.; and Sen, S. 2016. A simulation framework for measuring robustness of incentive mechanisms and its implementation in reputation systems. *Auton Agent Multi Agent Syst* 30(4):581–600.

Morales, J.; Lopez-Sanchez, M.; Rodriguez-Aguilar, J. A.; Wooldridge, M.; and Vasconcelos, W. 2013. Automated synthesis of normative systems. In *AAMAS'13*, 483–490. IFAAMAS.

Morales, J.; Wooldridge, M.; Rodríguez-Aguilar, J. A.; and López-Sánchez, M. 2017. Evolutionary synthesis of stable normative systems. In *AAMAS'17*, 1646–1648. IFAAMAS.

Nowak, M. A., and Sigmund, K. 2005. Evolution of indirect reciprocity. *Nature* 437(7063):1291–1298.

Nyborg, K.; Anderies, J. M.; Dannenberg, A.; Lindahl, T.; Schill, C.; Schlüter, M.; Adger, W. N.; Arrow, K. J.; Barrett, S.; Carpenter, S.; et al. 2016. Social norms as solutions. *Science* 354(6308):42–43.

Ohtsuki, H., and Iwasa, Y. 2006. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J Theor Biol* 239(4):435–444.

Ohtsuki, H.; Iwasa, Y.; and Nowak, M. A. 2015. Reputation effects in public and private interactions. *PLOS Comput Biol* 11(11):e1004527.

Peleteiro, A.; Burguillo, J. C.; and Chong, S. Y. 2014. Exploring indirect reciprocity in complex networks using coalitions and rewiring. In *AAMAS'14*, 669–676. IFAAMAS.

Pinyol, I., and Sabater-Mir, J. 2013. Computational trust and reputation models for open multi-agent systems: a review. *Artif Intell Rev* 40(1):1–25.

Rand, D. G., and Nowak, M. A. 2013. Human cooperation. *Trends Cogn Sci* 17(8):413–425.

Santos, F. P.; Pacheco, J. M.; and Santos, F. C. 2016. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci Rep* 6(37517).

Santos, F. C.; Pacheco, J. M.; and Skyrms, B. 2011. Co-evolution of pre-play signaling and cooperation. *J Theor Biol* 274(1):30–35.

Santos, F. P.; Santos, F. C.; and Pacheco, J. M. 2016. Social norms of cooperation in small-scale societies. *PLoS Comput Biol* 12(1):e1004709.

Sasaki, T.; Okada, I.; and Nakai, Y. 2016. Indirect reciprocity can overcome free-rider problems on costly moral assessment. *Biol Lett* 12(7):20160341.

Savarimuthu, B. T. R.; Arulanandam, R.; and Purvis, M. 2011. Aspects of active norm learning and the effect of lying on norm emergence in agent societies. In *PRIMA'11*, 36–50.

Semmann, D.; Krambeck, H.-J.; and Milinski, M. 2004. Strategic investment in reputation. *Behav Ecol Sociobiol* 56(3):248–252.

Sen, S., and Airiau, S. 2007. Emergence of norms through social learning. In *IJCAI'07*, 1512. AAAI Press.

Shoham, Y., and Tennenholtz, M. 1995. On social laws for artificial agent societies: off-line design. *Artif Intell* 73(1-2):231–252.

Shoham, Y., and Tennenholtz, M. 1997. On the emergence of social conventions: modeling, analysis, and simulations. *Artif Intell* 94(1):139–166.

Sigmund, K. 2010. *The calculus of selfishness*. Princeton University Press.

Suzuki, S., and Kimura, H. 2013. Indirect reciprocity is sensitive to costs of information transfer. *Sci Rep* 3(1435).

Villatoro, D.; Sen, S.; and Sabater-Mir, J. 2010. Of social norms and sanctioning: A game theoretical overview. *Int J Agent Technol Syst* 2(1):1–15.

Wooldridge, M. 2009. *An introduction to multiagent systems*. John Wiley & Sons.

Young, H. P. 2015. The evolution of social norms. *Annu Rev Econom* 7(1):359–387.