ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230361

Generosity and the Emergence of Forgiveness in the Donation Game

Nathan Griffiths[a;*] **and Nir Oren**[b;**]

[a]University of Warwick, Coventry, CV4 7AL, UK
[b]University of Aberdeen, AB24 3UE, Scotland
ORCiD ID: Nathan Griffiths https://orcid.org/0000-0002-6406-8632,
Nir Oren https://orcid.org/0000-0002-4854-9014

Abstract. Research has shown that cooperative action struggles to emerge in the noisy variant of the donation game, a simple model of noisy multi-agent systems where indirect reciprocity is required to maximise utility. Such noise can arise when agents may have an incorrect view of the reputation of their interaction partners, or when the actions themselves may fail. Concepts such as generosity, as well as the use of higher-order norms, have been investigated as mechanisms to facilitate cooperation in such environments, but often are not effective or require additional assumptions or infrastructure in the system to operate. In this paper, we demonstrate both analytically and empirically that a simple form of generosity when combined with fine grained reputation can help cooperation emerge. We also show that the use of individual forgiveness strategies rather than the presence of global generosity can support cooperation in such environments.

1 Introduction

Reputation, in the form of *indirect reciprocity*, is an effective mechanism for supporting cooperation in human societies, and acts as an incentive for individuals to engage in potentially costly cooperative actions towards others [5, 11, 24, 26]. In order to analyse the circumstances in which indirect reciprocity is effective, it is often studied in abstracted settings such as the *donation game* [12, 13, 16]. In this setting, pairs of individuals are selected, one as a potential donor and the other as a recipient [12]. The donor may choose to cooperate (by *donating*), based on their own strategy and the recipient's reputation, in which case the donor incurs a cost, the recipient receives a benefit, and the donor's reputation is increased. Agents are selected randomly to interact, and at the end of each generation offspring are produced using selection and/or mutation. Agents with higher fitness, as determined by the benefits they received less any costs of donating, are more likely to be chosen to produce the next generation.

An oft-investigated strategy for deciding whether or not to donate within the game is based on a simple reputation measure referred to as *image scoring*. Here, donation increases an agent's image and defection (non-donation) decreases it, and agents choose whether to donate or not based on the recipient's image score. Nowak and Sigmund demonstrated that in such settings the population goes through cycles of establishing cooperation, only to be undermined by defectors [12]. The basic donation game assumes that actions always succeed and that observations of interactions (to update image scores) are perfect. In this paper, we consider a noisy setting in which actions may fail and where observations are imperfect, i.e., we consider *action noise* and *perception noise*.

Image scoring has been shown to be unstable, and higher order norms governing donation (known as the *leading eight*) have been proposed in the literature [14] and shown to support cooperation. However, such work makes multiple limiting assumptions; if reputation is not publicly known, or noise is present in the system, such higher order norms are no longer effective [23].

Another potential approach to mitigate the impact of unreliable or non-public image scores, as suggested by Schmid *et al.*, is the use of generosity, which either judges an agent as having a higher image score than they do, or allows a donor to donate when it normally would not. While an empirical evaluation of generosity has shown it to be of only limited efficacy [23], this evaluation also assumed binary reputation. We hypothesise that this has insufficient nuance to enable image scoring to be effective. Generosity is also an 'aligned' mechanism, in the sense that there is a fixed probability of being generous throughout the lifetime of a system and no other information, such as the image score of a potential recipient, is considered.

In this paper, we (1) consider the impact of nuanced reputation on generosity and (2) propose two forms of forgiveness, namely *action* and *assessment* forgiveness. These are 'non-aligned' mechanisms in which individuals have their own strategies and serve as 'higher order' norms which consider others' image scores. In evaluating generosity and forgiveness we relax some of the assumptions made in previous work. Specifically, we include both action noise and perception noise, meaning that an attempted cooperative act might fail and observations are imperfect. We also assume that reputation is nuanced, as in the work of Nowak and Sigmund [12] rather than being binary in the case of generosity and the leading eight [14, 21, 23][1].

The contributions of this paper are as follows. We (i) investigate the impact of action and perception noise on image scoring, (ii) explore the impact of nuanced reputation on action and assessment generosity, (iii) present a theoretical analysis of the impact of generosity on cooperative behaviour, and (iv) propose and evaluate action and assessment forgiveness in a 'non-aligned' setting where forgiveness

* Corresponding Author. Email: Nathan.Griffiths@warwick.ac.uk.
** Corresponding Author. Email: n.oren@abdn.ac.uk.

[1] The code used to generate the results presented in this paper is available at: https://github.com/nathangriffiths/ECAI-Forgiveness.

strategies evolve alongside donation strategies.

The remainder of this paper is structured as follows. The next section situates our work within existing research. Section 3 details our experimental setting, while Section 4 provides an empirical evaluation of generosity in the presence of nuanced reputation scores. Section 5 provides an analytical analysis of generosity in this context. In Section 6 we introduce the notion of forgiveness in the context of the donation game and evaluate its effectiveness. Section 7 discusses our results, considers future work, and concludes.

2 Related Work

Indirect reciprocity through reputation has been shown to be a potential mechanism to support cooperation in populations of self-interested individuals, without requiring repeated encounters between individuals. Nowak and Sigmund describe how a simple *image scoring* reputation measure is able to support cooperation in a (noise free) *donation game* in which random pairs of agents are selected from a population $\mathcal{A}$, one as a potential donor and the other as a recipient [12]. The donor may choose, at cost c, to cooperate with the recipient, in which case the recipient receives a benefit b (such that $b > c > 0$). A donor's strategy is a threshold on the recipient's image score, such that a donor i with strategy k_i, will cooperate with a recipient j if $k_i \leq s_{ji}$, where s_{ji} is i's perceived image score of j. If a donor is observed to cooperate, the observer's image score of the donor is incremented, while it is decremented if they do not cooperate. The image score of a recipient is unchanged by an interaction.

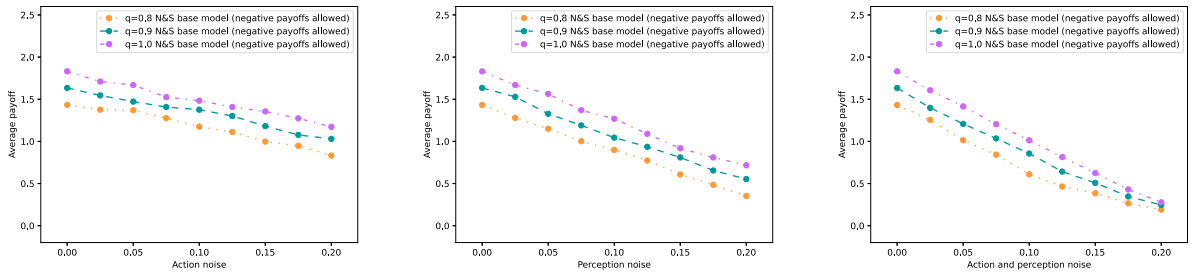
Nowak and Sigmund experiment with a population of $|\mathcal{A}| = n$ agents, such that for any agents $i, j \in \mathcal{A}$, $s_{ji} \in [-5, +5]$ and $k_i \in [-5, +6]$, with k_i initialised at random and s_{ji} initialised to 0. Each generation m donor-recipient pairs are randomly chosen, meaning that an agent will be involved in $2m/n$ interactions on average per generation. At the end of each generation agents produce offspring proportionally to their fitness, as determined by the benefits received less any costs of donating, with a small probability p_m of mutation causing an offspring to use a random strategy. In this setting, the population goes through cycles in which cooperative behaviour is established, only to be undermined by defectors, before becoming the dominant strategy again [12].

The effectiveness of image scoring is dependant on the ability of a donor to estimate a recipient's image score. After an interaction, other agents in the system observe the interaction with probability q and update their own perception of the donor's image score. Each such observer o keeps track of the image score s_{io} of a donor i, meaning that in the case of partial observation (i.e., $q < 1$) different observers may associate different image scores with a given donor. In Nowak and Sigmund's experiments, cooperation emerges in this setting provided that the probability of the donor knowing the image score of the recipient exceeds the cost-to-benefit ratio of the cooperative act [12]. However, Nowak and Sigmund's formulation assumes there is no noise, meaning that observers always perceive interactions perfectly, implying that image scores are accurate (if potentially incomplete), and that donation actions always succeed. In this paper, we relax the assumption of a noise-free setting, and consider two types of noise, namely perception noise and action noise.

The updating of reputation (i.e., an image score) by an observer, and its use in subsequently determining when the observer should cooperate can be viewed as a *social norm*. Such a social norm has two components, namely an *assessment rule* that defines how a donor's reputation is affected by its actions, and an *action rule* that specifies the circumstances under which a donor should cooperate with a given recipient [23]. The assessment rule for image scoring is that a donor's reputation is incremented for a cooperative action and decremented for defection (bounded to be in $[-5, +5]$). The action rule is that a donor i should cooperate with recipients who have sufficient reputation and defect otherwise, i.e., they should cooperate with recipient j if $k_i \leq s_{ji}$. In the image scoring social norm, reputation assessment depends only on the donor's actions, and so it is considered to be a 'first-order' norm. Image scoring has been shown to be unstable [12, 14, 23], as a result of the norm requiring individuals to defect against those with a reputation lower than their strategy, resulting in their own reputation being damaged. A potential solution to this lack of stability is for social norms to distinguish between justified and unjustified acts of defection [3, 18, 23].

Ohtsuki and Iwasa explore 'second order' and 'third order' norms, which consider the reputation of the donor and recipient, in addition to the donor's action [14]. Using a simplified setting where reputation is binary and publicly known (with all agents associating the same score with an individual) and there is no noise, Ohtsuki and Iwasa identified eight norms (the 'leading eight') which are able support cooperation. For interactions with a recipient of good reputation, each of these norms gives good reputation for cooperation and bad reputation for defection. However, they differ in how they assess interactions with bad recipients, for example norms such as 'stern judging' penalise cooperative behaviour towards a bad individual [17]. The leading eight show that consideration of second or third order norms is a potential solution to the instability of image scoring [4, 15, 17, 19, 20]. However, the assumptions of binary public reputation and a lack of noise are unrealistic, and it has been shown that if reputation is not public, i.e., in partially observable settings (where $q < 1$) the leading eight are no longer effective [7, 23]. In such circumstances, therefore, or in the presence of perception noise, alternative mechanisms are needed [9, 23].

Schmid *et al.* propose two types of generosity, *assessment generosity* and *action generosity* [23], as potential methods to mitigate the impact of unreliable or non-public image scores. Assessment generosity is where an agent sometimes assigns a good reputation to individuals who would normally be regarded as bad. Action generosity is where an agent sometimes cooperates with an individual with whom they would usually defect. Similarly to Ohtsuki and Iwasa, Schmid *et al.* use a simplified version of the donation game, in which reputation is binary (i.e., 'good' or 'bad'). However, unlike Ohtsuki and Iwasa, they consider partial observations ($q < 1$) which are subject to perception noise, such that an observation may be perceived incorrectly with some probability e_p (meaning that cooperation might be perceived as defection and vice versa). Through empirical experiments, Schmid *et al.* show that assessment generosity is not effective and reduces cooperation (by increasing noise and allowing defectors to proliferate) but that small amounts of action generosity can be helpful. However, it is important to note that these experiments assumed binary reputation and that the probability of generosity is uniform across all agents and interactions, i.e., it is an 'aligned' setting where agents are not more or less generous towards others with good or bad reputations. In this paper, we (i) consider the impact of nuanced reputation ($s_{ji} \in [-5, +5]$) on generosity in the presence of perception and action noise, and (ii) analyse a 'non-aligned' analogue of generosity, in the form of *forgiveness*, in which individuals have their own forgiveness strategies which evolve and consider the reputation of the recipient when determining whether to cooperate.



(a) Action noise (b) Perception noise (c) Action and perception noise

Figure 1. The impact of (a) action, (b) perception and (c) action and perception noise on Nowak and Sigmund's base model (using their setting of $n = 100$, $m = 300$ and $p_m = 0.001$ [12]), with negative payoffs allowed, i.e., without the additional payoff of 0.1 for each interaction. Results are averaged over 100 runs of $g = 10^5$ generations, and show the average payoff for different probabilities, q, of onlookers observing interactions.

3 Action and perception noise in the Donation Game

The donation game presented by Nowak and Sigmund is a noise-free environment, in which agents perceive other's actions perfectly and actions are always successful (i.e., a donation always succeeds). Since many real-world environments contain noise, we consider the impact of different types of noise on the effectiveness of using image scoring to facilitate cooperation. We consider the following two types of noise, which might occur in an interaction.

- *Action noise*: with some small probability e_a the donation action of donor i towards recipient j fails, such that although $k_i \leq s_{ji}$ the recipient does not receive benefit b. For simplicity, we assume that in this case the donor does not incur the cost c and that (in the absence of perception noise) observers perceive the actual interaction outcome, i.e., they perceive no donation rather than the 'intended' donation. Thus, we interpret a failed donation action as being equivalent to the donor choosing not to donate. We also assume that action noise is only associated with the donation action, and therefore a decision to not donate (i.e., where $k_i > s_{ji}$) is unaffected by action noise.
- *Perception noise*: with some small probability e_p an observer o of an interaction, in which a donor i is paired with recipient j, incorrectly perceives the donor's action such that donation is perceived as not donating and vice-versa. Thus, with probability e_p the image score s_{io} of the donor i as seen by the observer o is decremented for donation and incremented for non-donation (bounded such that $s_{io} \in [-5, +5]$).

While the impact of perception noise has previously been considered in the context of generosity (with binary image scores) [23], to the best of our knowledge the impact of action noise on image scoring has not been investigated. In order to understand the impact of noise in the donation game, we performed experiments by adding small amounts of action and perception noise, both individually and in combination, to the base model presented by Nowak and Sigmund. We varied the level of noise (e_a and e_p) between 0.0 and 0.2 in increments of 0.025, such that either a single type of noise was present or both types of noise had equal probabilities (i.e., $e_a = e_p$). For other parameters we adopted the values used by Nowak and Sigmund, namely $b = 1$ and $c = 0.1$ for the benefit and cost of donation, with population size $n = 100$ and $m = 300$ interactions. In their experiments, to avoid negative payoffs, a value of 0.1 is added to each interaction by Nowak and Sigmund [12]. This means that agents are artificially rewarded purely for being selected to interact (since they receive a reward of 0.1, even if there is no donation), and does not reflect that donation has a cost (i.e., a negative reward). Therefore, in our instantiation of the donation game we do not add these artificial rewards, and we allow rewards to be negative.

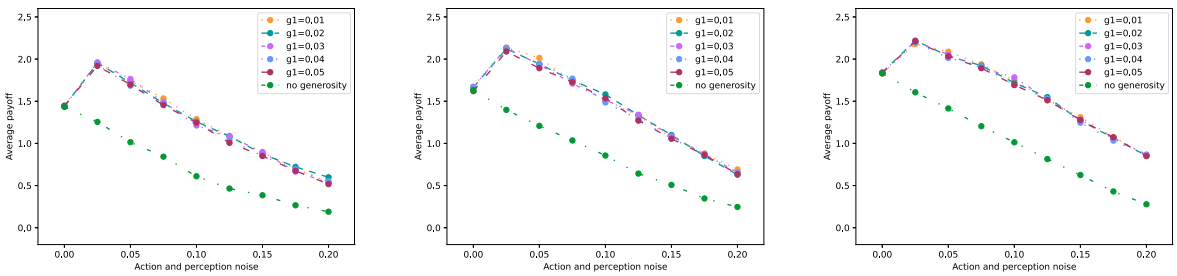
Figure 1 shows the impact on the average payoff of (a) action noise, (b) perception noise, and (c) the combination of action and perception noise on the base model with negative payoffs allowed[2], for full observation ($q = 1.0$) and partial observation ($q \in \{0.8, 0.9\}$). It is clear that while cooperation is still achieved in the presence of a small amount of noise, indicated by the positive average payoff, an increase in noise has a significant negative impact on cooperation (i.e., there are fewer donations, and so a lower average payoff). Perception noise has a greater individual impact than action noise, suggesting that the ability to observe behaviour to build reputation plays a fundamental role in achieving cooperation. This supports previous arguments that effective social norms require that reputation is publicly known [7, 23]. As is expected, when both types of noise are present the impact is the largest. Full observation ($q = 1$) yields the highest average payoff, with lower levels of observation resulting in lower payoff.

4 Generosity with nuanced reputation

Schmid *et al.* suggested *assessment generosity* and *action generosity* as potential methods to mitigate the impact of unreliable or non-public image scores [23]. They consider modified versions of the leading eight [14], which allow for generosity. In cases where the original leading eight norm would result in cooperation, the modified version also results in cooperation. Similarly, the assessment rule assigns a good reputation where the original version would do so. Assessment generosity is incorporated by assigning a good reputation with probability g_1 in cases where the original would assign a bad reputation. Action generosity is included by the modified version cooperating with probability g_2 in cases where the original would defect. Thus, generosity may cause cooperation when the original leading eight norms would defect (i.e., agents may be generous), but never causes agents to defect when the original would cooperate and never assigns a bad reputation if the original would assign good.

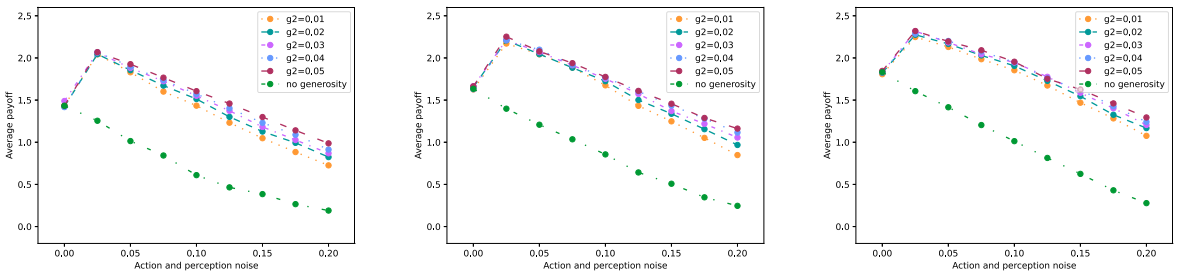
In the restricted setting of binary reputation, Schmid *et al.* found assessment generosity to be ineffective and reduce cooperation, while small amounts of action generosity could be helpful [23]. Since previous work points to the importance of reputation in supporting

[2] Note that we obtained similar results where negative payoffs are prevented, but they are omitted due to space constraints.



(a) $q = 0.8$ (b) $q = 0.9$ (c) $q = 1.0$

Figure 2. The effect of different probabilities of assessment generosity, g_1, for different levels of action and perception noise ($e_a = e_p$). Results are averaged over 100 runs of $g = 10^5$ generations, and show the average payoff for different probabilities, q, of onlookers observing interactions.



(a) $q = 0.8$ (b) $q = 0.9$ (c) $q = 1.0$

Figure 3. The effect of different probabilities of action generosity, g_2, for different levels of action and perception noise ($e_a = e_p$). Results are averaged over 100 runs of $g = 10^5$ generations, and show the average payoff for different probabilities, q, of onlookers observing interactions.

cooperation, we investigate replacing the binary reputation in Schmid *et al.*'s evaluation of generosity with more nuanced image scores ($s_{ji} \in [-5, +5]$). We also consider the impact of action noise.

Figures 2 and 3 show the impact of assessment and action generosity respectively on the average payoff, where image scores are in the range $[-5, +5]$. We consider the same generosity probabilities as Schmid *et al.*, namely $g_1, g_2 \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$, in the presence of equal amounts of action and perception noise (i.e., $e_a = e_p$). Our results differ significantly from those of Schmid *et al.*'s binary reputation case, with both assessment and action increasing average payoff, although we also see that action generosity is more effective than assessment generosity. We hypothesise that the increase in the reputation space, combined with the corresponding increase in strategy space, enables generosity to mitigate the impact of unreliable and partially observed image scores. As the probability of observation q increases, we see a corresponding increase in average payoff, suggesting that the increase in reliable reputation information from increased observation is supporting cooperation. For assessment generosity the probability of being generous, g_1, has little effect, with negligible difference between the average payoff for $g_1 = 0.01$ and $g_1 = 0.05$. Conversely, for action generosity the probability of being generous, g_2, has a small effect, with slightly higher average payoff for $g_2 = 0.05$ compared to lower probabilities.

An interesting characteristic of the results in Figures 2 and 3 is the increase in average payoff for a small amount of noise. As the level of noise increases the improvement in payoff rapidly reduces, as is expected, broadly at the same rate as the decrease in the base model without generosity. An exception to this is in the case of full observation $q = 1$ where the decrease in payoff as noise increases is at a lower rate than in the base case. One possibility for this behaviour is that the small level of noise provides addition mitigation from the impact of unreliable and partially observed image scores, given the larger reputation and strategy space.

5 Analysing Generosity

We now turn to a formal analysis of a simplified form of generosity. This simplification assumes that all agents are aware of, and agree on, a reputation value for each agent in the system and only action generosity (i.e., $q = 1, e_p = 0, g_1 = 0$). To mitigate against this simplification, we introduce a third type of noise in our analysis, namely *reputation noise*, denoted e_r. Here, with likelihood e_r, a donor will ignore a potential recipient's true image score, and instead view the recipient's image score as a random value, drawn uniformly from all possible image scores.

Our analysis builds on the approach introduced by [27], taking into account the wider range of image scores and strategies available to the agents, noise present in the system, and the possibility of (action) generosity. The goal of our analysis, following work such as [22], is to compute the *cooperation index* for a system, a value which indicates the likelihood that agents will play a cooperative action.

We begin by considering a donor agent following strategy k_d, and a recipient agent with image score i_r. Given reputation noise e_r, action noise e_a, and action generosity g_2, and denoting the total number of possible image scores in the system by $|R|$, the likelihood that the donor will successfully undertake a donation action is computed as follows.

$$p_D = (1-e_a)\times\begin{cases} 0 \text{ if } k_d > |R| \\ (1-e_r) + e_r(1 - \frac{|R|-k_d}{|R|} + g_2\frac{|R|-k_d}{|R|}) \text{ if } i_r \geq k_d \\ g_2(1-e_r) + e_r(1 - \frac{|R|-k_d}{|R|} + g_2\frac{r-k_d}{|R|}) \text{ otherwise} \end{cases}$$

As is commonly done [19, 22, 27], we assume that changes in strategy happen over significantly longer timescales than changes in reputation, meaning that the latter are stationary when the former occurs. Now consider a system with only two strategies, denoted s_1 and s_2. Assume that there are n agents in our system, with n_{s_1} following s_1, and $n_{s_2} = n - n_{s_1}$ following strategy s_2. We let the tuple pair $(I^1, I^2) = ((i^1_1, \ldots I^1_{|R|}), (i^2_1, \ldots i^2_{|R|}))$ where the value of i^1_j encodes the number of agents following strategy s_1 which have an image score indexed by j[3]. If we select agents for interaction at random, the likelihood of an agent with image i^a_j acting as a donor, and $i^b_{j'}$ acting as a recipient is computed as follows (note that $\hat{i}^b_{j'} = i^b_{j'}$ if $a \neq b$ or $j \neq j'$, and $\hat{i}^b_{j'} = i^b_{j'} - 1$ otherwise).

$$p_I = \frac{|i^a_m||\hat{i}^b_n|}{(|I^a| + |I^b|)(|I^a| + |I^b| - 1)}$$

Given a tuple pair, the values $p_D p_I$ and $(1 - p_D)p_I$ give us the probability of transitioning to a new tuple pair for which the image score of the donor is incremented or decremented to reflect donation or non-donation. We can thus construct a transition matrix whose entries describe the likelihood of transitioning between tuple pairs (I_1, I_2) and (I'_1, I'_2). This matrix is ergodic, and its fixed point captures the stationary distribution of image scores for the pair of strategies under consideration. Furthermore, using P_D, P_I, the computed likelihood of a tuple pair and the payoffs for donating and receiving a donation, we can compute the expected payoffs of each strategy. If n_{s_i} players play strategy i (leaving the other strategy implicit), we denote this expected payoff as $\pi_i(n_{s_i})$.

Over the longer term, agents can change strategies. In this timescale, we model strategy change as a Fermi process. That is, a random player is chosen from the population and, with likelihood μ, chooses a new strategy at random from the set of available strategies, while with likelihood $1 - \mu$, they select another player from the population and choose whether to adopt this latter player's strategy with probability

$$\frac{1}{1 + e^{-\mu(\pi_j - \pi_i)}}$$

Here π_i is the payoff of the agent choosing whether to change strategy, and π_j is the payoff of the observed agent.

Given a population of n agents where all but one follow a single strategy, the *fixation probability* determines the likelihood that this single strategy will take over the population [25]. In the case where strategies propagate following the Fermi process, the fixation probability for new strategy i and old strategy j is computed according to the following formula.

$$\rho_{ij} = \frac{1}{1 + \sum_{i=1}^{n-1} \Pi_{k=1}^{l} e^{-\mu(\pi_i(k) - \pi_j(k))}}$$

Using these fixation probabilities we can construct a transition matrix Λ such that

$$\Lambda_{ij} = \begin{cases} 1 - \epsilon \sum_{j \neq i} \rho_{ij} \text{ if i = j} \\ \epsilon\rho_{ij} \text{ otherwise} \end{cases}$$

[3] Thus, $\sum_j i^1_j = n_{s_1}$ and $\sum_j i^2_j = n_{s_2}$.

Here, ϵ is a small constant chosen to ensure that Λ_{ii} is positive. Furthermore, Λ is ergodic, and thus has a unique stationary distribution. This stationary distribution reflects the frequency with which the system is in a state where only one strategy exists under the assumption that mutations are rare [6]. We can therefore use this distribution to compute how often donation will occur by multiplying it with the likelihood of donation of a single strategy based on the image distribution determined by the image fixedpoints of the strategy playing against itself. The result of this analysis is a value referred to as the *cooperation index* [19].

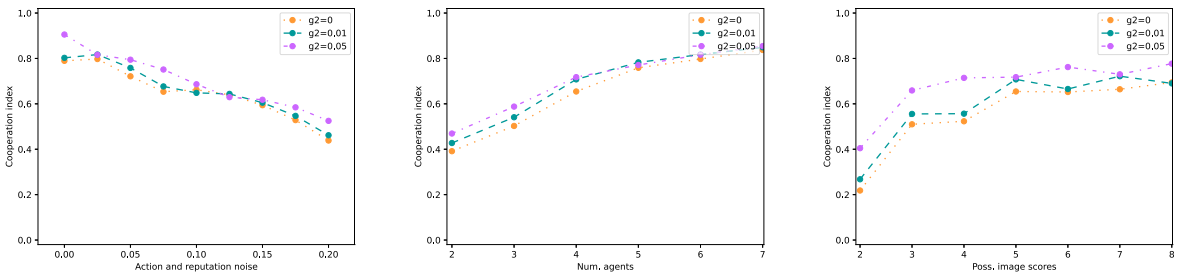
We ran our analysis on small systems with different parameters. Across all of our evaluations, as in other portions of this work, we set the cost of donating to 0.1, and the benefit of receiving a donation to to 1. We note that the number of tuples grows very quickly with the number of possible image scores and number of agents in the system, meaning that we had to restrict ourselves to very small systems. Figure 4 summarises our results.

Figure 4(a) considers the effect of increasing noise on the cooperation index under different generosity levels. Unsurprisingly, we see that cooperation decreases with increasing noise, but that generosity slightly mitigates this effect, as noted by Schmid *et al.* [23]. We note that for different numbers of agents and possible image scores our results show a significant uptick in cooperation in the presence of generosity at low noise levels, similar to the effects shown in Figure 3. However, as per Figure 4(a), this effect was not present in some system configurations. Figure 4(b) demonstrates that cooperation tends to rise with an increasing number of agents, and the effect of generosity decreases. This contrasts with the empirical results of Figure 3, and we hypothesise that this is possibly due to the low number of agents we evaluated against and more likely due to fewer possible image scores. This conclusion is supported by Figure 4(c) which both shows that cooperation increases when the possible number of image scores increases, and that increasing action generosity continues to affect the system. We believe that the irregularity of scores here (e.g., when $g_2 = 0.01$ and 0.05 for 5 agents) arises due to the low number of agents we evaluated against. Building on the ideas of [22], as future work we intend to perform a simulation-based analysis of the evolutionary dynamics for a larger number of agents and image scores than is possible with a purely formal analysis.

6 Forgiveness

Schmid *et al.* note that a limitation of generosity is the assumption of an 'aligned' setting, where all agents use the same fixed probabilities for assessment (g_1) and action (g_2) generosity [23]. This is unrealistic in a population of autonomous agents, in which we might expect individuals to make their own decisions about when to be generous. In this section, we propose assessment and action forgiveness which take a similar form to generosity, but in a 'non-aligned' manner, with each individual agent i having its own forgiveness strategy f_i. We assume that forgiveness strategies evolve in the population alongside donation strategies, such that when offspring are produced at the end of a generation (proportionally to their fitness in terms of payoff) an offspring inherits its parent's donation and forgiveness strategies, k_i and f_i (subject to the usual probability of mutation, p_m). Given the effectiveness of the leading eight in a restricted setting (with binary publicly known reputation and no noise), we hypothesise that a decision to forgive should depend on the potential beneficiary's image score, i.e., forgiveness should be a higher order norm.

In the donation game introduced by Nowak and Sigmund, the image score, s_{io} for a donor i perceived by an observer o is incremented



(a) (b) (c)

Figure 4. Results of the formal analysis of generosity showing the cooperation index for (a) different levels of action and reputation noise ($e_a = e_r$) for a system of 6 agents and 5 possible image scores; (b) changes in the number of agents under constant noise ($e_a = e_r = 0.025$) with 5 possible image scores; (c) changes in the number of possible reputation scores in a system with 4 agents under constant noise ($e_a = e_r = 0.025$).

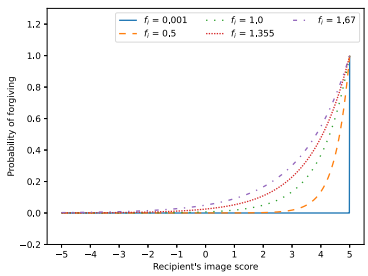


Figure 5. Probability of forgiving against the beneficiary's image score for different forgiveness strategies.

when i is paired with recipient j if i donates (i.e., if $k_i \leq s_{ji}$), and is decremented otherwise (bounded such that $s_{io} \in [-5, +5]$). Thus, in any future pairings where i takes the role of recipient, any previous choices not to donate reduce its chances of receiving a donation (since they reduce its reputation, i.e., its image score)[4]. In the absence of noise, an agent's image score is a direct result of its donation behaviour and, as shown by Nowak and Sigmund, this form of indirect reciprocity is able to support cooperation. However, as seen in Figure 1, the presence of noise causes a breakdown in cooperation and a reduction in the average reward received.

In human interactions, reputation is fundamental to establishing cooperation. However, in addition to reputation such environments typically also include a form of *forgiveness*, such that agents who do not cooperate are given a second chance [1]. Previous research on norm emergence in artificial environments has also suggested forgiveness to be important [8]. In Axelrod's seminal study of norm emergence, strategies which included an element of forgiveness were shown to be more effective [2]. We therefore investigate whether forgiveness is able to mitigate the impact of noise on cooperation in the context of a simple reputation mechanism, in the form of image scoring. Our hypothesis is that where a failure (or perceived failure) to donate is a result of noise, forgiveness will reduce (but not avoid) the fall in cooperation that results from using image score alone to determine whether to donate.

In this paper, we consider two kinds of forgiveness, as follows.

- Assessment forgiveness: when observing a non-donation an observer may forgive the donor and not decrement its image score.
- Action forgiveness: donor i may forgive recipient j by donating when j's image score is lower than i's strategy, i.e., $k_i \not\leq s_{ji}$.

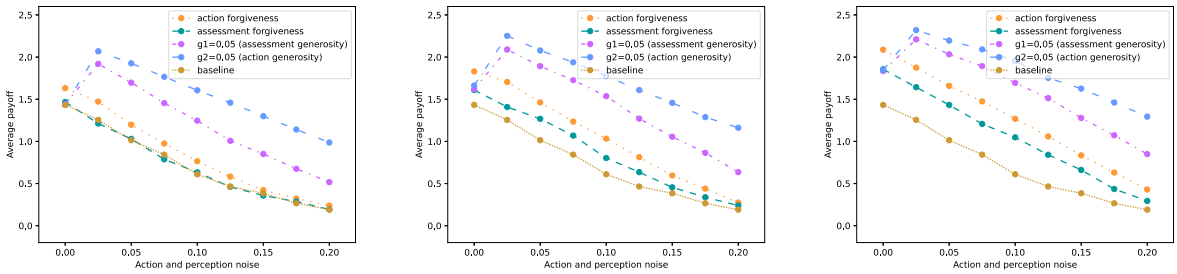
Assessment and action forgiveness are analogues of assessment and action generosity, with two key differences (in addition to our focus on nuanced reputation scores of $s_i \in [-5, +5]$ for agent i). First, each agent, i, has its own individual strategy for forgiveness, f_i, which evolves alongside donation strategies in the population. Second, as a way of approximating a recipient's intention, a donor considers the reputation of the recipient, such that those with a high reputation are assumed to be more likely to have good intentions, and those with a low reputation are more likely to have bad intentions. Previous work has shown that intention recognition can be an effective means of supporting forgiveness [1]. Thus, we use reputation as a proxy for intention, and an agent's forgiveness strategy determines the extent to which a recipient's reputation impacts on the probability of forgiveness. Past work on forgiveness suggested that apology is a prerequisite for forgiveness, as it gives an indication of intention, and that such apology should be costly as an indication of sincerity [10]. Here, we are interested in whether forgiveness can be effective in cases where there is no mechanism for apology. Our hypothesis is that using reputation as an approximation of a recipient's intention may enable an improvement in cooperation in noisy environments without the need for a formal mechanism for costly apology.

Assessment forgiveness defines the decision of an observer o regarding whether to forgive a donor i for an observed non-donation as a function of both the observer's forgiveness strategy f_o and their image score of the donor s_{io}. Specifically, the probability p_f of observer o forgiving donor i, i.e., not decrementing the image score s_{io} for a non-donation, is defined as:

$$p_f = e^{(s_{io} - s_{max})/f_o}$$

where s_{max} is the maximum possible image score (in our setting, and that of Nowak and Sigmund, $s_{max} = 5$). Thus, an observer is more likely to forgive if it has a high forgiveness strategy and the (defecting) donor has a high image score. We assume that each agent i has a forgiveness strategy $f_i \in \mathcal{F}$. In this paper we consider the set of possible forgiveness strategies $\mathcal{F} = \{0.001, 0.5, 1.0, 1.355, 1.67\}$ as representative values, with 0.001 corresponding to the probability of forgiving tending to 0 unless the defector has an image score of 5, and 1.67 meaning that the probability of forgiving is increasingly positive as the defector's image score becomes more positive. The values of 1.355 and 1.67 are selected such that in the initial case of all agents having a image score of 0, the probability of forgiving p_f

[4] Note that since image scoring is a first order social norm, this is the case even where an agent does not cooperate with another who has a bad reputation [14].



(a) $q = 0.8$ (b) $q = 0.9$ (c) $q = 1.0$

Figure 6. The effect of different probabilities of assessment and action forgiveness for different levels of action and perception noise ($e_a = e_p$). Results are averaged over 100 runs of $g = 10^5$ generations, and show the average payoff for different probabilities, q, of onlookers observing interactions. The best performing results for action and assessment generosity are also shown (i.e., the highest probability of generosity, $g1 = g2 = 0.05$).

will be approximately 0.25 and 0.5 respectively, meaning that p_f is in a similar range to the probability of assessment and action generosity, g_1 and g_2 considered by Schmid *et al.*. The impact of an agent's forgiveness strategy, f_i, on the probability of it forgiving can be seen in Figure 5, which shows the probability of forgiving for the possible image score values of the beneficiary.

Action forgiveness is defined similarly, affecting whether a donor will donate even if reputational information suggests it should not.

An agent's forgiveness strategy f_i, along with the donation strategy k_i, is propagated to subsequent generations as described in Section 2, such that with a small chance of mutation, p_m, an offspring i' will use random strategies for both $k_{i'}$ and $f_{i'}$.

Figure 6 shows the impact of assessment and action forgiveness on average payoff, where image scores are in $[-5, +5]$, for different probabilities q, of observing interactions. The set of possible forgiveness strategies is $\mathcal{F} = \{0.001, 0.5, 1.0, 1.355, 1.67\}$, as illustrated in Figure 5. For comparison, assessment and action generosity (with the best performing probabilities of $g_1 = g_2 = 0.05$) are also shown.

Our results show that assessment forgiveness has limited effect which is highly dependent on the level of observation. For a low level of observation ($q = 0.8$) there is negligible difference between the average payoff in the baseline case of Nowak and Sigmund's model, and assessment forgiveness. As observation levels increase ($q = 0.9$ and $q = 1$), assessment forgiveness is more effective, but gives significantly lower average payoff than action forgiveness or generosity. A possible reason for this is that assessment forgiveness causes image scores to become less accurate (i.e., by not reflecting each defection), which in turn may reduce cooperation. Since forgiveness is dependent on individuals' forgiveness strategies, which are updated each generation, assessment forgiveness has a lower average payoff than assessment generosity where the probability of not reducing the image score of a defector is fixed throughout the entire simulation.

Action forgiveness has a much greater positive impact on the average payoff than assessment forgiveness, again with a greater increase for higher levels of observation. The results in Figure 6 demonstrate that in a 'non-aligned' setting, with nuanced (i.e., non-binary) reputation, action forgiveness is moderately effective at mitigating the impact of unreliable non-public image scores and action and perception noise. Interestingly, the average payoff from action forgiveness is below both assessment and action generosity. Our hypothesis is that this is because action forgiveness enables a donor to cooperate with a recipient who is generally good, i.e., although the recipient's image score might be below the donor's strategy it is sufficiently positive that there is a chance of forgiving. However, since all forgiveness strategies result in a very small forgiveness probabilities for negative image scores, 'known bad' recipients are not forgiven. This is similar to the 'stern judging' norm which allows for justified defection, and has been previously shown to be effective [17]. In contrast, action generosity has a fixed probability of donating when the donor should defect and is not able to discriminate whether generosity is appropriate given the recipient's image score.

7 Discussion, Future Work and Conclusions

This paper makes several contributions to the investigation of reciprocity. First, we examine the effect of generosity in settings where agents have more than just "good" and "bad" reputation. Second, we provide an analytical analysis of cooperation in such settings. Finally, we introduce the notion of forgiveness and investigate its effects.

Our results — in contrast to the work of Schmid *et al.* — demonstrate that generosity has a positive effect on payoffs across the system, reflecting increased cooperative behaviour in the context of nuanced reputation values. Curiously, we note an improvement in cooperation in the presence of slight noise and generosity in contrast to the no-noise case, a result we are unable to fully explain, though we note that this result is somewhat supported by our analytical analysis. Turning to this analysis, which considers only a small number of agents and up to 8 different reputation values, we observed that cooperation increases as both the number of agents and reputation values increase, with the latter highlighting the utility of nuance in reputation. Turning to forgiveness we were surprised to discover that the use of this individualised strategy did not lead to an improvement in comparison to the use of global generosity. Since forgiveness could converge to a universal value (and therefore mirror generosity), we intend to investigate why this occurs in more detail as part of our future work. We are also pursuing several other directions of work. First, we intend to use a hybrid simulation/analytical approach, as described in [22], to investigate the behaviour of forgiveness and generosity in larger system with nuanced reputation and different types of noise. We are also investigating whether more complex strategies and strategy-learning mechanisms can improve cooperation in the system. Finally, we are examining the effects of reputation noise in conjunction with forgiveness strategies.

Acknowledgements

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

[1] H. T. Anh, L. M. Pereira, and F. C. Santos, 'Intention recognition promotes the emergence of cooperation', *Adaptive Behaviour*, **19**(4), 264–279, (2011).

[2] R. Axelrod, 'An evolutionary approach to norms', *The American Political Science Review*, **80**(4), 1095–1111, (1986).

[3] H. Brandt and K. Sigmund, 'The logic of reprobation: assessment and action rules for indirect reciprocation', *Journal of Theoretical Biology*, **231**(4), 475–486, (2004).

[4] H. Brandt and K. Sigmund, 'Indirect reciprocity, image scoring, and moral hazard', *PNAS*, **102**(7), 2666–2670, (2005).

[5] J. A. Cuesta, C. Gracia-Lázaro, A. Ferrer, Y. Moreno, and A. Sánchez, 'Reputation drives cooperative behaviour and network formation in human groups', *Scientific Reports*, **5**(7843), (2015).

[6] Drew Fudenberg and Lorens A. Imhof, 'Imitation processes with small mutations', *Journal of Economic Theory*, **131**, 251–262, (2006).

[7] C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, and M. A. Nowak, 'Indirect reciprocity with private, noisy, and incomplete information', *PNAS*, **115**(48), 12241–12246, (2018).

[8] C. D. Hollander and A. S. Wu, 'The current state of normative agent-based systems', *Journal of Artificial Societies and Social Simulation*, **14**(2), (2011).

[9] M. Krellner and H. T. Anh, 'Pleasing enhances indirect reciprocity based cooperation under private assessment', *Artificial Life*, **27**(3-4), 246–276, (2021).

[10] L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, and T. Lenaerts, 'Apology and forgiveness evolve to resolve failures in cooperative agreements', *Scientific Reports*, **5**(10639), (2015).

[11] M. Milinski, D. Semmann, and H.-J. Krambeck, 'Reputation helps solve the 'tragedy of the commons'', *Nature*, **415**, 424–426, (2002).

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

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

[14] H. Ohtsuki and Y. Iwasa, 'The leading eight: Social norms that can maintain cooperation by indirect reciprocity', *Journal of Theoretical Biology*, **239**, 435–444, (2006).

[15] H. Ohtsuki, Y. Iwasa, and M. A. Nowak, 'Indirect reciprocity provides only a narrow margin of efficiency for costly punishment', *Nature*, **457**(79-82), (2009).

[16] I. Okada, 'A review of theoretical studies on indirect reciprocity', *Games*, **11**(3), 27, (2020).

[17] J. M. Pacheco, F. C. Santos, and F. A. C. C. Chalub, 'Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity', *PLoS Computational Biology*, **2**(12), 1634–1638, (2006).

[18] K. Panchanathan and R. Boyd, 'A tale of two defectors: the importance of standing for evolution of indirect reciprocity', *Journal of Theoretical Biology*, **224**(1), 115–126, (2003).

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

[20] T. Sasaki, I. Okada, and Y. Nakai, 'The evolution of conditional moral assessment in indirect reciprocity', *Scientific Reports*, **7**, 41870, (2017).

[21] L. Schmid, K. Chatterjee, C. Hilbe, and M. A. Nowak, 'A unified framework of direct and indirect reciprocity', *Nature Human Behaviour*, **5**, 1292–1302, (2021).

[22] L. Schmid, F. Ekbatani, C. Hilbe, and K. Chatterjee, 'Quantitative assessment can stabilize indirect reciprocity under imperfect information', *Nature Communications*, **14**(1), 2086, (2023).

[23] L. Schmid, P. Shati, C. Hilbe, and K. Chatterjee, 'The evolution of indirect reciprocity under action and assessment generosity', *Scientific Reports*, **11**(17443), (2021).

[24] D. Semmann, H.-J. Krambeck, and M. Milinski, 'Strategic investment in reputation', *Behavioral Ecology and Sociobiology*, **56**(3), 248–252, (2004).

[25] A. Traulsen and C. Hauert, 'Stochastic evolutionary game dynamics', *Reviews of nonlinear dynamics and complexity*, **2**, 25–61, (2009).

[26] C. Wedekind and V. A. Braithwaite, 'The long-term benefits of human generosity in indirect reciprocity', *Current Biology*, **12**(12), 1012–1015, (2002).

[27] J. Xu, J. García, and T. Handfield, 'Cooperation with bottom-up reputation dynamics', in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, p. 269–276, (2019).