Evolution of Cooperation via Joint Commitments and Reputation

Marcus Krellner[1,2] and The Anh Han[1]
[1]Teesside University, UK, [2]University of St. Andrews, UK
E-Mail: Krellner.Marcus@gmx.de, T.Han@tees.ac.uk

Introduction

Joint commitments play such an essential role for human behaviour that they have been argued to "make our social world" (Gilbert, 2014). The ability to arrange joint commitments with others was shown to be a fundamental difference to other primates (Tomasello et al., 2012; Tomasello, 2019). A commitment is commonly defined as a form of a promise to do a certain thing (or to refrain from doing it). A joint commitment is a special form which involves two parties promising something, but the promises are only binding if both commit. For example, in a marriage ceremony, both have to say "I do" and only afterwards their commitments are valid.

Yet, the prevalence of joint commitments is puzzling. When we just need to coordinate for the best mutual outcome, any (single) commitment is beneficial. When we are tempted to free-ride (i.e. in social dilemmas), however, commitment serves no obvious purpose. Our commitment does not remove the incentive for our partner to exploit us (Nesse, 2001; Han, 2022; Han et al., 2013), and since we are also still incentivized to exploit them, our partner has good reason to doubt our commitment.

A solution to this problem is a reputation system. These have been studied extensively for so-called indirect reciprocity (Nowak and Sigmund, 1998; Ohtsuki and Iwasa, 2006; Okada, 2020). Whereas those models investigated what rules should determine if somebody is worthy of cooperation, Krellner and Han (2023b) study what makes a person worthy of trust. Trust determines if individuals enter joint commitments, and commitment determines if they cooperate. This important intermediate step has the potential to solve the problem of disagreement (Panchanathan and Boyd, 2003; Krellner and Han, 2022), which has occupied the research on indirect reciprocity for a long time (Uchida, 2010; Hilbe et al., 2018; Radzvilavicius et al., 2019; Krellner and Han, 2021, 2023a; Perret et al., 2021) .

Herein we summarise a recent work showing that a reputation system, which judges action in social dilemmas only after joint commitment, can prevent free-riding (Krellner and Han, 2023b). The study proposes the following principles: Keeping commitments builds trust. We can selectively enter joint commitments with trustworthy individuals. Making them enter such commitments aims to ensure their cooperation (since they will now be judged by whether they uphold their commitment). We simply do not commit to cooperate with those we do not trust, and hence we can freely defect in these situations without losing the trust of others.

Results

Krellner and Han (2023b) use a model akin to the latest advancements in indirect reciprocity, in which players hold private opinions about all other players (Uchida, 2010; Hilbe et al., 2018), using methods from evolutionary game theory (Sigmund, 2016) and agent-based simulations (Sayama, 2015; Adami et al., 2016). They analytically predict average reputation values (Fujimoto and Ohtsuki, 2022), and use them to calculate payoffs.

Krellner and Han (2023b) consider 9 possible strategies in co-presence. One, named 'RA', upholds both principles of joint commitment: Commit only when meeting a good player, and cooperate only in joint commitments (see Figure 1). The other strategies are less discriminatory, committing always or never, and/or cooperating always or never. It's shown that RA fares best, resulting in high levels of cooperation over time ($>90\%$). It can easily invade defective strategies, which always defect but never commit, and also faking strategies, which commit but never follow through. Yet RA is not fully stable, since a few more naive cooperators have a slight advantage over it. They are in turn overtaken by fakers or defectors, causing a somewhat cyclic behaviour of the population. However, the periods of abundance of RA are lasting.

The study demonstrates that reputation and joint commitments can sustain cooperation in the Prisoner's Dilemma even without repeated interactions. The only condition is that the cost of setting up joint commitments is lower than the net benefit of cooperation (i.e. benefit for the partner minus the cost of cooperation for oneself). In fact, a higher cost of joint commitments tends to stabilize RA and increase the amount of cooperation.

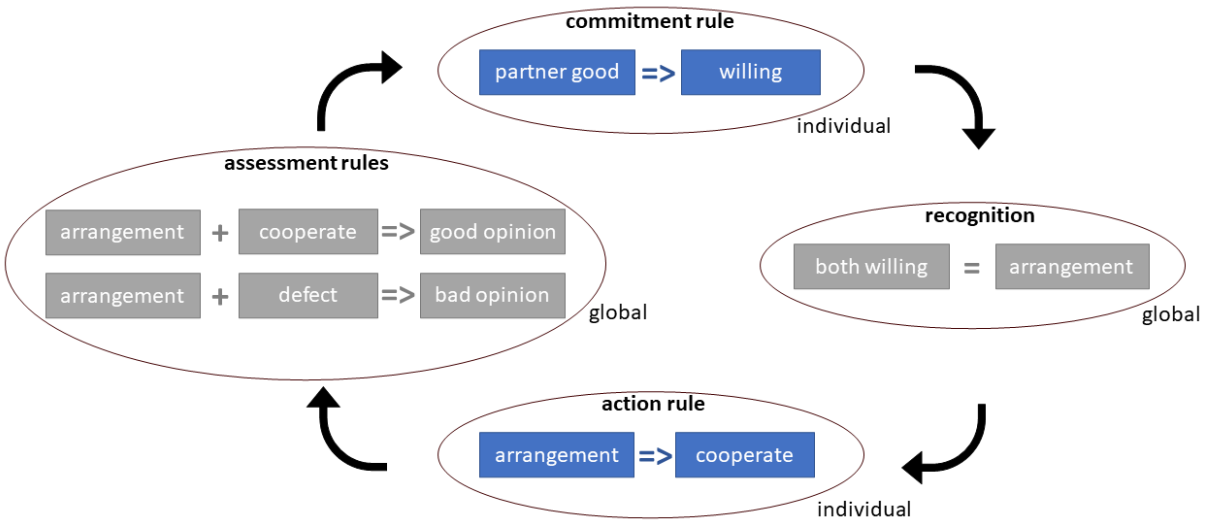

Figure 1: Principles of Joint Commitments. Different strategies were analysed in an evolutionary setting, only one of which upholds all principles of joint commitment. In particular, individual principles (blue) were varied among the strategies studied, whereas all strategies shared the global aspects (grey). However, not applying individual principles can render global principles practically meaningless (e.g. a strategy that never enters joint commitment has no meaningful use of trust). The symbol '=>' indicates 'if and only if', so that rules could be simplified (i.e. these rules were not listed: partner bad => not willing, no arrangement => defect, no arrangement + any action => no assessment.)

Discussion

Overall, Krellner and Han (2023b) demonstrated the potential of joint commitments in fostering stable cooperation, even if interactions are one-shot and even if there is no additional mechanism to enforce commitments. This suggests that joint commitments could have played an important role even before the existence of law enforcement. Written contracts are a special form of joint commitments. Today, we rely on courts and other instruments of modern states to make parties fulfil their commitments. But, such contracts might be based on a long tradition of joint commitments that have become second nature to us.

The amount of cooperation observed in Krellner and Han (2023b) surpasses that observed with unaltered indirect reciprocity under similar conditions, i.e. private assessments (Hilbe et al., 2018), by far. However, indirect reciprocity does not only work for the Prisoner Dilemma, but also for the donation game. This game has a crucial difference to the conditions studied here. In it, only one player acts. Therefore, a player cannot use joint commitment to alter the behaviour of their partner, rendering commitment pointless again. However, considering one simple alteration, the results also hold for the donation game. The results still apply, if players first meet, decide to commit, and only then it is revealed who is donor and who is recipient in the game. Such conditions are realistic, for example in everyday friendships, in which it is not determined who might need help moving their sofa next. It also holds in more elaborate joint commitments, such as defensive alliances like NATO, in which it is not clear, which nation might be attacked in the future.

Moreover, it is shown that cooperation rates declined slightly, when the costs of joint commitments decreased. This hints that some extraordinary rituals of joint commitment, such as marriage ceremonies, are not only expensive to ensure that the commitment of both parties is broadcasted as widely as possible, but that spending itself may serve a function.

In conclusion, Krellner and Han (2023b)'s results clearly demonstrated the potential for joint commitment and reputation to fostering cooperation, shedding light on their significance predating modern enforcement mechanisms. These findings have implications for the evolution of human behaviour and could also function as blueprints to establish cooperation between artificial agents.

Acknowledgements

TAH acknowledges generous support from EPSRC (grant EP/Y00857X/1).

References

Adami, C., Schossau, J., and Hintze, A. (2016). Evolutionary game theory using agent-based methods. *Physics of life reviews*, 19:1–26.

Fujimoto, Y. and Ohtsuki, H. (2022). Reputation structure in indirect reciprocity under noisy and private assessment. *Scientific Reports*, 12(1):1–13.

Gilbert, M. (2014). *Joint Commitment: How We Make the Social World*. Oxford University Press, New York.

Han, T. A. (2022). Institutional incentives for the evolution of committed cooperation: ensuring participation is as important as enhancing compliance. *Journal of The Royal Society Interface*, 19(188):20220036.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013). Good agreements make good friends. *Scientific reports*, 3(1):2695.

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., and Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, page 201810565.

Krellner, M. and Han, T. A. (2021). Pleasing Enhances Indirect Reciprocity-Based Cooperation Under Private Assessment. *Artificial Life*, pages 1–31.

Krellner, M. and Han, T. A. (2022). The Last One Standing? - Recent Findings on the Feasibility of Indirect Reciprocity under Private Assessment. In *The 2022 Conference on Artificial Life*, volume 1, Cambridge, MA. MIT Press.

Krellner, M. and Han, T. A. (2023a). We both think you did wrong – How agreement shapes and is shaped by indirect reciprocity.

Krellner, M. and Han, T. A. (2023b). Words are not Wind – How Joint Commitment and Reputation Solve Social Dilemmas, without Repeated Interactions or Enforcement by Third Parties. *Preprint arxiv: http://arxiv.org/abs/2307.06898*.

Nesse, R. (2001). *Evolution and the capacity for commitment*. Russel Sage Foundation.

Nowak, M. A. and Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, 194(4):561–574.

Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4):435–444.

Okada, I. (2020). A Review of Theoretical Studies on Indirect Reciprocity. *Games*, 11(3):27.

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

Perret, C., Krellner, M., and Han, T. A. (2021). The evolution of moral rules in a model of indirect reciprocity with private assessment. *Scientific Reports*, 11(1):23581.

Radzvilavicius, A. L., Stewart, A. J., and Plotkin, J. B. (2019). Evolution of empathetic moral evaluation. *eLife*, 8:8–10.

Sayama, H. (2015). *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks [Imprint].

Sigmund, K. (2016). *The calculus of selflessness*. Princeton University Press.

Tomasello, M. (2019). The Moral Psychology of Obligation. *Behavioral and Brain Sciences*.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., and Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence Hypothesis. *Current Anthropology*, 53(6):673–692.

Uchida, S. (2010). Effect of private information on indirect reciprocity. *Physical Review E*, 82(3):036111.