HHAI 2026
M. Alimardani et al. (Eds.)

doi:10.3233/FAIA260514

Formalising Trust as Reduced Monitoring in Human–AI Interaction

Cedric PERRET [a], The Anh HAN [b], Elias FERNÁNDEZ DOMINGOS [c], Theodor CIMPEANU [d] and Simon T. POWERS [d,1]

[a] *University of Lausanne*
[b] *Teesside University*
[c] *Vrije Universiteit Brussel*
[d] *University of Stirling*

Abstract. There are many debates over what trust means in the context of human-AI interactions. However, this discussion has often lacked formal theories of trust that can provide testable predictions in different domains. To address this, we develop a game-theoretic formalisation of a prominent view that equates trust with reduced monitoring over time, capturing folk psychology intuition of "once I trust you then I don't have to keep checking what you're doing". This provides a behavioural measure of trust – how often one agent monitors to observe a partner's action. Using evolutionary game theory, we analyse the effects of trust on the frequency of cooperation between agents in canonical social dilemmas. We show that trust heuristics, which reduce monitoring once frequent cooperation has been observed, facilitate cooperation in two ways. First, when monitoring is costly, trust promotes cooperation in Prisoner's Dilemmas where the temptation to defect is high. Second, when agents can make errors, trust increases cooperation even in Stag-Hunt coordination interactions. Our results disentangle the effects of trust on cooperation, and provide a trust measure not limited to human-human interactions. We discuss the implications for designing auditing and monitoring systems in human-AI interactions, and for experimentally measuring trust in AI.

Keywords. trust, cooperation, evolutionary game theory, auditing

1. Introduction

Trust has proven difficult to define and operationalise in the context of interactions between humans and AI. Some argue that trust is best understood as a cognitive-affective state, and so by definition can only apply to interactions between humans. A "trust-like" relationship between humans and machines, including AI agents, is then said to be one of mere reliance [1]. However, other authors argue for taking a functionalist perspective on trust, where trust is understood in terms of the adaptive function that it performs for the trustor [2]. On these accounts, trust is an adaptive heuristic for regulating individual social interactions. For example, Luhmann [3] argues that the function of trust is to save an individual from having to constantly reason about the likelihood of every pos-

[1] Corresponding Author: Simon T. Powers, s.t.powers@stir.ac.uk

sible outcome from an interaction. In this sense, trust can be considered as a heuristic that helps individuals cope with the overwhelming information processing demands of everyday life. Viewed through this functionalist, evolutionary, lens there is nothing that says that trust can only be between humans. In accordance with the intentional stance of Dennett [4], trust is a heuristic that can apply to any type of social interaction, regardless of whether the interaction partner is human. Consequently, the concept of trust not only becomes meaningful in human-AI interactions, but provides a theoretical framework that can inform their design. By understanding when people will use a trust heuristic, we can ensure that human-AI interactions support people to use this appropriately. That is, to help people calibrate their trust to not over trust – risking exploitation – but also not under trust and risk losing out on potential benefits of human-AI interactions.

Recent work has built on this by operationalising trust behaviours as reduced monitoring of an interaction partner [5,6]. Crucially, trust-as-reduced-monitoring provides a *behavioural* measure of trust – how often an agent chooses not to monitor. Evolutionary game theory (EGT) models have demonstrated the adaptive benefit of this heuristic. Game theory is appropriate because users and developers of AI systems often have conflicting goals, leading to conflicts of interest in the interaction between users and AI agents [7]. For example, when a human interacts with a large language model (LLM), the human has the goal of obtaining accurate answers, while the goal of the LLM's developer is typically to maximise user engagement, which can manifest in sycophantic LLM behaviour [8].

2. Trust-as-reduced-monitoring is an adaptive heuristic in games of conflicting interests

EGT can analyse the (evolutionary) benefit of trust-as-reduced monitoring in pairwise repeated social dilemmas [6,9], defined by the following parameterised payoff matrix [10,11]:

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R & S \\ T & P \end{pmatrix}. \end{array}$$

Two individuals who cooperate receive the reward for mutual cooperation (R). If one individual cooperates while its partner defects, it receives the sucker's payoff (S), while the partner receives the temptation payoff (T). If both defect, both receive the punishment for mutual defection, (P). Varying S and T captures all possible pairwise symmetric social dilemmas [11,12] – Prisoner's Dilemma (PD): $T > 1 > 0 > S$; Snowdrift (SD): $T > 1 > S > 0$ and $T + S < 2$; Stag-Hunt (SH): $1 > T > 0 > S$.

The PD captures a strong conflict of interest, the SD moderate, and the SH a risky coordination situation (both do better by coordinating on cooperation, but only if the other does). A key strategy in repeated games is Tit-for-Tat (TFT): cooperate on the first round; thereafter monitor what the partner did, and mirror their action next time. Crucially, TFT always monitors. The cost of monitoring is usually assumed to be zero. However, in reality, it would typically be a cost $\varepsilon > 0$, representing time and energy. This paves the way for a trust heuristic (TUC): initially play TFT and monitor every round; once the partner has cooperated for θ rounds, switch to always cooperate and only

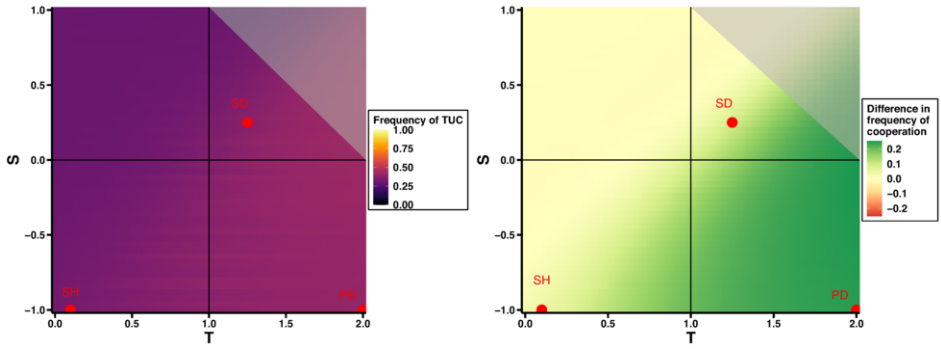

Figure 1. Frequency of the TUC strategy (left) and the impact of considering a trust heuristic on the frequency of cooperation (right). Cost of monitoring $\varepsilon = 0.25$. Adapted from [6].

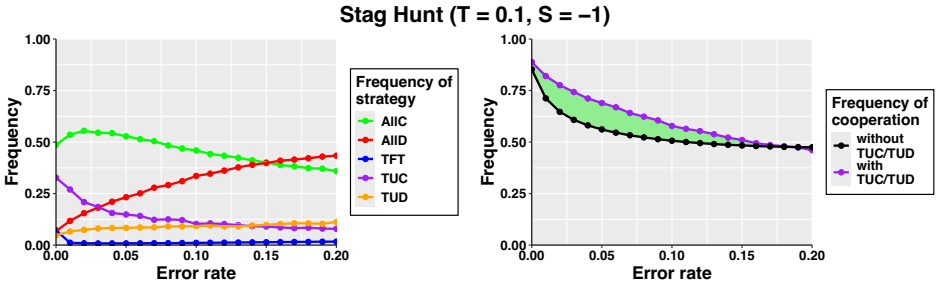

Figure 2. Trust evolves (left) and increases cooperation (right) if agents can make errors in the Stag-Hunt (adapted from [6])

monitor with probability p (and switch back to TFT if the partner is caught defecting when monitored). Figure 1 shows evolution in a finite population. TUC evolves under all social dilemmas (left), and promotes cooperation (right) strongly in the PD, less so in the SD, and with little effect in SH coordination games. However, if agents can make errors, then a trust heuristic is beneficial even in the SH (Figure 2).

3. Implications

Current experiments on trust in human-AI interactions lack a widely shared theoretical framework [13,14]. Trust-as-reduced-monitoring, analysed via game-theoretic methods, could provide one such framework. Trust can be objectively measured by how often a participant monitors, e.g. verifies the accuracy of output of an LLM, or watches a robotic partner. EGT models could be calibrated to this, with parameters such as the importance of the interaction, or the prior experience a participant has with the AI. These kinds of models can make testable predictions, allowing for the theoretically-informed design of trust in human-AI interactions. This is in contrast to much work that measures trust subjectively via surveys. Trust-as-reduced-monitoring can also inform the design of AI control strategies to prevent scheming – LLM agents that could appear aligned during testing, but then become misaligned once deployed, exploiting reduced oversight [15]. A TUD strategy (switch to defection once θ is reached) could capture these dynamics [9].

References

[1] Bolton ML. Trust is Not a Virtue: Why We Should Not Trust Trust. Ergonomics in Design. 2024 Oct;32(4):4-11.

[2] Lewis PR, Marsh S. What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. Cognitive Systems Research. 2022 Mar;72:33-49.

[3] Luhmann N. Trust and Power. Chichester: John Wiley & Sons; 1979.

[4] Dennett DC. Intentional Systems. The Journal of Philosophy. 1971;68(4):87-106.

[5] Ferrario A, Loi M, Viganò E. Trust does not need to be human: It is possible to trust medical AI. Journal of Medical Ethics. 2021 Jun;47(6):437-8.

[6] Perret C, Han TA, Fernández Domingos E, Cimpeanu T, Powers ST. Disentangling trust from cooperation: Evolution of trust as reduced monitoring in social dilemmas. Chaos, Solitons & Fractals. 2026 Jul;208:118130.

[7] Manzini A, Keeling G, Marchal N, McKee KR, Rieser V, Gabriel I. Should users trust advanced AI assistants? Justified trust as a function of competence and alignment. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. Rio de Janeiro Brazil: ACM; 2024. p. 1174-86.

[8] Sun Y, Wang T. Be friendly, not friends: How LLM sycophancy shapes user trust. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. CHI '26. New York, NY, USA: Association for Computing Machinery; 2026. p. 1-15.

[9] Han TA, Perret C, Powers ST. When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. Cognitive Systems Research. 2021 Aug;68:111-24.

[10] Wang Z, Kokubo S, Jusup M, Tanimoto J. Universal scaling for the dilemma strength in evolutionary games. Physics of Life Reviews. 2015;14:1-30.

[11] Santos FC, Pacheco JM, Lenaerts T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(9):3490-4.

[12] Macy MW, Flache A. Learning dynamics in social dilemmas. Proceedings of the National Academy of Sciences of the United States of America. 2002;99:7229-36.

[13] Leichtmann B, Nitsch V, Mara M. Crisis Ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement. Frontiers in Robotics and AI. 2022 Mar;9.

[14] Leichtmann B, Humer C, Hinterreiter A, Streit M, Mara M. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. Computers in Human Behavior. 2023 Feb;139:107539.

[15] Meinke A, Schoen B, Scheurer J, Balesni M, Shah R, Hobbhahn M. Frontier models are capable of in-context scheming. arXiv; 2025. ArXiv:2412.04984 [cs].