An evolutionary game theory analysis of trust in repeated games and lessons for human-machine interactions

The Anh Han[1], Cedric Perret [2] and Simon T. Powers [3]

[1] School of Computing, Engineering and Digital Technologies, Teesside University
[2] College of life and environmental sciences, University of Exeter
[3] School of Computing, Edinburgh Napier University
Emails: t.han@tees.ac.uk, cedric.perret.research@gmail.com, S.Powers@napier.ac.uk

Introduction The actions of intelligent agents, such as chatbots, recommender systems, and virtual assistants are typically not fully transparent to the user (Beldad et al., 2016; Chung et al., 2017). Consequently, users take the risk that such agents act in ways opposed to the users' preferences or goals (Luhmann, 1979). It is often argued that people use trust as a cognitive shortcut to reduce the complexity of such interactions. In our recent work (Han et al., 2021), we study this by using the methods of evolutionary game theory (EGT) to examine the viability of trust-based strategies in the context of an iterated prisoner's dilemma (IPD) game (Axelrod, 1984a; Sigmund, 2010). We show that these strategies can reduce the opportunity cost of verifying whether the action of their co-player was actually cooperative, and out-compete strategies that are always conditional, such as Tit-for-Tat. We argue that the opportunity cost of checking the action of the co-player is likely to be greater when the interaction is between people and intelligent artificial agents, because of the reduced transparency of the agent.

Trust-based strategies In our work, trust-based strategies are reciprocal strategies that cooperate as long as the other player is observed to be cooperating. Unlike classic reciprocal strategies, once mutual cooperation has been observed for a threshold number of rounds they stop checking their co-player's behaviour every round, and instead only check it with some probability. By doing so, they reduce the opportunity cost of verifying whether the action of their co-player was actually cooperative. We demonstrate that these trust-based strategies can out-compete strategies that are always conditional, such as Tit-for-Tat, when the opportunity cost is non-negligible.

We argue that this cost is likely to be greater when the interaction is between people and intelligent agents, since the interaction becomes less transparent to the user (e.g. when it is done over the internet (Grabner-Kraeuter, 2002)), and artificial agents have limited capacity to explain their actions compared to humans (Pu and Chen, 2007). Consequently, we expect people to use trust-based strategies more frequently in interactions with intelligent agents. Our results provide new, important insights into the design of mechanisms for facilitating interactions between humans and intelligent agents, where trust is an essential factor. Note that previous EGT models studying conditional strategies in repeated games usually ignore this cost or assume it to be very small (Han et al., 2011; Martinez-Vaquero et al., 2015; Hilbe et al., 2017; Qu et al., 2019; Kurokawa, 2017).

Model We consider a finite population of constant size N. At each time step, or generation, a random pair of players are chosen to play with each other. Interactions are modelled as a repeated symmetric two-player Prisoner's Dilemma game, defined by the following payoff matrix (for row player)

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}. \end{array}$$

Players can choose either to cooperate (C) or defect (D) in each round. After each round, there is a probability w that yet another round of the game will take place. In such repeated interactions, the strategy called tit-for-tat (TFT) has been shown to be particularly successful (Axelrod, 1984b; Axelrod and Hamilton, 1981). *TFT* starts by cooperating, and does whatever the opponent did in the previous round. As a conditional strategy, TFT incurs an additional opportunity cost, denoted by ϵ, compared to the unconditional strategies, namely, ALLC (always cooperate) and ALLD (always defect). This cost involves a cognitive cost (to memorise previous interaction outcomes with co-players and make a decision based on them). But crucially, it also involves a cost of revealing the actual actions of co-players – did the co-player act cooperatively or not?

We consider a new trust-based strategy, called TUC, that is capable of switching off the costly deliberation process and the checking of co-players' actions when it trusts its co-players enough. This strategy starts an IPD interaction as a TFT player. When its ongoing trust level towards the co-player—defined here as the difference between the number of cooperative and defect moves from the co-player so

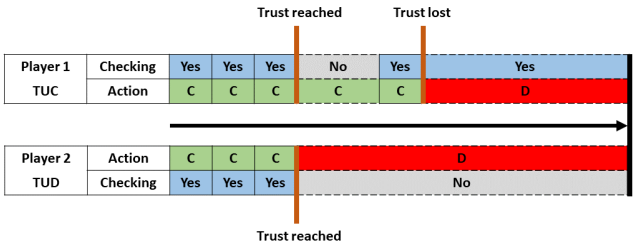

Figure 1: Diagram representing repeated interactions between a trust-based cooperator TUC and a trust-based defector TUD. In this case, $\theta = 3$.

far in the IPD—reaches a certain threshold, denoted by θ, the strategy will play cooperate unconditionally. We assume that TUC will check, with a probability p, the co-player's actions after switching off. If the co-player is found out to defect, TUC will revert to its initial strategy TFT and will not trust again its co-player. As a (defect) counterpart of TUC, we consider a strategy called TUD that whenever the ongoing trust level reaches the threshold θ, switches to playing defect unconditionally. These two strategies are illustrated interacting with each other in Figure 1.

We investigate the evolutionary success of trust-based strategies (TUC and TUD) in the presence of three other strategies: AllC, AllD and TFT. We use methods previously developed to analyse evolutionary game dynamics with finite populations and assuming small mutation rates (Imhof et al., 2005), to calculate the average time the population spends in using each of the possible strategies. For full description of the method, see (Traulsen et al., 2006; Sigmund, 2010).

Results Can we expect individuals to use trust? The top panel of Figure 2 shows that TUC is the most common strategy for a low to intermediate opportunity cost ϵ (between 0 and 0.3). When the opportunity cost ϵ is zero, both TUC and TFT are successful strategies and the population is composed of either one of them for most of the time. The success of TUC and TFT is explained by the capacity of these strategies to maintain high levels of cooperation within their homogeneous populations, while avoiding exploitation by AllD. Yet, the success of TFT is limited by the opportunity cost paid to check its partner's actions. This is shown in the results by the population being mostly AllD when the opportunity cost ϵ is high. Compared to TFT, TUC can limit this opportunity cost by reducing its attention to its partner's actions once trust is reached. This is why as the opportunity cost increases, the frequency of TFT plummets while TUC becomes more commonly observed.

Does the presence of trust increase the frequency of cooperation? To answer this question, we compare the frequency of cooperation between populations where trust strategies are allowed or not. The bottom panel of Figure 2 shows

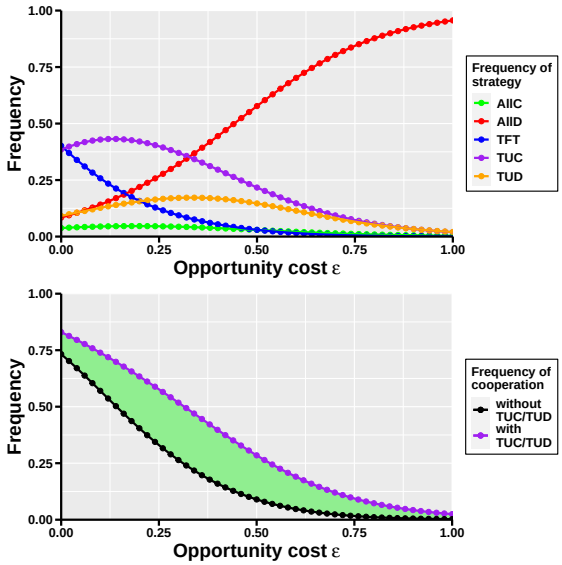

Figure 2: **Top**: Frequency of strategies as a function of the opportunity cost, ϵ. **Bottom**: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ. Other parameters: $\theta = 3$, $N = 100$.

that the presence of trust-based strategies increases the frequency of cooperation. Importantly, this increase happens even when the opportunity cost ϵ is high, and not only when TUC is the most frequent, i.e. for low ϵ. This is because a high frequency of cooperation is already reached for a low opportunity cost due to TFT. The presence of TUC has a more important effect on cooperation when the opportunity cost increases, since in that case the performance of TFT significantly reduces.

Conclusions Trust is a commonly observed mechanism in human interactions, and discussions on the role of trust are being extended to social interactions between humans and intelligent machines (Andras et al., 2018). It is therefore important to understand how people behave when interacting with those machines; particularly, whether and when they might exhibit trust behaviour towards them? Answering this is crucial for designing mechanisms to facilitate human-intelligent machine interactions, e.g. in engineering prosociality in a hybrid society of humans and machines (Paiva et al., 2018). To this extent, we have summarised our recent analysis showing that trust-based cooperation is a particularly common strategy, especially in interactions with moderate opportunity cost, and promotes cooperation for a large range of opportunity costs (Han et al., 2021).

Acknowledgements T.A.H. acknowledges support from Future of Life Institute (grant RFP2-154) and Leverhulme Research Fellowship (RF-2020-603/9).

References

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., Urquhart, N., and Wells, S. (2018). Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine*, 37(4):76–83.

Axelrod, R. (1984a). *The Evolution of Cooperation*. Basic Books, Inc., NY.

Axelrod, R. (1984b). *The Evolution of Cooperation*. Basic Books, ISBN 0-465-02122-2.

Axelrod, R. and Hamilton, W. (1981). The evolution of cooperation. *Science*, 211:1390–1396.

Beldad, A., Hegner, S., and Hoppen, J. (2016). The effect of virtual sales agent (VSA) gender – product gender congruence on product advice credibility, trust in VSA and online vendor, and purchase intention. *Computers in Human Behavior*, 60:62–72.

Chung, H., Iorga, M., Voas, J., and Lee, S. (2017). Alexa, can i trust you? *Computer*, 50(9):100–104. Conference Name: Computer.

Grabner-Kraeuter, S. (2002). The role of consumers' trust in online-shopping. *Journal of Business Ethics*, 39(1):43–50.

Han, T. A., Moniz Pereira, L., and Santos, F. C. (2011). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, 19(4):264–279.

Han, T. A., Perret, C., and Powers, S. T. (2021). When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Systems Research*, 68:111–124.

Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K., and Nowak, M. A. (2017). Memory-n strategies of direct reciprocity. *Proceedings of the National Academy of Sciences*, 114(18):4715–4720.

Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2005). Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences of the United States of America*, 102:10797–10800.

Kurokawa, S. (2017). Persistence extends reciprocity. *Mathematical Biosciences*, 286:94–103.

Luhmann, N. (1979). *Trust and Power*. John Wiley & Sons, Chichester.

Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., and Lenaerts, T. (2015). Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific reports*, 5:10639.

Paiva, A., Santos, F. P., and Santos, F. C. (2018). Engineering pro-sociality with autonomous agents. In *Thirty-second AAAI conference on artificial intelligence*.

Pu, P. and Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556.

Qu, X., Cao, Z., Yang, X., and Han, T. A. (2019). How group cohesion promotes the emergence of cooperation in public goods game under conditional dissociation. *Journal of Artificial Societies and Social Simulation*, 22(3):5.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton University Press.

Traulsen, A., Nowak, M. A., and Pacheco, J. M. (2006). Stochastic dynamics of invasion and fixation. *Phys. Rev. E*, 74:11909.