Indirect reciprocity with stochastic and dual reputation updates

Yohsuke Murase[1,2] and Christian Hilbe[2]

[1]RIKEN Center for Computational Science, Japan
[2]Max Planck Research Group 'Dynamics of Social Behavior', Max Planck Institute for Evolutionary Biology, Germany
yohsuke.murase@gmail.com

Abstract

Cooperation is essential for both human and artificial life societies, yet understanding how to promote it remains a complex challenge. Indirect reciprocity, where individuals cooperate to maintain a good reputation, is one mechanism to encourage cooperation. To promote stable cooperation, society needs social norms that stipulate how individuals should behave and how they should evaluate others. Previous research has identified a set of effective social norms, called the "leading eight", for achieving evolutionarily stable cooperation. In this study, we expand on a classical framework in two significant ways. First, we include norms that update the reputations of passive receivers. Second, we introduce stochasticity to social norms. We theoretically derived the necessary and sufficient conditions for evolutionarily stable norms that result in full cooperation within this generalized model. Our findings offer a new perspective on prior research and provide a foundation for future studies in this field.

Introduction

Cooperation is essential for both human and artificial life societies, as both exhibit a remarkable ability to work together. In these contexts, cooperative interactions can be rationalized when they occur publicly, as cooperation may help individuals or agents gain a good reputation, which can be valuable in future interactions. This process, known as indirect reciprocity, is one of the most fundamental mechanisms for maintaining cooperation.

In models of indirect reciprocity, the interplay between an individual's actions and the resulting reputation is governed by a community's social norm. Social norms can be conceptualized as a combination of an assessment rule and an action rule. The assessment rule determines how reputations are assigned to community members, depending on who did what to whom. The norm's action rule determines how people should act, which may depend on their own reputation and the reputation of their interaction partner. One of the aims of indirect reciprocity studies is to identify social norms that lead to stable cooperation.

A seminal work by Ohtsuki and Iwasa identified a set of effective norms known as the 'leading eight' (Ohtsuki and Iwasa, 2004, 2006). A few examples of the leading eight norms are shown in Table 1. These norms enable full cooperation in the limit of low errors and also form strict Nash equilibria. Here, we refer to norms that meet these criteria as cooperative ESS (CESS). Because of the simplicity and effectiveness of the leading eight, they have become the primary reference for many subsequent theoretical studies.

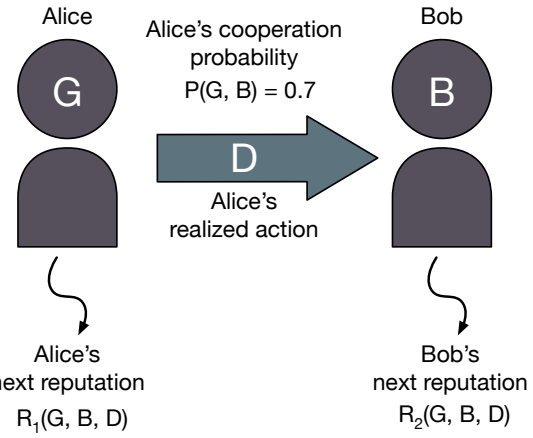

Figure 1: A schematic diagram of the model. At each time step, a donor (Alice) and a recipient (Bob) are randomly selected from the population. Suppose their current reputations are Good (G) and Bad (B), respectively. Alice decides her action according to her action rule P, which returns the probability of cooperation. In this example, the cooperation probability is $P(G, B) = 0.7$, and the realized action is defection (D). The reputation of Alice and Bob are updated according to the assessment rules R_1 and R_2. Alice and Bob get G reputation with probability $R_1(G, B, D)$ and $R_2(G, B, D)$, respectively. This process iterates, changing the donor-recipient pairs, and the payoffs are accumulated over time. Reprinted from (Murase and Hilbe, 2023).

In this work (Murase and Hilbe, 2023), we extend this previous work in two ways. Our first extension addresses the reputations of recipients. In previous studies, social norms only determined how the donor's reputation is updated while the recipient's reputation remained constant. However, some

Table 1: The prescriptions of the leading eight norms. We only show two of these: Simple Standing (L3) and Stern Judging (L6) for simplicity. The reputations are represented by G (good) and B (bad), and the actions are cooperation (C) and defection (D). The following reputations are assigned to the donor depending on the reputation of the recipient and the action taken by the donor.

	L3	L6
cooperation with G	G	G
defection against G	B	B
cooperation with B	G	B
defection against B	G	G

empirical works (Jordan and Kouchaki, 2021) as well as everyday experience suggest that some social interactions also affect the reputations of passive receivers. As illustrated in Fig. 1, we introduce a "dual reputation update" framework, where the reputations of both the donor and the recipient are updated by assessment rules R_1 and R_2, respectively.

Our second extension introduces stochasticity into the model. Most previous studies presume social norms to be deterministic. Deterministic norms have the formal advantage that they can be enumerated, and hence they can be studied exhaustively. However, in reality, actions and assessments may not need to be deterministic for various reasons. As shown in Fig. 1, the action rule P and the assessment rules R_1 and R_2 return the probabilities of cooperation and reputation updates, respectively. Our model includes the deterministic norms as special cases, and we recover the deterministic norms, including the leading eight, within the respective limits.

Main Results

Although our generalized model has an infinite number of norms preventing us from a comprehensive enumeration, we theoretically derive the necessary and sufficient conditions for CESS norms in this model. For the specific forms and the derivations of the conditions, see Eq. (29) and Eq. (34) in (Murase and Hilbe, 2023). Since we do not have space to present all the results here, we focus on a few main findings.

First, we consider a special case where the rules are deterministic without updating the recipient's reputation. This special case falls back to the classical model of indirect reciprocity. In this case, we comprehensively identified all the CESS norms from the results obtained for the generalized model. As expected, we reproduce the leading eight norms. In addition, we find another set of CESS norms that have not been identified in the previous works. These norms, called the "secondary sixteeen", are also effective in promoting cooperation when the benefit of cooperation is $b/c > 2$. We also showed that the leading eight and the secondary sixteen are the only CESS norms when the rules are deterministic and the recipient's reputation is kept constant. It is impossible to construct another CESS norm even for a higher b/c.

Second, we explore the space of all deterministic norms, including those that update the recipient's reputation. In total, there are $524,800$ deterministic norms. Among, we find $2,944$ CESS norms. These CESS norms fall into several discrete classes based on the minimal b/c ratio required for cooperation and the speed at which cooperation recovers from erroneous defections. We find that $R_2(G, B, D)$ (how a bad recipient is assessed when he/she is defected against) is particularly important. When a bad recipient recovers their reputation after being punished, society has a quicker recovery from errors. However, this also entails a drawback: The minimal b/c ratio required for cooperation increases. This trade-off is not limited to the deterministic norms but also applies to stochastic norms, indicating that it is a fundamental feature of indirect reciprocity.

Lastly, we consider stochastic norms. In deterministic CESS norms, defection against a good recipient is always regarded as bad, punishment by a good donor towards a bad recipient is always justified, and cooperation from a bad donor towards a good recipient (interpreted as an apology) is always accepted. However, we found that these behaviors are not necessarily consistent in stochastic norms. Defection may not always be assessed as bad, punishment may not always be justified, and apologies may not always be accepted. If the probabilities of these events meet certain conditions, the norm can still promote cooperation. Some of these behaviors are empirically observed as well. For instance, punishment towards anti-social individuals is sometimes only partially justified (Yamamoto et al., 2020).

Overall, our results provide a solid theoretical foundation and a new perspective for understanding the norms that society should follow to maintain cooperation. In future theoretical and experimental studies, our results would serve as a reference for exploring the space of social norms and understanding the trade-offs between different norms.

References

Jordan, J. J. and Kouchaki, M. (2021). Virtuous victims. *Science Advances*, 7(42):eabg5902.

Murase, Y. and Hilbe, C. (2023). Indirect reciprocity with stochastic and dual reputation updates. *PLOS Computational Biology*, 19(7):e1011271.

Ohtsuki, H. and Iwasa, Y. (2004). How should we define goodness? – reputation dynamics in indirect reciprocity. *J. Theor. Biol.*, 231(1):107–120.

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

Yamamoto, H., Suzuki, T., and Umetani, R. (2020). Justified defection is neither justified nor unjustified in indirect reciprocity. *PloS One*, 15(6):e0235137.