Empathy, conformism and consensus in reputation-based cooperation with private information

Henrique Correia da Fonseca
GAIPS/INESC-ID and Instituto Superior Técnico,
Universidade de Lisboa, Portugal

Abstract
Explaining the emergence of cooperation remains an open challenge in theoretical ecology and human evolution. Indirect Reciprocity (IR) constitutes one of the most elaborate mechanisms that provide a way out of what is known as the cooperation dilemma, relying on concepts such as social norms and reputations. In the last decades, several social norms have been identified as active promoters of cooperation in reputation-based settings. However, such norms often rely on the assumption that every action is observed by every member of a population, an assumption that cannot always be translated into real scenarios. A consequence of lifting this assumption is that reputations cease to be absolute and objective, i.e., independent opinions on agents' reputations will tend to diverge throughout a population, defeating the advantages offered by IR. Here we propose an Evolutionary Game Theory computational model with a ternary reputation system, configurable observability of interactions and two consensus promoting social mechanisms: empathy and conformism. We show that changes in the consensus of agents' opinions have direct impacts in cooperation levels and that well mixed populations of empathetic, conformist and egocentric observers can sustain significant levels of cooperation under certain social norms in scenarios of imperfect information.

Introduction
The widespread presence of cooperation in human society is as yet an eyebrow-raising question. If to cooperate is to provide a benefit at personal expense, as is typically defined in its simpler frame, the incentives to do so are still a diffuse and obscure matter researchers endeavour to shed some light into. Whereas widespread cooperation benefits everyone, its associated cost tends to lead individuals into defection. Common theories regarding this incentive to cooperate are often based on kin selection, group selection, or reciprocal altruism (Nowak and Sigmund 2005; Rand and Nowak 2013; Sigmund 2010). Among the varieties of this last one is Indirect Reciprocity (IR), which proposes that cooperation pays because it confers the image of a valuable community member to the cooperator (Nowak and Sigmund 1998), leading this individual to become a target of future cooperation. In other words, in IR a donor provides help if the recipient has a reputation that suggests she will help others too. Thus, the cost of an altruistic act is decreased by an increased chance of later being the target of another altruistic act, netting a positive outcome for the cooperator in the long run, even when the cooperative act is rather costly. Moreover, IR’s importance (and attractiveness) is not limited to its theoretical effectiveness: on one hand, the concepts of reputation and social norms are believed to constitute the biological foundations of morality (Pacheco, Santos, and Chalub 2006), helping to provide answers to social and economic problems like the tragedy of the commons (Milinski, Semmann, and Krambeck 2002). On the other hand, human agents are known to tend towards altruistic behaviour when feeling observed (Haley and Fessler 2005; Okada, Sasaki, and Nakai 2017).

IR has been proven to be an effective and elegant mechanism dedicated to promoting cooperation in a great number of works (Nowak and Sigmund 1998; Pacheco, Santos, and Chalub 2006; Milinski, Semmann, and Krambeck 2002; Nowak and Sigmund 2005; Panchanathan and Boyd 2003; Mohtashemi and Mui 2003; Ohtsuki and Iwasa 2004). Despite IR’s effectiveness, most of these analytic or computational models naturally make simplifying assumptions regarding how information about reputations (a core factor of IR) circulates within populations. One such assumption is that of observability: that reputations are public, and all actions are observed by all other agents in the population. The abandonment of this assumption often hinders cooperation in most Evolutionary Game Theory (EGT) scenarios (Uchida 2010; Radzvilavicius, Stewart, and Plotkin 2019; Mohtashemi and Mui 2003; Hilbe et al. 2018). Indeed, if agents cannot evaluate or trust their assessment of other agents’ reputations, the incentive to cooperate ceases to exist since agents no longer have the prospect of getting recognised as someone who is likely to help. However, different social norms deal with observability (or lack thereof) in different ways. Indeed, its easy to assume that social norms of a more punishing nature tend to struggle at promoting cooperation in low visibility scenarios: if one punishes another based on controversial information, discord quickly takes over the population, thwarting cooperation.

Stern Judgement (SJ), the most effective and resilient
cooperation supporting Social Norm (SN) (Kandori 1992; Pacheco, Santos, and Chalub 2006; Santos, Pacheco, and Santos 2016; Santos, Santos, and Pacheco 2018), has a particularly hard time keeping its cooperation levels high in such scenarios (Uchida 2010). This norm’s success relies on (a) its simplicity: SJ, a rather simple second order SN, leads to more cooperation than more complex norms of higher orders (Santos, Santos, and Pacheco 2018); (b) its straightforwardness: for every possible situation there is always an action that will yield a good or a bad reputation to the donor; and (c) its robustness: SJ is highly resilient at maintaining considerable levels of cooperation even when in scenarios riddled with noise inducing errors, such as reputation assessment, intended action execution or reputation attribution errors (Santos, Santos, and Pacheco 2016; Hilbe et al. 2018). It affirms that bad individuals should always be punished, while at the same time easily forgiven, should they choose to punish an individual with bad reputation as well. Furthermore, SJ’s set of rules seems to closely resemble behaviour such as the one found in pre-verbal infants. In fact, as concluded by experiments with toddlers as young as five months old, these subjects are capable of a considerable degree of abstract and complex reasoning in their first year of life (Hamlin et al. 2011). The authors (2011, p.3) state that:

[...] infants prefer an individual who harms a wrong-doer over one who helps, because they view the wrong-doer as deserving punishment.

However, private reputations may lead to disagreement under SJ: as opinions on individuals diverge within a population, what some see as a sanctioned punishment of bad individuals is regarded by others as unfounded defection. Private information, however, suggests that (i) some individuals may not be known to a large fraction of the population and (ii) evaluations can become subjective and potentially influenced by mechanisms such as empathy, conformism or egocentrism (all defined below). The interplay of those mechanisms in affecting cooperation under different SN remains unclear.

In this work we address the problem of imperfect information with an EGT computational model centred on IR, the donor’s game and variable action observability. With such a model, we intend to answer the following research questions:

1. How should one evaluate actions towards strangers for cooperation to thrive?
2. How do empathy, conformism and ego-centrism influence the efficiency of indirect reciprocity under private reputation systems?

In order to answer the first research question, our model comprehends a ternary reputation system instead of the typically binary system, where we keep the reputations of the latter and add a third possible reputation, the unknown. Strategies in the context of binary reputations are limited to all binary combinations in a two entry vector: AllC [C, C], Disc [C, D], pDisc [D, C] and AllD [D, D], where the first entry prescribes an action towards a good agent and the second towards a bad one. In a ternary system, each of the previous strategies can be extended to a tolerant or intolerant counterpart: for example, the intolerant variation of AllC would be the vector [C, C, D], stating that the agent should always cooperate with good or bad agents, but refuse to do so with a stranger i.e., an individual with an unknown reputation. Additionally, binary norms denote which reputation to attribute to an Donor who helped or refused help towards a good or bad recipient, in that order. Therefore, a norm like SJ [G, B, B, G] gets four variations, regarding how to judge someone who cooperated or not with an unknown agent. Of interest are the combinations that also discriminate based on unknown reputations.

As for the second research question, we propose that an effective way to promote cooperation in this scenario is by introducing consensus homogenising social mechanisms as agent traits, mechanisms such as empathy (Radzvilavicius, Stewart, and Plotkin 2019): the ability to take the observed individual’s perspective forgoing the observer’s own perspective (see Figure 1, top); and conformism: the capacity of figuring out an agent’s public image, i.e. to act or judge on the majority’s perspective (see Figure 1, bottom), in opposition to the default trait of ego-centrism: to only take into account to all binary combinations in a two entry vector: AllC [C, C], Disc [C, D], pDisc [D, C] and AllD [D, D], where the first entry prescribes an action towards a good agent and the second towards a bad one. In a ternary system, each of the previous strategies can be extended to a tolerant or intolerant counterpart: for example, the intolerant variation of AllC would be the vector [C, C, D], stating that the agent should always cooperate with good or bad agents, but refuse to do so with a stranger i.e., an individual with an unknown reputation. Additionally, binary norms denote which reputation to attribute to an Donor who helped or refused help towards a good or bad recipient, in that order. Therefore, a norm like SJ [G, B, B, G] gets four variations, regarding how to judge someone who cooperated or not with an unknown agent. Of interest are the combinations that also discriminate based on unknown reputations.

Figure 1: Illustration of the empathetic (top) and conformist (bottom) judgement. Top: When evaluating Alice’s refusal to help Bob in an empathetic way, Chloe practices a form of “theory of mind”, understanding the intention of Alice under the present SN, and evaluates her action accordingly. Bottom: when evaluating Alice’s refusal to help Bob in a conformist way, Chloe assesses how the majority of Chloe’s peers feel towards Bob, and updating the her own attitude towards Alice accordingly.

As for the second research question, we propose that an effective way to promote cooperation in this scenario is by introducing consensus homogenising social mechanisms as agent traits, mechanisms such as empathy (Radzvilavicius, Stewart, and Plotkin 2019): the ability to take the observed individual’s perspective forgoing the observer’s own perspective (see Figure 1, top); and conformism: the capacity of figuring out an agent’s public image, i.e. to act or judge on the majority’s perspective (see Figure 1, bottom), in opposition to the default trait of ego-centrism: to only take into
account the single observer’s perspective.

**Related Work**

Social norms are paramount pillars of human societies (Young 2015). While definitions abound (Bicchieri 2006), social norms are often seen as prescriptive rules, defining how individuals should act and be regarded according to their actions. In the specific context of indirect reciprocity, social norms are defined as rules that explicitly define how should reputations be attributed in each possible context. Norms in indirect reciprocity can span different complexity levels (Santos, Santos, and Pacheco 2018).

<table>
<thead>
<tr>
<th>Name</th>
<th>$d_{CG}$</th>
<th>$d_{DG}$</th>
<th>$d_{CB}$</th>
<th>$d_{DB}$</th>
<th>Decimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Bad</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Shunning</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Stern Judging</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Image Score</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Simple Standing</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>All Good</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Formal description of two illustrative norms (All Bad and All Good) and the so-called leading four social norms in a binary reputation system, where each entry $d_{ij}$ details the probability that an agent receives a good reputation after practising an action $i$ towards a recipient with a reputation $j$.

In order to account for judgements that possible discriminate based on action (C/D) and recipient’s reputation (G/B), one may use second-order social norms. There are 16 possible norms in this space, however, since G and B are pure labels and possess no a priori meaning, one can take into account symmetry and slice the set into only 10 distinct norms. Out of these, works are typically only concerned with the norms that attribute a reputation of G to agents who cooperate with recipients holding a reputation of G. Thus, only four social norms are identified as points of interest for the study of IR, the norms known as Shunning (SH), SJ or Kandori, Image Scoring (IS) and Simple Standing (SS), present and detailed in Table 1.

The study of IR under a non-binary reputation system, although uncommon, is not a unprecedented modification. Nowak and Sigmund studied the cooperation dilemma in 1998 through IR, with such a model where the image of an agent is a quantitative score which the agent builds through his actions: agents start the simulation with an image score of 0, then an agent that chooses to cooperate increase his score by 1 and an agent that refuses to do so decreases his by 1, with the lower bound of -5 and upper bound of 5 (Nowak and Sigmund 1998). In turn, strategies are simplified to a value $k$, determining that an agent with a strategy $k$ cooperates only, and only if, the image score of the potential recipient is at least $k$. If we reduce this interval until we only have two reputations, 0 and 1 for instance, this model’s reputation update rules correspond to the IS SN: cooperation yields good reputations, defection results in bad reputations, regardless of the recipient’s image.

More recently, systems of ternary reputations have also been studied (Tanabe, Suzuki, and Masuda 2013). In this work, Tanabe et al. present the reader with three possible reputations: good, neutral and bad. This leads to the existence of 2^3 action rules (or strategies). While strategies like AIJC and AILD try to implement actions C and D, respectively, regardless of the recipient’s reputation, gDisc (generous discriminator) only gives help to recipients with good or neutral reputations. At the same time, rDisc (rigorous discriminator) refuses help to either neutral and bad players. Akin to Nowak and Sigmund’s 1998 work and contrary to most works on IR, reputations in this model contain an inherent order. Within the former, labels like G or B are void of any a priori concrete significance, while in contrast, within this current work, social norms prescribe a positive or negative reputation shift, instead of directly prescribing a new reputation. These increases and decreases of reputation give this reputation system an inherent order: $B < N < G$. Tanabe et al. conclude that, for small errors, the maximum level of cooperation happens when a population where the gDisc (generous discriminator) strategy is prevalent, under a 3-bit variation of SJ where neutral agents are considered good, i.e., where defection towards neutral players is condemned.

Nowak and Sigmund’s work not only focuses on non-binary reputations, but also on incomplete information, arguing that IR based cooperation relies on the “ability of a player to estimate the image score of the opponent”, (1998, p. 575) (Nowak and Sigmund 1998). Then the authors proceed to study a scenario considered to be more realistic: where interactions between two individuals are only observed by a small subset of the population. Introducing the Image Matrix (IM), containing the image score each agent believes every other agent has, the authors account for the possibility of agents having different perceptions on the image score of other agents, the concept we present as reputation asymmetry. The authors conclude that in order to have cooperative strategies as evolutionarily stable strategies, the probability of knowing the image of the recipient must exceed the cost-to-benefit ratio of the altruistic act.

The challenge of incomplete information has been further studied (Uchida 2010; Hilbe et al. 2018) and its challenge stands: noisy, asymmetric, incomplete or private information hinders cooperation since it leads individuals to disagree on each others’ reputations, disagreements which proliferate over time, albeit differently under each norm. S. Uchida also explores how different social norms sustain cooperation in such scenario, specifically for three of the leading-four social norms: IS, SS and SJ (Uchida 2010). Again, the authors consider a parameter $q$ controlling the percentage of the population that is selected to observe each interaction and apply the concept of the image matrix. The authors conclude that there are significant differences in how SS and SJ struggle to foster cooperation in the scenario of private information: while SS’s level of sustained cooperation gradually decreases with the value of $q$, being able to promote a significant amount of cooperation for medium levels of observability ($q = 0.5$), SJ fails to hold its seat as the highest level of cooperation sustaining norm - in fact, the slightest decrease in observability (of 0.01, for instance) brings the
image matrix into a chaotic state. The author argues that the difference between these two assessment rules is that while SS allows for a mismatch in the opinion between two observers to be repaired, the stern rule SJ does not: if two observers using SS have different opinions towards a donor, this difference is removed whenever such donor chooses to cooperate – cooperation always leads to good reputation under SS. Yet, under SJ, no action can repair this mismatch of opinions, leading disagreements to proliferate over time. Therefore, this reasoning can be interpreted in the following manner: social norms that hold the highest level of information consensus in the face of private information are the ones whose sustained cooperation suffer the least from this adversity. The opposite is also true: social norms who fail to hold significant levels of consensus are the ones whose cooperation suffers the most.

A possible corollary of this theory is that if there are mechanisms dedicated to promote consensus, cooperation can be sustained in a scenario of private information. We assume the problem detailed within the following premises: (i) IR is regarded as one of the most elegant and simple solutions to the cooperation dilemma; (ii) information about other individuals’ reputations isn’t objective and publicly available in real life; (iii) IR’s impact on cooperation declines significantly in the face of the second premise; (iv) cooperation in society is ubiquitous.

Looking for inspiration in the realm of the social intelligence of humanity, one can identify social mechanisms such as gossip, empathy and conformism as mechanisms dedicated to overlook personal disagreements in favour of collective harmony (Dunbar 1996; Batson and Moran 1999). Thus, by emulating such behaviours into a computational model, one can assume that cooperation can be boosted through the homogenisation of consensus through such mechanisms. The role of empathy in the realm of EGT has been recently studied, and its effects are bright: on one hand, empathy boosts cooperation in populations devoid of public and complete information; on the other hand, empathy can be evolutionarily selected into prominence if allowed to be learned or imitated by agents through social learning (Radzvilavičius, Stewart, and Plotkin 2019). Also, as stated by Dunbar within the previous quote, gossip has been mathematically modelled and established as a powerful mechanism for detaining free riders in a population of cooperators (Enquist and Leimar 1993). Sommerfeld et al. have experimentally studied gossip as a means of transmission of possibly imperfect social information (Sommerfeld et al. 2007) and concluded that gossiped information has a strong influence on participants’ decisions, even when these directly observe the relevant interaction. As for conformism, the mechanism has been studied within the context of cultural transmission (Peña 2008; Efferson et al. 2008) and of the study of EGT through social networks (Peña et al. 2009).

For the purpose of studying the impact of conformity in this work (as a reputation homogenising mechanism), we propose its division into two different approaches: (1) observation based conformism and (2) action based conformism. The former acts within observers and their observing and reputation-updating process, whereas the latter is applied into the donor’s decision making process. This work focuses especially on the former: how conformism affects the judgement of an action by third parties. This way, we propose to simplify and incorporate gossip into the conformity mechanism: we assume that when a conformist agent judges a donor accordingly, the observer does so on information on the donor we assume was promptly shared by the observer’s peers by a form of gossip. This simplification relies on the implication that conformism depends on gossip: an individual’s decision on how to act towards another individual or judge one always takes into account information that was previously shared, likely through gossip (Sommerfeld et al. 2007). On the other hand, conformism can be seen as a form of group empathy, as in an individual’s ability to empathise and quickly assess not a single individual’s opinion, but a group’s opinion as a whole on any given matter. In fact, as was previously mentioned, humans are known to act altruistic when under acknowledged observation (Haley and Fessler 2005; Okada, Sasaki, and Nakai 2017). Thus, when under observation by a large group, conformism can be seen as a propensity to second-guess our first reaction (to act egocentrically) and adopt the most sanctioned behaviour in the group (choose the less polarising action i.e., act like the majority of the population would act or judge like most agents judge).

Lastly, some works identify partner-selection as one of the main purposes of IR: agents strive to keep good reputations in order to avoid being shunned by other agents and stay as potential targets of further cooperation. This can be modelled as a higher probability of choosing good individuals over others as recipients of future interactions. We study the impacts of partner-selection for the following reason: while partner-selection is known to boost cooperation by itself, this mechanism has no direct effects on information consensus. Thus, while we expect to see cooperation boosted by partner-selection, the lack of a reputation homogenising component makes it a poor candidate to solve IR’s observability problem.

Model

Background

Our model considers a finite population of \( Z = 50 \) agents, each holding (a) a strategy, corresponding to a programmed behaviour towards other types of agents, prescribing which action (of Cooperation or Defection) to execute towards each reputation (we only consider pure strategies - probabilities are either 0 or 1); (b) a value of fitness, a quantitative unit representative of the individual’s relative ability to reproduce against other individuals’; and (c) a personal bias corresponding to the way the agent judges others’ actions: the agent is either egocentric, empathetic or conformist. The distribution of the frequency of each judgement type in a population is controlled by parameters \( \delta, \epsilon \) and \( \sigma \), respectively. While most models rely on a vector of binary reputations, this model stands on a matrix of ternary reputations, the Image Matrix, \( \beta \), storing the way each agent perceives every other agent, with the Good \( G \), Bad \( B \) or Unknown \( U \) reputation. Let \( D \equiv \) Donor, \( R \equiv \) Recipient,
With the complementary probability \( \explore \), an agent \( X \) is picked to explore a new strategy, which happens with probability \( \mu \). With the complementary probability \( 1 - \mu \), the agent revises his strategy through social learning, a process that requires the following events: agent \( X \) and another randomly selected agent \( Y \) play \( 2Z \) Donor's Games with randomly selected opponents. This selection, however, can be influenced by partner-selection: a parameter to which we attribute the selected opponents. This selection, however, can be influenced by partner-selection: a parameter to which we attribute the Greek letter \( \psi \) to control the added probability of agent \( X \) selecting an agent who \( X \) sees as a good agent to play the Donation Game with. To illustrate, \( \psi = 0 \) corresponds to a completely random partner-selection, while \( \psi = 1 \) sees agents with a good image being selected twice as much as other agents. For every game, there's equal probability of assuming the roles of Recipient or Donor. Next, the Donor agent accesses the Image Matrix's entry corresponding to the Donor's opinion of the Recipient, \( \beta_{D,R} \), on which the Donor's strategy acts; an act of cooperation yields the Recipient's fitness with a benefit \( b \) and costs the Donor's a value of \( c \); both agents proceed to compare their respective average payoff obtained during the \( 2Z \) donation games. A complete simulation occurs over \( G \) time steps, or generations.

Errors

Some of the events of the donation game are subject to noise inducing errors: the donor's act of assessing the recipient's reputation in order to apply its strategy results in the opposite effect. An agent \( X \) acts with probability \( \epsilon \), called assessment error. Additionally, while the agent's strategy dictates its action towards each possible reputation, if this action is that of cooperation, the agent defects with probability \( \varepsilon \), called the execution error. This error resembles the inability of an agent to execute the action prescribed by its strategy and it is usually implemented only as failed intended cooperation due to lack of resources or opportunity to donate (Santos, Santos, and Pacheco 2016). Thus, making an intention to defect is always carried out.

The effects of each type of error on each norm and their resilience to such have already been studied (Santos, Pacheco, and Santos 2016). During our simulations, we will fix the values of the errors at \( \varepsilon = \chi = 0.01 \) as a small noise inducing factor. We consider that, over time, agents change their strategies, either by evaluating their relative success and imitating the most successful, or by random mutations. These two processes, further explained, correspond to the processes of social learning, through which successful strategies are propagated, and of strategy exploration, through which a population lowers the probability of settling in a strategy that provides a local maximum of payoffs: combining the effects of both processes, if an agent randomly mutates into a more successful strategy, it may quickly invade the population through social learning.

Dynamics of the reputations on the population rely on each time step: either an agent explores a new strategy (with probability \( \mu \)), or revises it through social learning (with probability \( 1 - \mu \)), leading to the events that also trigger reputation dynamics. The effects of different values of \( \mu \) on each specific SN have already been studied (Santos, Pacheco, and Santos 2016), leading us to fix a value of \( \mu = 10^{-2}/Z = 0.002 \). This means that within any time step, with a probability of \( 0.2\% \), the selected agent forgoes his or her strategy for a random new one from the whole spectrum of possible strategies, regardless of his or her current success. On the other hand, with a probability of \( 99.8\% \), the selected agent \( X \) plays the donation game. Then it proceeds to compare its fitness \( f_X \) (the average payoff obtained during the \( 2Z \) donation games) with the second agent’s (\( Y \)) fitness \( f_Y \). The higher the fitness difference, the higher the probability of \( X \) changing its strategy to that of \( Y \). This probability is given by the pairwise comparison rule (Equation 1).

\[
P_{X \rightarrow Y} = (1 + e^{-\beta(f_Y - f_X)})^{-1}
\]

Dynamics of Opinions/Reputations

The IM, containing the opinions each agent has on every other agent, changes constantly during each generation. Each donation game, a number of agents \( q Z \), is selected to observe. Each observer \( O \) processes its own type of judgement, under the same social norm \( d \), according to its own observer-trait, in the following manner:

- an Egocentric agent evaluates the Donor’s action from the observer’s point of view: \( \beta_{O,D} = d(A, \beta_{O,D}) \)
- an Empathetic agent evaluates the Donor’s action as seen from the Donor: \( \beta_{O,D} = d(A, \beta_{D,R}) \)
- a Conformist agent sacrifices its own opinion for the most shared opinion of the Donor: \( \gamma = \text{public reputation of } D, \beta_{O,D} = d(A, \gamma) \)

The simulations begin with a completely consensual IM, i.e., each of its columns bears the same reputation value for each of its rows. Thus, opinions on each agent are consistent throughout the population. This remains true for simulations where \( q = 1 \), meaning that every action is observed by the entirety of the population. Whenever two agents play the Donor’s Game, all other agents are selected to observe and, regardless of the observers’ judgement types, since \( \beta_{O,R} = \beta_{D,R}, \forall O \in Z \), either all opinions towards the donor change or remain unchanged.

On the other hand, if \( q \neq 1 \), opinions naturally start to diverge: observers of an occurring game may change their views on the acting Donor while the remaining non-observant population has no opportunity to revise theirs. As a result, consensus in the population decreases, and the same
is expected to happen to cooperation. This decrease, however, happens differently for each SN. In the mathematical context of this model, a norm is defined by a vector \( d \) (see Table 1), with the appendage of two entries, \( d_{CU} \), \( d_{DU} \), denoting the probability that the donor’s reputation becomes \( G \) after acting towards a recipient with the unknown reputation.

**Methods**

In order to evaluate the results of our model in relation to the ones of previous works, we focus on two main metrics: the average consensus rate \( \eta \), and the average consensus of the image matrix, \( AC \). Each simulation is repeated \( R = 50 \) times (runs), accounting for the stochastic nature of the process, guaranteed that the pseudo-random number generator has a different seed each run. Thus, both the aforementioned metrics are averaged over the total number of runs.

Since \( q = 0 \) stands for no action observation at all, such a scenario sees no reputation dynamics and thus IR ceases to effectively exist. However, as we intend to study the reputation dynamics of extreme scenarios such as those of minimum observability, the results we show for \( q = 0 \) are in fact for an arbitrarily low value of observability (\( q = 0.01 \)).

To calculate each run’s results, we first allow for a concentration period of 10% of the total number of generations, after which the simulation is expected to become stable. Only after these \( 0.1 \ast G \) generations actions start to get registered in order to calculate metrics such as the following ones.

**Average Cooperation Rate**

The previously detailed model is simulated for \( G = 10^6 \) generations. Each simulation runs for 10% of its total amount of generations, starting with randomly attributed strategies and reputations, allowing for the simulation to enter a stationary situation. Equation 2 accounts only for the subsequent 90% of generations, during which every cooperative act is counted against the total number of interactions.

The average cooperation rate \( \eta_i \) in run \( i \) is calculated by the ratio between the total number of cooperative acts \( C_i \) and the total number of games \( K_i \):

$$\eta = R^{-1} \sum_{i=1}^{R} C_i / K_i = R^{-1} \sum_{i=1}^{R} \mu_i \quad (2)$$

For each tested combination of parameters within this work, we consider the average cooperation rate \( \eta \) of all runs of the simulator.

**Average Consensus**

In order to study how opinion divergence impacts cooperation, we use a metric that evaluates how much opinion divergence there is in the image matrix. Let \( G_i \) and \( B_i \) be the number of agents that regard agent \( i \) as good and bad:

$$AC = Z^{-2} \sum_{i=1}^{Z} |G_i - B_i| \quad (3)$$

The average consensus on the image matrix is computed by calculating the consensus rate on each individual image (consensus = \( |G - N| \) for each agent) and then calculating its average over the population (Equation 3). The consensus value of an agent \( X \) is maximum (100%) when all agents agree on \( X \)’s reputation, and minimum (0%) when it is the most divergent, i.e., when half the agents in the population see \( X \) as a good individual and the other half considers \( X \) a bad one. For the purpose of this metric, unknown images are not considered, for its inherent semantic value: if an agent \( Y \) considers agent \( X \) to be unknown, agent \( Y \)’s opinion doesn’t go against the opinion of and agent \( W \) who sees \( X \) as good, for instance, since an unknown image is considered a lack of opinion, instead of a disagreement, while the average consensus rate only evaluates the level of disagreement in a population.

**Results**

Using the previously described model, we set to answer our research questions. In Figure 2 we show the cooperation levels sustained by the ternary variations of the leading four social norms in a scenario of imperfect information, where all agents judge interactions egocentrically, as in most previous related works. While the norm SJ (like SH) seems unaffected by the way unknown individuals are treated, the norm SS clearly benefits from regarding unknown individuals as good ones (the stranger-tolerant variation), in opposition to the second ternary variation which punishes individuals for cooperating with unknown ones (the stranger-intolerant variation). This positive difference between the cooperation sustained by the tolerant and intolerant variations is also observable for the the norm IS, despite being as marginal as the negative difference seen for SH.

![Figure 2: Cooperation sustained under tolerant and intolerant ternary variations of the leading four social norms, for q = 0.6 in a fully egocentric population (δ = 1, σ_{obs} = ε = 0). For each presented norm, the leftmost bars correspond to the amount of cooperation sustained by the tolerant variations of each norm (regard unknown individuals as good) while the rightmost ones the intolerant variation. The single letter within each bar denotes the most popular action towards strangers. Other parameters: Z = 50, G = 10^5, μ = 0.02, ε = λ = 0.01, φ = λ = σ_{act} = 0.](image-url)
the simulation reaches a stationary state and actions start being registered. In other words, while unknown agents can have their images updated into good or bad reputations, no state can lead back to an unknown reputation. Thus, the only explanation towards such a stark difference in the cooperation promoted by SS’s variations relies on the assessment error $\chi$, which was set to 0.01 in our simulations. Hence, after reaching a stationary state (when no agent regards another with an unknown reputation), the donor incurs in a probability of $\chi \ast \frac{1}{2}$ of assuming the recipient is unknown to him, acting and being judged accordingly. Assuming such an event happens once every 200 games and knowing that a complete simulation occurs over $2 \ast Z \ast G = 2 \ast 50 \ast 10^5 =$ ten million games, an action towards a stranger is still executed and judged roughly fifty thousand times per simulation. Regarding such cases, under the tolerant variation of the norm SS $d=(1,0,1,1,0)$, acting towards strangers like one would act towards good individuals clearly boosts overall cooperation over its second variation $d=(1,0,1,1,0,1)$ – regarding strangers as bad individuals. We also studied the frequencies each strategies had in the population by the end of each simulation for each norm. Figure 3 shows how each variation of a social norm advocates strategies in the population: the intolerant variation of the norm SH, the only one showing a marginal benefit from not regarding strangers as good individuals, is also the only norm whose frequency difference between its tolerant and intolerant variations is non-marginally negative, i.e., where the intolerant behaviour is the most popular. Under any other norm, whichever the variation, regarding strangers as good individuals is the most evolutionarily popular strategy. This difference is the most stark within the variations of the AllC strategy in populations ruled by the tolerant variations of both SS and SH.

In order to study the impacts of empathy and conformity on different levels of observability, we first studied cooperation and consensus on homogeneous populations, i.e., populations comprised of only one type of judging traits. In Figures 4 we show that both empathetic and conformist populations sustain a larger percentage of information consensus than egocentric populations, an insight also observable for cooperation under SJ. In fact, empathetic populations sustain remarkable amounts of consensus even in extreme scenarios of observability ($q = 0.01$), which is clearly translated into remarkable levels of cooperation. Additionally, we find that, for both norms, the amplitude of consensus variation for each interval of $q$ closely resembles the amplitude of cooperation variation within the same intervals, leading us to hypothesise there is a mathematical correlation between consensus decay and cooperation decay.

Despite suggesting interesting insights into the effects of each observation type, this scenario of homogeneous populations is not a realistic one, for there are no populations consisting on only one type of observer. Thus, we set to study how populations consisting on different combinations of observation profiles fare in sustaining cooperation under each norm, i.e., when populations detain members with different observation traits. In order to do so, we fixed two values of observability - $q = 0.25$ for a scenario of low observability and $q = 0.75$ for a scenario of higher observability - and performed the simulation for all possible distributions of observation traits, with a step of 0.1. This results in
a ternary plot, where each section of the triangle is colour coded for the amount of cooperation promoted by the corresponding combination of observation trait values, a combination denoted by a vector \( c = (\delta, e, \sigma) \), where the \( \delta \) stands for the ratio of egocentric observers, \( e \) for the ratio of empathetic observers and \( \sigma \) for the ratio of conformist observers. Hence, while each vertex of the triangle pertains a combination corresponding to each different homogeneous population, each side corresponds to populations devoid of the opposing-vertex type of observers, with complementary frequencies of the two adjacent vertices.

In Figure 5 we show the results for differently mixed populations, for both observability scenarios, under the second variation of SJ. The most immediate insights these plots suggest is of the prevalence of cooperation near the empathy vertex, which under a higher visibility extends to the conformist one. Indeed, this latter observer type vertex clearly falls shorter on the promotion of cooperation when there is less visibility. Thus, observing the areas near each vertex, we conclude that highly empathetic populations manage to maintain their high levels of cooperation under loss of observability, whereas populations dominated by conformist observers lack such resilience. However, the scenario of high - albeit still imperfect - observability shows that certain populations where egocentric observers dominate can still thrive with significant levels of cooperation (assuming a threshold of \( q = 50\% \)), with populations such as the ones constituted by the distributions \((0.5, 0.5, 0), (0.5, 0.4, 0.1)\) and \((0.5, 0.3, 0.2)\).

We presented, however, two possible paths to study conformism: observer-level conformism, where observers judge interactions through the participants’ public image, and action-level conformism, where agents act on such public image. While both types of conformism were expected to boost the level on consensus in the image matrix, the degree to which each type of conformism affects cooperation through consensus under imperfect information differs greatly, which can be seen in Figures 6(a) and 6(b). These plots show that while both types have very similar impacts on cooperation, consensus is much more dependant on observation-level conformism than its action-level counterpart. This suggest that while observation-level conformism is much more capable of promoting cooperation through information consensus, action-level conformism promotes cooperation while having marginal impacts on consensus. On the other hand, the combination of both types of conformism is clearly more effective in the promotion of both cooperation and consensus.

Lastly, an additional mechanism was introduced into the model in order to ascertain to importance of partner-selection (Santos, Pacheco, and Lenaerts 2006) in scenarios of imperfect information. We compared the level of cooperation sustained under all four norms by both \( \psi = 0 \) and \( \psi = 1 \) (where \( \psi \) represents the added probability of an agent selecting a Good partner), in scenarios of imperfect information with both low and high levels of consensus. As illustrated in Figure 7, when the population is fully egocentric, partner-selection has marginal effects on all norms, with the exception of IS. This exemption can be explained by the fact that this norm is no more than a first order social norm: cooperation leads to good reputations, defection leads to bad ones, regardless of the target. However, when the population is capable of sustaining higher levels of consensus, through the addition of some empathetic and some conformist agents, partner-selection shows non-marginal benefits, with the exception of the norm SH.

### Conclusion

In this work we set to answer questions regarding the incompatibility of IR as a cooperation promoting mechanism and the reality imperfect of information. We showed how recent works regarding the latter challenge the accomplished results suggested by the study of the former. Additionally, both our results and works on the impacts of imperfect information suggest that cooperation is deeply dependant on information consensus: the more a group disagrees on someone’s reputation, the more cooperation is hindered. Hence, we posed the following question: If cooperation is deeply linked with consensus (or the lack of it), can cooperation be fostered in a scenario of imperfect information through
additional social mechanisms capable of effectively disseminating opinions?

In order to do so, we first include a third reputation: individuals can be regarded as Good, Bad, or Unknown to others. Since the scope of this work is centred on small populations (Z < 100) and social norms never prescribe an unknown reputation, interactions with unknown individuals quickly become rare, albeit not necessarily insignificant. In such a scenario, where interactions with strangers happen occasionally, we show how different social norms show different behaviours when promoting cooperation: regarding strangers as individuals with good reputations is mostly popular but has marginal impacts on cooperation at best (SH, IS and SJ); however, when such a behaviour promotes cooperation, it makes a remarkable difference, as the one seen under SS.

Subsequently, we studied the impacts of adding two meta-mechanisms to the mechanism of IR. The first one was empathy, whose modelling in EGT’s framework was first devised by Radzvilavicius et al. (Radzvilavicius, Stewart, and Plotkin 2019): the ability to let go of our personal opinion on Bob and observe him through the eyes of Alice, allowing our opinion on Alice to change according to her judgement of the situation instead of ours. We confirm the author’s conclusions on the impacts of empathy: it is a remarkably powerful tool in the promotion of cooperation in scenarios of imperfect information: even a non dominant distribution of empathetic agents (such as 40%) is enough to turn SJ’s observability woes into significant amounts of cooperation (over 50%). In a scenario of a fully empathetic population, cooperation can even be sustained above 90% for any level of observability, which shows how powerful empathy can be as a consensus homogenisation tool. In fact, information consensus seems to decay ever so slightly with observability, when supported by empathetic agents. Additionally, when comparing how consensus is fostered by each population type (egocentric, empathetic and conformist), there is a clear ranking on their effectiveness, with straightforward translation on cooperation: empathy provides a higher degree of consensus than conformism, which in turn sustains a higher degree than our egocentric baseline. On the mechanism of conformism, we suggested two ways to study it: conformism at observer level and at action level. At the former, when agents evaluate the donor’s action towards a recipient, they take the public image of the recipient into consideration, while at the latter, donor’s take this same image into consideration when choosing how to act. We find that while observer-level conformism also has a mostly positive impact on both cooperation and consensus, its promoted cooperation shows a starker decline with observability than the one provided by the mechanism of empathy. On the other hand, when accounting for action-level conformism, cooperation can be boosted without the promotion of consensus: in fact, if an acting donor can effectively figure out how the majority of the observers regard the current recipient, strategies sanctioned by the underlying social norm quickly become more popular. We also show that the combination of both levels of conformism is extremely effective at the promotion of cooperation.

While our model shows that both empathy and conformism are powerful “meta-mechanisms” capable of promoting cooperation where IR seems to fail, such as scenarios of imperfect information, it still leaves some ques-
tions to be asked. For instance, while Radzvilavicius et al. allow for empathy to evolve as trait through social learning (Radzvilavicius, Stewart, and Plotkin 2019), the same could be applied to conformism. A number of added complexities can be added to this computational framework in order to study the complexities of human social interaction. For instance, humans can be capable of high degrees of empathy but choose not to act on it in order to achieve a higher social status by empathising with the observers and figuring which action leads to better reputations and consequent future cooperation. In other words, empathy itself can be used to exploit a system of IR. Lastly, another example of added complexity is that of costly reputation building, which has been studied by Santos et al. (Santos, Pacheco, and Santos 2018), and that of opinion-inertia: the reporting of actions is often costly, and humans often stick to their first impression.

Acknowledgments

This research was supported by FCT-Portugal grant PTDC/MAT-STA/3358/2014.

References


