When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games

The Anh Han[a], Cedric Perret[b], Simon T. Powers[c,*]

[a] *Teesside University*
[b] *Teesside University*
[c] *Edinburgh Napier University*

Abstract

The actions of intelligent agents, such as chatbots, recommender systems, and virtual assistants are typically not fully transparent to the user. Consequently, users take the risk that such agents act in ways opposed to the users' preferences or goals. It is often argued that people use trust as a cognitive shortcut to reduce the complexity of such interactions. Here we formalise this by using the methods of evolutionary game theory to study the viability of trust-based strategies in repeated games. These 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 outcompete strategies that are always conditional, such as Tit-for-Tat, when the opportunity cost is non-negligible. We argue that

[⋆]This is the accepted version of the manuscript published in *Cognitive Systems Research*. The published version of record is available at https://doi.org/10.1016/j.cogsys.2021.02.003.

[*]Corresponding author

Email addresses: theanhhan.vn@gmail.com (The Anh Han), cedric.perret.research@gmail.com (Cedric Perret), S.Powers@napier.ac.uk (Simon T. Powers)

this cost is likely to be greater when the interaction is between people and intelligent agents, because of the reduced transparency of the agent. 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.

Keywords: Trust, evolutionary game theory, intelligent agents, cooperation, prisoner's dilemma, repeated games

---

1. Introduction

Artificial intelligence is undoubtedly becoming more integrated into our every day lives. While much attention has recently been paid to deep machine learning, intelligent agents that exhibit goal directed behaviour (Wooldridge, 2009) have also come of age. These range from purely software systems such as videogame characters or chatbots, through to cyberphysical systems such as smart fridges or autonomous vehicles. We are delegating more and more aspects of our daily lives to these agents, from the virtual sales agent that recommends products and services to us on an e-commerce website (Beldad et al., 2016), to the intelligent virtual assistant (e.g. Amazon Alexa, Apple Siri, Google Home) that plans our route to work and orders goods and services for us on command (Chung et al., 2017). But in all of these cases, the operation of the agent is not fully transparent to the end user. Although research in explainable AI is beginning to address these issues (Nunes and Jannach, 2017), it seems unlikely that a user will ever be able to get complete information about how and why the agent has taken a particular decision. Consequently, using such an agent, and accepting its recommendations, necessarily involves the user placing some degree of trust in the agent. In the broadest sense, trust is willingness to take risk under uncertainty (Luhmann, 1979). Here the risk is that the agent will act in a way opposed to our own goals, and the uncertainty comes from us lacking complete information about the behaviour of the agent to be able to ascertain this.

For example, consider again the virtual sales agent operating on the website of an e-commerce company (Chattaraman et al., 2012), which sells products to customers based on the information that it learns about the customer through a chat dialogue, i.e. an agent-based recommender system (Pu and Chen, 2007; Yoo et al., 2012; Jugovac and Jannach, 2017). When a customer interacts with this virtual sales agent it does not have complete information about why product A from company X is being recommended as opposed to product B from company Y (Grabner-Kraeuter, 2002). Thus, if the customer is going to use the virtual sales agent, they must take some degree of risk, for example, that the virtual sales agent recommends more expensive products, or those from manufacturers that the seller has a preferential relationship with, or does not provide full information about the quality of the product. Without a full understanding of the virtual sales agent's source code, the specifications of the alternative products, and the relationships between the sellers and manufacturers (Akerlof, 1970; Mahadevan, 2000; Lewis, 2011), some degree of risk and hence trust must be involved (Luhmann, 1979; Grabner-Kraeuter, 2002; Kumar et al., 2020). Similarly, when a virtual assistant gives us directions, we do not have complete information either about the route planning algorithm that it is using, or about relevant environmental conditions such as traffic levels. Again, this means that the use of such systems necessarily involves some degree of risk and hence trust.

This raises the question: how will people behave when interacting with these kinds of intelligent agents? How will they handle the complexity of the interaction? Ultimately, this question will need to be answered by empirical work. However, to guide the empirical work it is necessary to generate hypotheses about how we expect people to behave. Because intelligent agents exhibit goal directed behaviour, and their goals (as programmed by their designers) may be in conflict with the goals of their users, evolutionary game theory (EGT) (Maynard Smith, 1982; Sigmund, 2010) provides a suitable formal framework for modelling the strategic interaction and understanding behavioural dynamics (Shoham, 2008). This is because not only is the interaction strategic, but there is empirical evidence that people use a standard set of social scripts whether they are interacting with a person or a machine in a particular social situation (Nass and Moon, 2000). This suggests that predictions from game theoretical studies about human behaviour in traditional (e-)commerce, for example (e.g. Laaksonen et al. 2009; Dahlstrom et al. 2014), can also be useful when the interaction is between a human and an intelligent agent representing another entity (individual, firm, organisation), rather than with that entity directly.

In light of this, we propose that the types of interaction discussed above can be modelled as repeated games between the user and the agent (acting to fulfil the goals of its designer). Moreover, in important cases the actions available to the agent and the user correspond to "cooperate" and "defect".

Cooperation between players represents both the user and agent behaving honestly, reliably and transparently with each other. For example, cooperation would be a virtual sales agent selling products that match the preferences that the user has revealed in the conversation, while defection might correspond to trying to upsell products or warranties. On the side of the user, cooperation could represent continued use of the agent, which benefits the seller by reducing their opportunity costs of answering customer enquiries themselves. Defection would then represent refusing to use the agent and instead speaking directly to a human sales advisor.

The folk theorem of repeated games tells us that the key to cooperative outcomes, which benefit both sides, is sufficient information for the players to be able to condition their actions on the past actions of the other player(s) (Fudenberg and Maskin, 1986). This allows for reciprocal strategies, e.g. cooperate if the other player cooperated in the previous interaction, as exemplified by the Tit-for-Tat strategy (Axelrod, 1984). However, the use of reciprocal strategies necessarily carries an opportunity cost. Part of this comes from devoting cognitive resources to remembering a history of past actions, and processing this when deciding how to act. But in addition to this, reciprocal strategies also involve *verifying* whether the observed action of the other player actually was cooperative or not. In traditional face-to-face interactions between humans verifying whether the other player cooperated might involve, for example, checking the quality and specification of goods that have been purchased, or that the correct amount of change has been given. However, these costs are usually assumed to be low compared to the benefit and cost of cooperation (Ho, 1996; Imhof et al., 2005a; Han, 2013), and are mostly omitted in (evolutionary) game theoretic models (McNally et al., 2012; Han et al., 2013b; Martinez-Vaquero et al., 2015; Garcia and van Veelen, 2018; Hilbe et al., 2017; Glynatsi and Knight, 2020; Han et al., 2020). But the transition to interactions over the internet increases these costs (Grabner-Kraeuter, 2002), since the increased separation in space and time over the course of the interaction makes verifying the action of the other player more costly. The move to interacting with intelligent agents increases these costs even more, since the interaction becomes less transparent to the user, and artificial agents have limited capacity to explain their action compared to humans. This issue becomes even more relevant when considering hybrid societies of humans and intelligent agents (Paiva et al., 2018; Santos et al., 2019).

It is often argued that humans use trust as a cognitive shortcut, to reduce the complexity of the interaction that they need to reason about (Luhmann, 1979; Grabner-Kraeuter, 2002; Petruzzi et al., 2014). In this paper, we formalise this in EGT by introducing trust-based strategies in repeated games, and study their evolutionary viability when competing with other strategies in repeated games, in a similar fashion to Imhof et al. (2005a). Unlike traditional Tit-for-Tat, trust-based strategies only check a co-player's actions occasionally after a trust threshold has been reached, i.e. after their co-player has cooperated for a certain number of rounds. By doing so, they reduce the opportunity cost of verifying the action of their co-player every round. We demonstrate that trust-based strategies can be more successful than Tit-for-Tat when the opportunity cost of using a conditional strategy is non-negligible. Moreover, one may ask under what kinds of interaction or business at hand are trust-based strategies more likely to be used by the parties involved? For instance, will users trust in a chatbot to handle highly important interactions such as a multi-million dollar transaction? We show that trust-based strategies are most successful when the interaction is of intermediate importance, and the interaction is repeated over many rounds. These results provide game theoretic support for the theory that humans use trust to reduce the complexity of interactions, and suggest that people are likely to behave even more in this manner when interactions are with intelligent agents, since the opportunity costs of verifying the actions of intelligent agents are likely to be greater.

2. Models and Methods

2.1. Models

We consider a population of constant size N. At each time step or generation, a random pair of players are chosen to play with each other.

2.2. Interaction between Individuals

Interactions are modelled as a symmetric two-player prisoner's dilemma game, defined by the following payoff matrix (for row player)

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

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker's payoff S, whereas the defecting player gains the temptation to defect, T. Mutual cooperation (resp., defection) yields the reward R (resp., punishment P) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise (Macy and Flache, 2002; Santos et al., 2006). Namely, in this work we are concerned with the prisoner's dilemma (PD), where $T > R > P > S$. In a single round, it is always best to defect, but cooperation may be rewarded if the game is repeated. In repeated PD, it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ($2R > T + S$); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation. For convenience and a clear representation of results, we later mostly use the Donation game (Sigmund, 2010)—a famous special case of the PD—where $T = b,\ R = b - c,\ P = 0,\ S = -c$, satisfying that $b > c > 0$, where b and c stand respectively for "benefit" and "cost" (of cooperation).

In addition, in order to understand how the duration of the interaction or business at hand impacts the evolutionary viability of trust-based strategies in relation to others, we model how important or beneficial an interaction is using parameter $\gamma > 0$ (Han et al., 2013a). Hence, the payoff matrix becomes

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

In a population of N individuals interacting via a repeated (or iterated) PD, whenever two specific strategies are present in the population, say **A** and **B**, the fitness of an individual with a strategy **A** in a population with k **A**s and $(N - k)$ **B**s can be written as

$$\Pi_A(k) = \frac{1}{r(N-1)} \sum_{j=1}^{r} [(k-1)\pi_{A,A}(j) + (N-k)\pi_{A,B}(j)], \qquad (1)$$

where $\pi_{A,A}(j)$ ($\pi_{A,B}(j)$) stands for the payoff obtained from a round j as a result of their mutual behavior of an **A** strategist in an interaction with a **A** (**B**) strategist (as specified by the payoff matrix above), and r is the total number of rounds of the PD. As usual, instead of considering a fixed number of rounds, upon completion of each round, there is a probability w that yet another round of the game will take place, resulting in an average number of $r = (1 - w)^{-1}$ rounds per interaction (Sigmund, 2010). In the following, all values of Π will be computed analytically.

2.3. Strategies in IPD and the opportunity cost

The repeated (or iterated) PD is usually known as a story of tit-for-tat (TFT), which won both Axelrod's tournaments (Axelrod, 1984; Axelrod and Hamilton, 1981). *TFT* starts by cooperating, and does whatever the opponent did in the previous round. It will cooperate if the opponent cooperated, and will defect if the opponent defected.

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) and moreover, a cost of revealing the actual actions of co-players (cf. Introduction). The latter cost is usually ignored in previous works of IPD, but it can be non-trivial and thus significantly influence the nature of interactions. For instance, this cost is crucial to be considered in the context of human-machine interactions. For example, it might be quite costly and time consuming to check if one was charged the right amount when pay online/by Card/on ATM/ and whether the quality of the coffee produced by your coffee machine is reducing (and to what extent). This cost is even greater when interacting with intelligent agents whose operation and hence goals are less transparent, and which might, for example, be designed to hide pertinent information from users.

Trust-based strategies

We consider a new trust-based strategy that is capable of switching off the costly deliberation process when it trusts its co-players enough [1]. Namely, 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 defective moves from the co-player so far in the IPD—reaches a certain threshold, denoted by θ, it will play C unconditionally. We denote this strategy by TUC. TUC is illustrated in the Figure 1 representing one game between TUC and TFT.

[1]Our modelling approach is in accordance with the definition of trust often adopted in various multi-agent research, e.g. (Dasgupta, 2000; Ramchurn et al., 2004). That is, trust is a belief an agent has that the other party will do what it says it will (being honest and reliable) or reciprocate (being reciprocal for the common good of both), given an opportunity to defect to get higher payoffs.

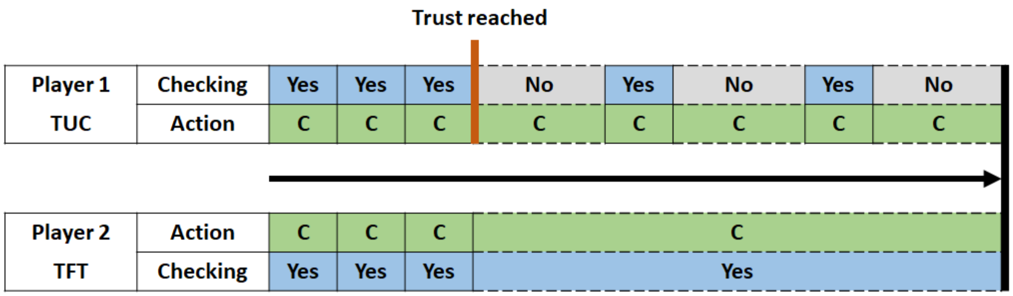

Figure 1: Diagram representing repeated interactions between a trust-based cooperator TUC and a tit-for-tat TFT. First, both strategies cooperate and check other player's action. After θ rounds (here $\theta = 3$), trust is reached and TUC now checks the action of TFT occasionally with a probability p. Because TFT continues to cooperate, TUC continues to trust and to cooperate.

Given the possibility of being exploited, but still to avoid costly deliberation, we assume that TUC will check, with a probability p, the co-player's actions after switching off [2]. If the co-player is found out to defect, TUC will revert to its initial strategy, i.e. TFT. As a counterpart of TUC, we consider TUD that whenever the ongoing trust level reaches the threshold θ, switches to playing D unconditionally. TUD is illustrated in the Figure 2 representing one game between TUC and TUD.

The payoff matrix for the five strategies ALLC, ALLD, TFT, TUC and TUD, can be given as follows

$$
\begin{array}{l}
\\
\textbf{AllC}\\
\textbf{AllD}\\
\textbf{TFT}\\
\textbf{TUC}\\
\textbf{TUD}
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\textbf{AllC} & \textbf{AllD} & \textbf{TFT} & \textbf{TUC} & \textbf{TUD}
\end{array}\\
\begin{pmatrix}
R & S & R & R & \frac{\theta R+(r-\theta)S}{r}\\
T & P & \frac{T+(r-1)P}{r} & \frac{T+(r-1)P}{r} & \frac{T+(r-1)P}{r}\\
R-\varepsilon & \frac{S+(r-1)P}{r}-\varepsilon & R-\varepsilon & R-\varepsilon & \frac{\theta R+S+(r-\theta-1)P}{r}-\varepsilon\\
R-\frac{\theta\varepsilon}{r}-\frac{p(r-\theta)\varepsilon}{r} & \frac{S+(r-1)P}{r}-\varepsilon & R-\frac{\theta\varepsilon}{r}-\frac{p(r-\theta)\varepsilon}{r} & R-\frac{\theta\varepsilon}{r}-\frac{p(r-\theta)\varepsilon}{r} & \Pi_{TUC,TUD}\\
\frac{\theta R+(r-\theta)T-\theta\varepsilon}{r} & \frac{S+(r-1)P}{r}-\varepsilon & \frac{\theta R+T+(r-\theta-1)P-\theta\varepsilon}{r} & \Pi_{TUD,TUC} & \frac{\theta R+(r-\theta)P-\theta\varepsilon}{r}
\end{pmatrix}
\end{array}
\qquad (2)
$$

For clarity, we write the payoff of TUC against TUD and TUD against

[2]We assume that, given a sufficient cost of checking, TUC can always correctly find out the co-player's actions.

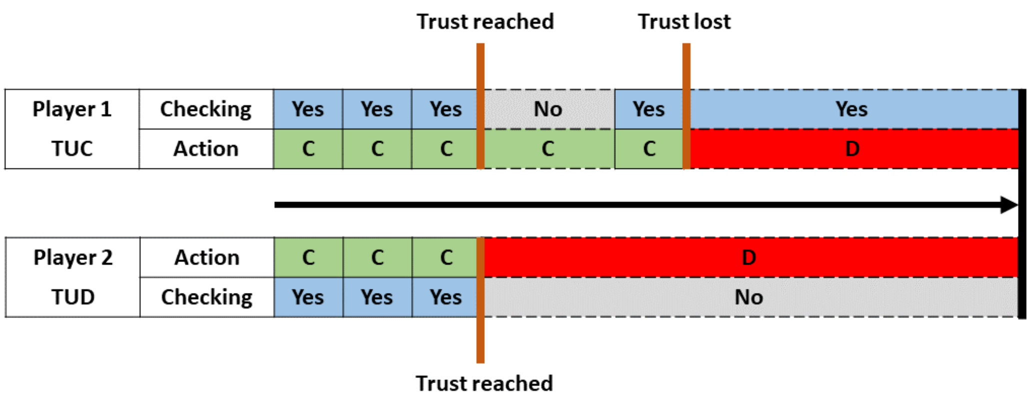

Figure 2: Diagram representing repeated interactions between a trust-based cooperator TUC and a trust-based defector TUD. First, both strategies cooperate and check the other player's action. After θ rounds (here $\theta = 3$), trust is reached for both strategies. TUC now cooperates and TUD defects. This continues until TUC checks and realises that TUD defects. After that, TUC looses trust, plays as a TFT and defects.

TUC separately. Namely, the payoff of TUC against TUD is given by

$$\begin{aligned}
&\Pi_{TUC,TUD} = \\
&\frac{1}{r}\left(\theta R - \theta\varepsilon + S + p(r'-1)(P-\epsilon) + (1-p)(S + p(r'-2)(P-\epsilon) + (1-p)[....])\right) \\
&= \frac{1}{r}\left(\theta R - \theta\varepsilon + S\sum_{i=0}^{r'-1}(1-p)^i + p(P-\epsilon)\sum_{i=0}^{r'-1}(r'-i-1)(1-p)^i\right) \\
&= \frac{\theta R - \theta\varepsilon}{r} + \frac{1}{r}\left(\frac{S(1-(1-p)^{r-\theta})}{p} + \frac{(P-\varepsilon)((1-p)^{r-\theta} + (r-\theta)p - 1))}{p}\right)
\end{aligned}$$

where $r' = r - \theta$. Similarly,

$$\Pi_{TUD,TUC} = \frac{\theta R - \theta\varepsilon}{r} + \frac{1}{r}\left(\frac{T(1-(1-p)^{r-\theta}}{p} + \frac{P((1-p)^{r-\theta} + (r-\theta)p - 1)}{p}\right)$$

The payoff formulas can be explained as follows. In the first θ rounds both TUC and TUD play C and keep checking, so they obtain in each round $R-\epsilon$. As trust is reached, from next rounds TUC will check only occasionally with probability p. For example, if in the next round TUC checks, it obtains S in that round and $P-\epsilon$ in the remaining rounds since it will play TFT. Otherwise, i.e. if TUC does not check in that round (with probability $1-p$), the process above is iterated for the payoffs calculation.

2.4. *Evolutionary dynamics in finite populations*

We resort in this paper to Evolutionary Game Theory methods for finite populations understanding evolutionary dynamics of trust-based behaviours, in relation to other strategies (Imhof et al., 2005b). In this context, agents' payoff represents their *fitness* or social *success*, and evolutionary dynamics is shaped by social learning (Sigmund, 2010), assuming that more successful agents will tend to be imitated more often by the others. We adapt here the pairwise comparison rule (Traulsen et al., 2006) to model social learning, where an agent A with fitness f_A adopts the strategy of another agent B with fitness f_B with probability given by the Fermi function,

$$P_{A \to B} = \left(1 + e^{-\beta(f_B - f_A)}\right)^{-1}.$$

The parameter β stands for the imitation strength or intensity of selection, i.e., how strongly agents base their decision to imitate on fitness comparison, where with $\beta = 0$, the imitation decision is random, while for increasing β, imitation becomes increasingly deterministic.

In the absence of behavioural exploration or mutations, end states of evolution inevitably are monomorphic. That is, whenever such a state is reached, it cannot be escaped via imitation. Thus, we further assume that, with some mutation probability, an agent can freely explore its behavioural space. In the limit of small mutation rates, the behavioural dynamics can be conveniently described by a Markov Chain, where each state represents a monomorphic population, whereas the transition probabilities are given by the fixation probability of a single mutant. The resulting Markov Chain has a stationary distribution, which characterises the average time the population spends in each of these monomorphic end states.

Suppose there exist at most two strategies in the population, say, k agents using strategy A ($0 \le k \le N$) and $(N-k)$ agents using strategies B. Let us denote by $\pi_{X,Y}$ the payoff an agent using strategy X obtained in an interaction with another individual using strategy Y (as given in the payoff matrix (2)). Hence, the (average) payoff of the agent that uses A and B can be written as follows, respectively,

$$\begin{aligned} \Pi_A(k) &= \frac{(k-1)\pi_{A,A} + (N-k)\pi_{A,B}}{N-1}, \\ \Pi_B(k) &= \frac{k\pi_{B,A} + (N-k-1)\pi_{B,B}}{N-1}, \end{aligned} \tag{3}$$

Now, the probability to change the number k of agents using strategy A by $\pm$ one in each time step can be written as (Traulsen et al., 2006)

$$T^{\pm}(k) = \frac{N-k}{N}\frac{k}{N}\left[1 + e^{\mp\beta[\Pi_A(k)-\Pi_B(k)]}\right]^{-1}. \quad (4)$$

The fixation probability of a single mutant with a strategy A in a population of $(N-1)$ agents using B is given by (Traulsen et al., 2006; Karlin and Taylor, 1975; Imhof et al., 2005b)

$$\rho_{B,A} = \left(1 + \sum_{i=1}^{N-1}\prod_{j=1}^{i}\frac{T^-(j)}{T^+(j)}\right)^{-1}. \quad (5)$$

When considering a set $\{1, ..., s\}$ of distinct strategies, these fixation probabilities determine the Markov Chain transition matrix $M = \{T_{ij}\}_{i,j=1}^{s}$, with $T_{ij,j\neq i} = \rho_{ji}/(s-1)$ and $T_{ii} = 1 - \sum_{j=1,j\neq i}^{s} T_{ij}$. The normalised eigenvector of the transposed of M associated with the eigenvalue 1 provides the above described stationary distribution (Imhof et al., 2005b), which defines the relative time the population spends while adopting each of the strategies.

3. Results

We use the model defined above to answer two questions. First, when will individuals use trust? To answer this question, we investigate under which conditions trust is an evolutionary viable strategy. We measure the success of trust by the frequency of the trust-based cooperative strategy (TUC), i.e the proportion of time the population is composed of only TUC. Second, when should there be trust? This is measured by how well the prevalence of trust-based behaviour enhances cooperation outcomes. We investigate the second question by looking under which conditions the presence of trust-based strategies (both TUC and TUD) increase the frequency of cooperation in the population.

The default values of the parameters, unless otherwise specified, are for the game payoffs $R = 1, S = -1, T = 2, P = 0$ (i.e. $b = 2$ and $c = 1$ in Donation game), the importance of the game $\gamma = 1$, the number of rounds $r = 50$, the population size $N = 100$, the intensity of selection $\beta = 0.1$, the trust threshold $\theta = 3$, the probability of checking its partner $p = 0.25$ and the opportunity cost $\epsilon = 0.25$. The analysis of the model has been implemented using the package EGTTools (Fernández Domingos, 2020).

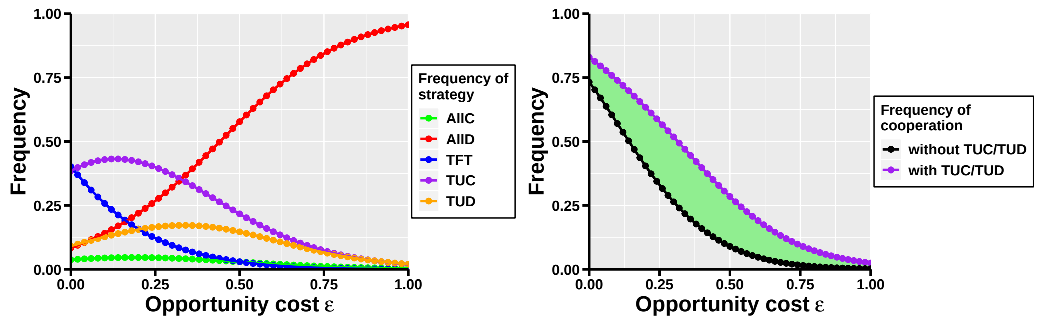

Figure 3: **Left: Frequency of strategies as a function of the opportunity cost ϵ. Right: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ.** The difference in frequency of cooperation between the two scenario is shaded in green when positive and red when negative.

3.1. Trust as a mechanism to reduce opportunity costs

The intuitive benefit of trusting something is to limit the cost of monitoring their actions in a long-term interaction, providing a shortcut in the decision making process. This is in line with several common definitions and theories of trust (Luhmann, 1979; Grabner-Kraeuter, 2002; Petruzzi et al., 2014). Thus, we explore first the effect of opportunity cost ϵ on the strategies employed by individuals and the resulting frequency of cooperation.

The left panel of Figure 3 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.

The right panel of Figure 3 shows that the presence of trust-based strategies increases the frequency of cooperation [3]. Importantly, this increase happens even when the opportunity cost ϵ is high ($\epsilon \approx c$), and not only when TUC is the most frequent, e.g. 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.

To conclude, trust-based cooperation is a particularly common strategy, in particular in interactions with moderate opportunity cost, and it promotes cooperation for a large range of opportunity costs.

3.2. Length of interactions and importance of the game

We now investigate (i) the importance of the game γ because this affects the *relative* cost of checking the other player and (ii) the number of rounds, because this affects the *relative* time that is required for trust to be established. The results are presented in Figure 4. First, we discuss the cases on the left column, where repeated interactions are short (expected number of rounds $r = 20$). The top left panel of Figure 4 shows that in such conditions, TUC is successful for medium values of the importance of the game parameter. TUC is also the most frequent strategy for a large range of the importance of the game parameter (note that the results presented are on a logarithmic scale). When the importance of the game is very low e.g. $\gamma = 0.1$, the most frequent strategy is AllD. In this condition, the opportunity cost is too high relative to the benefit provided by cooperation for either of the conditionally cooperative strategies, TUC or TFT, to thrive. When the importance of the game is very high, e.g. $\gamma = 1000$, TUC is almost never observed and TUD is, by far, the most frequent strategy. When the importance of game is high, defecting while the other player cooperates provides a huge benefit. AllD gets this benefit on the first round played with AllC, TFT and TUC. On the other hand, TUD obtains this advantageous payoff at least on the round after trust is established when interacting with TUC and TFT. This advantage by TUD is hard to recover through reciprocity if

[3] It is noteworthy that we compare the overall cooperation in our model to a baseline model that includes AllC, AllD and TFT. That is, there are three out of five cooperative strategies in our model, in comparison to two out of three in the baseline model. Thus, under neutrality (i.e. when all strategies have the same strategies, when $\beta \to 0$), it would be 60% cooperation vs 66.6% cooperation which is not in favour of our model.

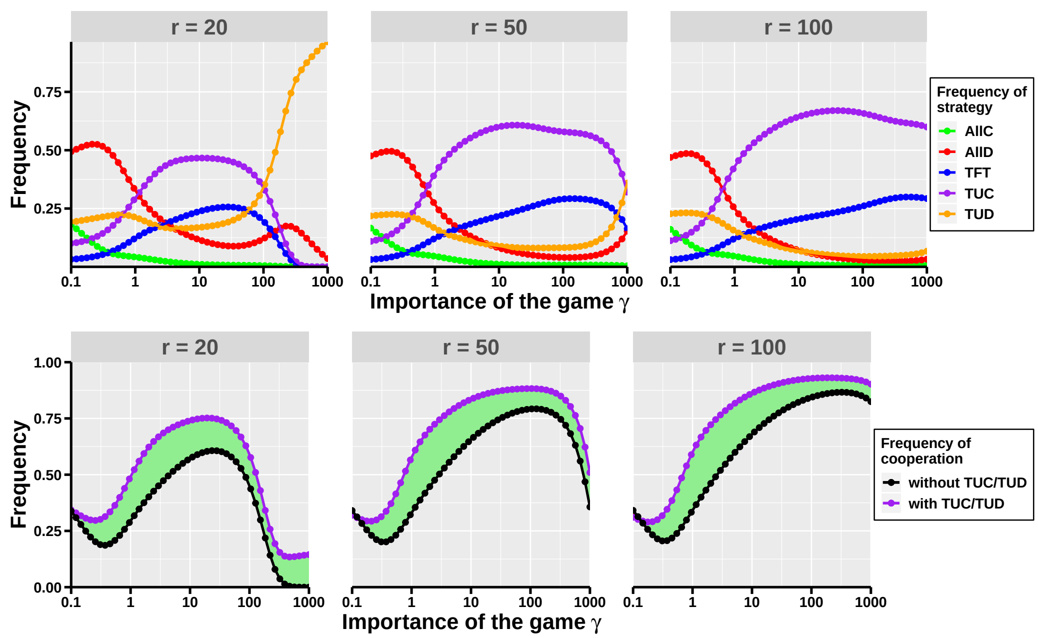

Figure 4: **Top: Frequency of strategies as a function of the number of rounds r and importance of the game γ (logarithmic scale); Bottom: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the number of rounds r and importance of the game γ (logarithmic scale).** For clarity, the difference in frequency of cooperation is shaded in green when positive and red when negative.

the number of rounds is not sufficiently high. It is noteworthy that β and the game importance parameter γ do not have the same effect. The former scales the whole fitness function, while the latter only scales the entries of the PD payoff matrix. Thus, β also scales the opportunity cost ϵ. A supplementary figure in appendix B shows that a higher β steepens the relationship between opportunity cost and frequencies or cooperation. In addition, a high intensity of selection leads to trust evolving only for a low opportunity cost. However, beside this expected effect, the qualitative results remain similar for a wide range of reasonable values of β (0.05 to 1).

This result is dependent of the length of the interactions. The top part of Figure 4 shows that a higher number of rounds r leads to (i) a higher frequency of TUC and (ii) the prevalence of TUC for a wider range of importance of game γ. TUC remains the most frequent strategy even when the importance of the game is high if the interactions are sufficiently long. This is because the high number of rounds where both individuals cooperate make up for the few initial rounds where TUC is exploited by TUD (and on a lesser extent, AllD).

The bottom part of Figure 4 shows that the presence of trust-based strategies increases the frequency of cooperation for all conditions examined. The highest frequency of cooperation is obtained for long interactions and high importance of the game. As shown by the similar shape of the curves, the higher frequency of cooperation appears to result from the high frequency of TUC. There is one notable exception. As shown in the bottom left figure (low r and high γ), the presence of trust based strategies also increases cooperation when TUC is not present. This is because TUD strategies cooperate more (for θ rounds) than AllD strategies which never cooperate.

In conclusion, trust is favoured for long-term interactions and can be observed on a wide range of importance of the game. The presence of trust-based strategies increases the frequency of cooperation for the whole set of parameter values studied.

3.3. Trustfulness and TUD

We have seen from the above results that trust-based cooperators are vulnerable to exploitation by TUD players, which are specifically tailored to take advantage of unconditional trust. This vulnerability was limited so far as we considered a rather careful truster with a $p = 0.25$. We now look at what is the effect of the probability of checking p on the success of TUC and the frequency of cooperation. For clarity, we present the result as a function

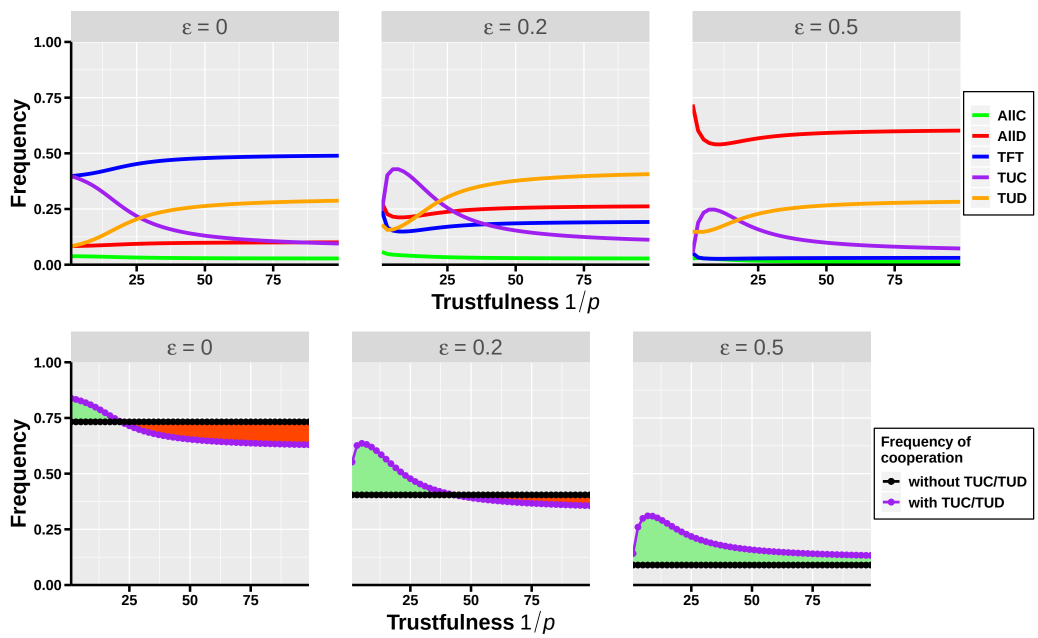

Figure 5: **Top: Frequency of strategies as a function of the opportunity cost ϵ and trustfulness $1/p$ (average number of rounds between checking event). Bottom: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ and trustfulness $1/p$ (average number of rounds between checking event)**. For clarity, the difference in frequency of cooperation is shaded in green when positive and red when negative.

of $1/p$, which approximates the trustfulness (which is larger for a smaller probability of checking) of TUC on the overall game, rather than p, which represents the carefulness of TUC on a single round.

The top part of Figure 5 first confirms that it is important for TUC's success to continue checking after trust is reached as TUC is much less frequent for a high value of trustfulness (i.e high $1/p$). If TUC is too trustful, the game is either dominated by TFT when the opportunity cost is small, by TUD when the opportunity cost is intermediate, and and AllD when the opportunity cost is high. There is an intermediate optimal trustfulness $1/p$ at which TUC is the most frequent strategy (except for zero opportunity costs where the lowest trustfulness and the highest probability of checking is the best strategy, which is equivalent to TFT). On the one hand, low trustfulness makes TUC less successful because TUC checks its partner often and so pays a higher opportunity cost. On the other hand, high trustfulness makes TUC vulnerable to exploitation by TUD for a longer number of rounds. The results show that there can be an optimal level of trust resulting from this trade-off.

The bottom part of Figure 5 shows that the presence of trust-based strategies increases the frequency of cooperation when the opportunity cost is moderate or high. This cooperation improvement is the highest for the optimal trustfulness at which TUC is very frequent. Again, the presence of trust-based strategies can lead to an increase in the frequency of cooperation even if they are not the most frequent strategy e.g. for very high opportunity costs ϵ. Unlike previously, the results also show that the presence of trust-based strategies can reduce the frequency of cooperation. This happens when the opportunity cost ϵ is low and the trustfulness $1/p$ is high. In these conditions, trustful and careless TUC players get exploited by TUD players, which increases the frequency of TUD, making cooperation a less viable option (evolutionarily). In the absence of trust-based strategies, TFT is careful enough to avoid this pitfall.

To conclude, unconditional trust is a viable strategy only in limited conditions and how much TUC relies on trust can have significant effect on the success of the strategy.

4. Discussion

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 these questions is crucial for providing suitable designs of mechanisms or infrastructures to facilitate human-intelligent machine interactions, e.g. in engineering pro-sociality in a hybrid society of humans and machines (Paiva et al., 2018). To this end, this paper provides a game theoretic analysis, where we formalised trust as a behavioural strategy and integrated it into an EGT model to study (i) its success in competition with non-trusting strategies and (ii) its effect on the level of cooperation.

Our results show first that trust is expected to be a pervasive cooperation enabling strategy. It is a frequent strategy for a large range of parameters, even in the presence of other strategies that are traditionally successful such as unconditional defection and Tit-for-Tat. Second, our results show that trust is a desirable mechanism in social systems because the presence of trust-based strategies increases the level of cooperation for a wide range of parameters. Finally, we show that trust-based cooperators are vulnerable to trust-based defectors, which are specialised to exploit them. However, our results also suggest that a minimum carefulness after trust is reached (low p) strongly limits this vulnerability.

Overall, our analysis shows that trust can emerge because it reduces the opportunity costs paid by individuals during interactions. It is a form of cognitive shortcut that, while exposing the player to some risks, can allow individuals to cooperate at lesser cost. If the pitfalls of trust have often been discussed, our results underlie the importance of taking into account both the benefits and the risk that the use of trust involves. Understanding the balance between these two is a first step to optimise the benefits of trust in intelligent machines while limiting the costs. On this line, further work could expand the model to look at different forms of trust based cooperation and defection strategies, how they co-evolve, and how exploitation of trust-based cooperators can be avoided. It is noteworthy that our additional (numerical) analyses have shown that all the observations described above (i.e. in Figures 3 - 5) are robust, e.g. for different values of the threshold number of rounds required for trust to be established (i.e. θ) (see Appendix). Moreover, note that in the current model we consider that TUC and TUD have the same θ, which is the worst case scenario for the evolution of TUC (and cooperation), since it represents the situation where TUD can perfectly recognise when TUC starts trusting a cooperative co-player and therefore becomes less vigilant of exploitation. More realistically, TUD might need to spend extra resources to gather information about TUC (e.g. providers learn about their customers' preferences and behaviours) to determine what is TUC's θ. On the other hand, TUC should not easily reveal or make available their information (that can be used to infer their θ), to better deal with TUD. Future work should address how these aspects might change the outcome of the evolutionary dynamics.

One of the most famous previous formalisations of trust is an experiment from behavioural economics called the trust game (Berg et al., 1995). This game consists of one individual receiving an endowment of money, of which it must choose a certain amount (which can be zero) to send to the other player. The amount sent to the other player is tripled by the experimenter (so that sending money represents an investment). The other player then decides what amount of this money (if any) to send back to the first player (so that there is risk in the first player sending money). While the Nash equilibrium is for the first player to send nothing to the other player, in experiments individuals usually deviate from this by sending a positive amount (Berg et al., 1995). The amount that the first player sends can be understood as measuring how much the first player trusts the second to reciprocate. However, in contrast to our formalisation, the trust game measures more a willingness to take risks blindly, as interactions are between anonymous individuals and are played only once. By contrast, we have considered repeated interactions between the same individuals, which has enabled us to look at the success of strategies that build up trust over time.

Within the context of the trust game, it has been shown recently that trust and trustworthiness cannot evolve in well-mixed and spatial networks with a homogeneous structure; they can evolve only in heterogeneous networks under highly advantageous conditions (Kumar et al., 2020). Moreover, within an overall grand challenge to understand the evolution of moral behaviour (beyond that of cooperation) (Capraro and Perc, 2018), the role of network topology in promoting honesty has been studied in (Capraro et al., 2019), extending works on honest signalling (Smith, 1991). Similarly, lying behaviour have recently been looked at within the context of well-mixed populations (Capraro et al., 2020).

In line with our approach, trust enabling strategies were previously considered in the context of repeated games (Han et al., 2011), where trust is built over time as a component of a larger decision making process, for prediction of opponents' behaviour. Trust was also considered for enabling cooperation in a one-shot prisoner's dilemma (Janssen, 2008; McNamara et al., 2009), but it was assumed that players can recognise how trustworthy a co-player is based on additional cues such as signalling. Our work differs from these approaches in that we consider trust as a cognitive shortcut to avoid deliberation and having to check the outcomes of previous interaction(s), thereby limiting the opportunity cost of conditional strategies. More generally, the role of an opportunity cost of monitoring the action of a co-player on the equilibrium level of cooperation has been studied in classic game theory models (for instance, see Lehrer and Solan (2018)). Using an evolutionary game theory approach, we complement this literature by showing that trust-based strategies are likely to emerge to deal with the cost of monitoring, even when players are short-sighted and only care about their immediate payoffs. We show conditions under which the presence of trust-based strategies can promote a high level of cooperation in comparison to the case where trust-based strategies are not available, e.g. where only classic reciprocal strategies such as TFT are possible.

In addition, trust has been used extensively in various computerised systems, such as in multi-agent open and distributed systems, to facilitate agents' interactions. Agents may have limited computational and storage capabilities that restrict their control over interactions, and trust is used to minimise uncertainty associated with the interactions, especially when agents inhabit in uncertain and constantly changing environments (Ramchurn et al., 2004; Falcone and Castelfranchi, 2001). This is the case for various applications including peer-to-peer computing, smart-grids, e-commerce, etc (Kamhoua et al., 2011; Ramchurn et al., 2004; Papadopoulou et al., 2001; Petruzzi et al., 2014; Brooks et al., 2020). These studies utilise trust for the purpose of regulating individual and collective behaviours, formalising different aspects of trust (such as reputation and belief) (Castelfranchi, 1997; Castelfranchi and Falcone, 2010). Our results and approach provide novel insights into the design of such computerised and hybrid systems as these require trust to ensure high levels of cooperation or efficient collaboration within a group or team of agents, including human-machine hybrid interactions. For instance, our results show that the importance of the business at hand (relative to the opportunity cost) needs to be taken into account to ensure a desired level of cooperation. Also, the system needs to be designed so that the opportunity cost of verifying the actions of an intelligent machine is sufficiently low (relative to the benefit and cost of the game) to enable a long-term trusting relationship with customers, e.g. making the activities transparent either directly to the user or to expert auditors that follow professional codes of ethics (Andras et al., 2018).

In the current work, since our goal was to explore the effect of trust-based strategies and a non-trivial opportunity cost in the context of human-intelligent machine interactions, we have based our analysis on a baseline model of IPD with three strategies (i.e. AllC, AllD and TFT), as described (Imhof et al., 2005a). There are other important strategies in this context, such as win-stay-lose-shift, grim and generous TFT, which are particularly relevant if errors in players' behavioural implementation is taken into account (Nowak and Sigmund, 1993; Imhof et al., 2007; Sigmund, 2010). For example, forgiveness is an important mechanism to deal with such errors, e.g. to resolve conflicts and avoid long-term retaliation in a long-term relationships. We will explore how trust-based strategies can be enhanced with forgiveness, as for instance errors might lead to difficulty in building initial trust and/or destroying built mutual trust not on purpose, and thus more forgiving trust-based strategies might better promote cooperation. On the other hand, these more forgiving strategies might be subject to more exploitation. In general, it is important to investigate a larger space of strategic behaviours in the context of IPD as it might influence the outcome of evolutionary dynamics.

An assumption made in our work is that the mutation or behavioural exploration is rare (Sigmund, 2010; Traulsen et al., 2006), allowing us to conveniently calculate the long-term frequencies (i.e. stationary distribution) of the strategies in the population. In reality, the mutation rate might be non-negligible and might have an effect on the evolutionary dynamics (Traulsen et al., 2009; Rand et al., 2013; Duong and Han, 2019). In general, larger mutation rates add more randomness to the system dynamics and might enable cooperation in situations where it is difficult to evolve otherwise, and vice versa (Hauert et al., 2007; Han et al., 2012; Rand et al., 2013; García and Traulsen, 2012). We aim to study the effect of mutation in future work.

In the current work we have focused on the prisoner's dilemma as it represents the hardest (pairwise) scenario for cooperation to emerge. Many other scenarios might be represented using other social dilemmas such as coordination or snowdrift games (Santos et al., 2006; Sigmund, 2010). Considering such games where it is easier for cooperation to emerge has the potential to open new windows of opportunity for long-term trust-based relationships to be established. Our future work will study how trust-based strategies (as we have modelled) evolve in the context of other social dilemmas. Moreover, given the importance of population structures in the emergence of trust and trustworthiness in the context of the trust game (Kumar et al., 2020), our future work will examine how different population structures influence the outcome of trustful behaviours in our context. Finally, we have assumed that agents always pay the cost of checking, but an alternative is that, they might choose not to check when it is too difficult or costly to do so (for example, checking if an AI development company complies with safety requirements in the development process within a competition to reach technology supremacy (Han et al., 2020)).

5. Conclusion

We have demonstrated in this paper that evolutionary game theory provides a valuable framework to study trust. Social interactions often result in complex dynamics with unexpected consequences, which a quantitative model is able to shed light on. Our model provides formal support for the theory that trust is a cognitive shortcut which people use to reduce the complexity of their interactions. The results of the model provide new insights into the questions of whether and when humans might trust intelligent machines, generating reasonable behavioural hypotheses that empirical studies can test.

Acknowledgements

We thank all members of the TIM 2019 Workshop, particularly, Lukas Esterle and Stephen Marsh, for useful discussion. TAH and CP are supported by Future of Life Institute (grant RFP2-154).

References

Akerlof, G. (1970). The market for ‘lemons’: Quality uncertainty and the market mechanism. *quarterly Journal of Economics*, 84(3):488–500.

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. (1984). *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.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142.

Brooks, N. A., Powers, S. T., and Borg, J. M. (2020). A mechanism to promote social behaviour in household load balancing. *Artificial Life Conference Proceedings*, 32:95–103.

Capraro, V. and Perc, M. (2018). Grand challenges in social physics: In pursuit of moral behavior. *Frontiers in Physics*, 6:107.

Capraro, V., Perc, M., and Vilone, D. (2019). The evolution of lying in well-mixed populations. *Journal of the Royal Society Interface*, 16(156):20190211.

Capraro, V., Perc, M., and Vilone, D. (2020). Lying on networks: The role of structure and topology in promoting honesty. *Physical Review E*, 101(3):032305.

Castelfranchi, C. (1997). Modeling social action for AI agents. *IJCAI International Joint Conference on Artificial Intelligence*, 2(c):1567–1576.

Castelfranchi, C. and Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons.

Chattaraman, V., Kwon, W.-S., and Gilbert, J. E. (2012). Virtual agents in retail web sites: Benefits of simulated social interaction for older users. *Computers in Human Behavior*, 28(6):2055–2066.

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

Dahlstrom, R., Nygaard, A., Kimasheva, M., and M. Ulvnes, A. (2014). How to recover trust in the banking industry? A game theory approach to empirical analyses of bank and corporate customer relationships. *International Journal of Bank Marketing*, 32(4):268–278. Publisher: Emerald Group Publishing Limited.

Dasgupta, P. (2000). Trust as a commodity. *Trust: Making and breaking cooperative relations*, 4:49–72.

Duong, M. H. and Han, T. A. (2019). On equilibrium properties of the replicator–mutator equation in deterministic and random games. *Dynamic Games and Applications*, pages 1–23.

Falcone, R. and Castelfranchi, C. (2001). Social trust: A cognitive approach. In *Trust and deception in virtual societies*, pages 55–90. Springer.

Fernández Domingos, E. (2020). Egttools: Toolbox for evolutionary game theory. https://github.com/Socrats/EGTTools.

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incompete information. *Econometrica*, 54:533–554.

García, J. and Traulsen, A. (2012). The structure of mutations and the evolution of cooperation. *PloS one*, 7(4):e35287.

Garcia, J. and van Veelen, M. (2018). No strategy can win in the repeated prisoner's dilemma: linking game theory and computer simulations. *Frontiers in Robotics and AI*, 5:102.

Glynatsi, N. and Knight, V. (2020). Using a theory of mind to find best responses to memory-one strategies. *Scientific Reports*, 17287:10.

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

Han, T. A. (2013). *Intention recognition, commitment and their roles in the evolution of cooperation: From Artificial intelligence techniques to evolutionary game theory models.* Springer SAPERE, vol. 9.

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., Pereira, L. M., and Santos, F. C. (2012). The emergence of commitments and cooperation. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'2012)*, pages 559–566. ACM.

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

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013b). Why is it so hard to say sorry? evolution of apology with commitments in the iterated prisoner's dilemma. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 177–183.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2020). To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research, pre-print available at arXiv:1907.12393.* In Press.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., and Sigmund, K. (2007). Via freedom to coercion: the emergence of costly punishment. *science*, 316(5833):1905–1907.

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.

Ho, T.-H. (1996). Finite automata play repeated prisoner's dilemma with information processing costs. *Journal of economic dynamics and control*, 20(1-3):173–207.

Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2005a). Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):10797–10800.

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

Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2007). Tit-for-tat or win-stay, lose-shift? *Journal of Theoretical Biology*, 247(3):574 – 580.

Janssen, M. A. (2008). Evolution of cooperation in a one-shot prisoner's dilemma based on recognition of trustworthy and untrustworthy agents. *Journal of Economic Behavior & Organization*, 65(3-4):458–471.

Jugovac, M. and Jannach, D. (2017). Interacting with Recommenders:Overview and research directions. *ACM Transactions on Interactive Intelligent Systems*, 7(3):10:1–10:46.

Kamhoua, C. A., Pissinou, N., and Makki, K. (2011). Game theoretic modeling and evolution of trust in autonomous multi-hop networks: Application to network security and privacy. In *2011 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Karlin, S. and Taylor, H. E. (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

Kumar, A., Capraro, V., and Perc, M. (2020). The evolution of trust and trustworthiness. *Journal of the Royal Society Interface*, 17(169):20200491.

Laaksonen, T., Jarimo, T., and Kulmala, H. I. (2009). Cooperative strategies in customer–supplier relationships: The role of interfirm trust. *International Journal of Production Economics*, 120(1):79–87.

Lehrer, E. and Solan, E. (2018). High frequency repeated games with costly monitoring. *Theoretical Economics*, 13(1):87–113.

Lewis, G. (2011). Asymmetric information, adverse selection and online disclosure: The case of eBay Motors. *American Economic Review*, 101(4):1535–1546.

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

Macy, M. W. and Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7229–7236.

Mahadevan, B. (2000). Business models for internet-based e-commerce: An anatomy. *California Management Review*, 42(4):55–69. Publisher: SAGE Publications Inc.

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.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK.

McNally, L., Brown, S. P., and Jackson, A. L. (2012). Cooperation and the evolution of intelligence. *Proceedings of the Royal Society B: Biological Sciences*, 279(1740):3027–3034.

McNamara, J. M., Stephens, P. A., Dall, S. R., and Houston, A. I. (2009). Evolution of trust and trustworthiness: social awareness favours personality differences. *Proceedings of the Royal Society B: Biological Sciences*, 276(1657):605–613.

Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Nowak, M. A. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in prisoner's dilemma. *Nature*, 364:56–58.

Nunes, I. and Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3):393–444.

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

Papadopoulou, P., Andreou, A., Kanellis, P., and Martakos, D. (2001). Trust and relationship building in electronic commerce. *Internet research*.

Petruzzi, P. E., Busquets, D., and Pitt, J. (2014). Experiments with social capital in multi-agent systems. In Dam, H. K., Pitt, J., Xu, Y., Governatori, G., and Ito, T., editors, *PRIMA 2014: Principles and Practice of Multi-Agent Systems*, Lecture Notes in Computer Science, pages 18–33, Cham. Springer International Publishing.

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

Ramchurn, S. D., Huynh, D., and Jennings, N. R. (2004). Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., and Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences*, 110(7):2581–2586.

Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences of the United States of America*, 103:3490–3494.

Santos, F. P., Pacheco, J. M., Paiva, A., and Santos, F. C. (2019). Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6146–6153.

Shoham, Y. (2008). Computer science and game theory. *Communications of the ACM*, 51(8):74–79.

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

Smith, J. M. (1991). Honest signalling: the philip sidney game. *Animal Behaviour*.

Traulsen, A., Hauert, C., De Silva, H., Nowak, M. A., and Sigmund, K. (2009). Exploration dynamics in evolutionary games. *Proceedings of the National Academy of Sciences*, 106(3):709–712.

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

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

Yoo, K.-H., Gretzel, U., and Zanker, M. (2012). *Persuasive Recommender Systems: Conceptual Background and Implications*. Springer Publishing Company, Incorporated, 1st edition.

Competing interests statement

The authors have not competing interests.

Author contributions

The Anh Han: Conceptualization, Methodology, Formal Analysis, Writing - Original draft preparation
Cedric Perret: Conceptualization, Methodology, Formal Analysis, Software, Writing - Original draft preparation
Simon T. Powers: Conceptualization, Methodology, Writing - Original draft preparation

Appendices

Appendix A: Results for different trust threshold θ

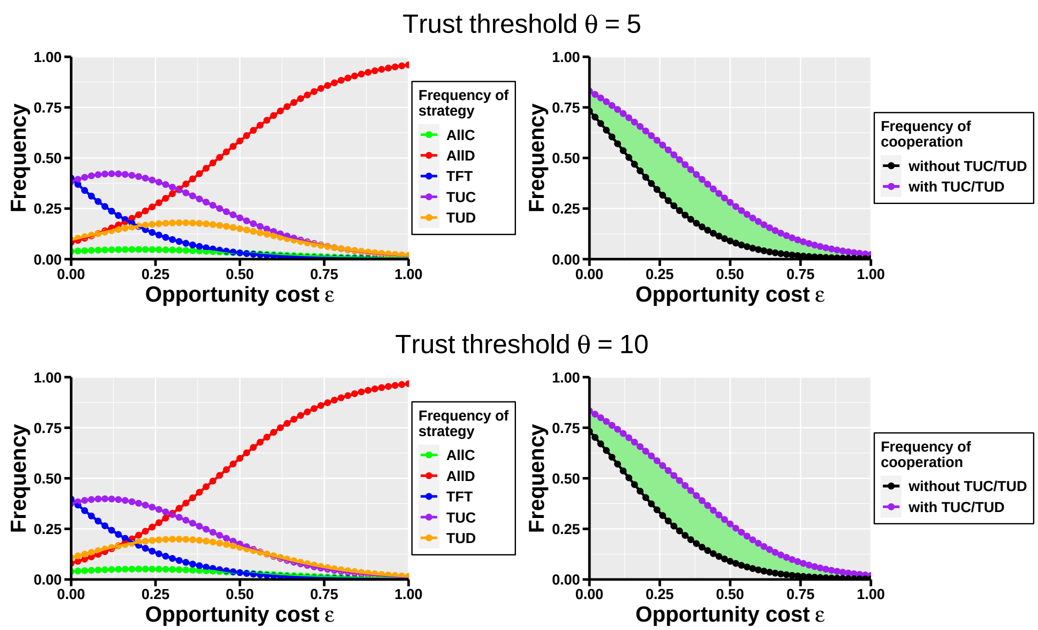

Figure 6: **Left: Frequency of strategies as a function of the opportunity cost ϵ. Right: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ.** The difference in frequency of cooperation between the two scenario is shaded in green when positive and red when negative. Each results are presented for different trust threshold $\theta = 5$ and $\theta = 10$. Parameters: $\beta = 0.1$, $N = 100$, $\gamma = 1$, $r = 50$, $p = 0.25$, $R = 1, S = -1, T = 2, P = 0$.

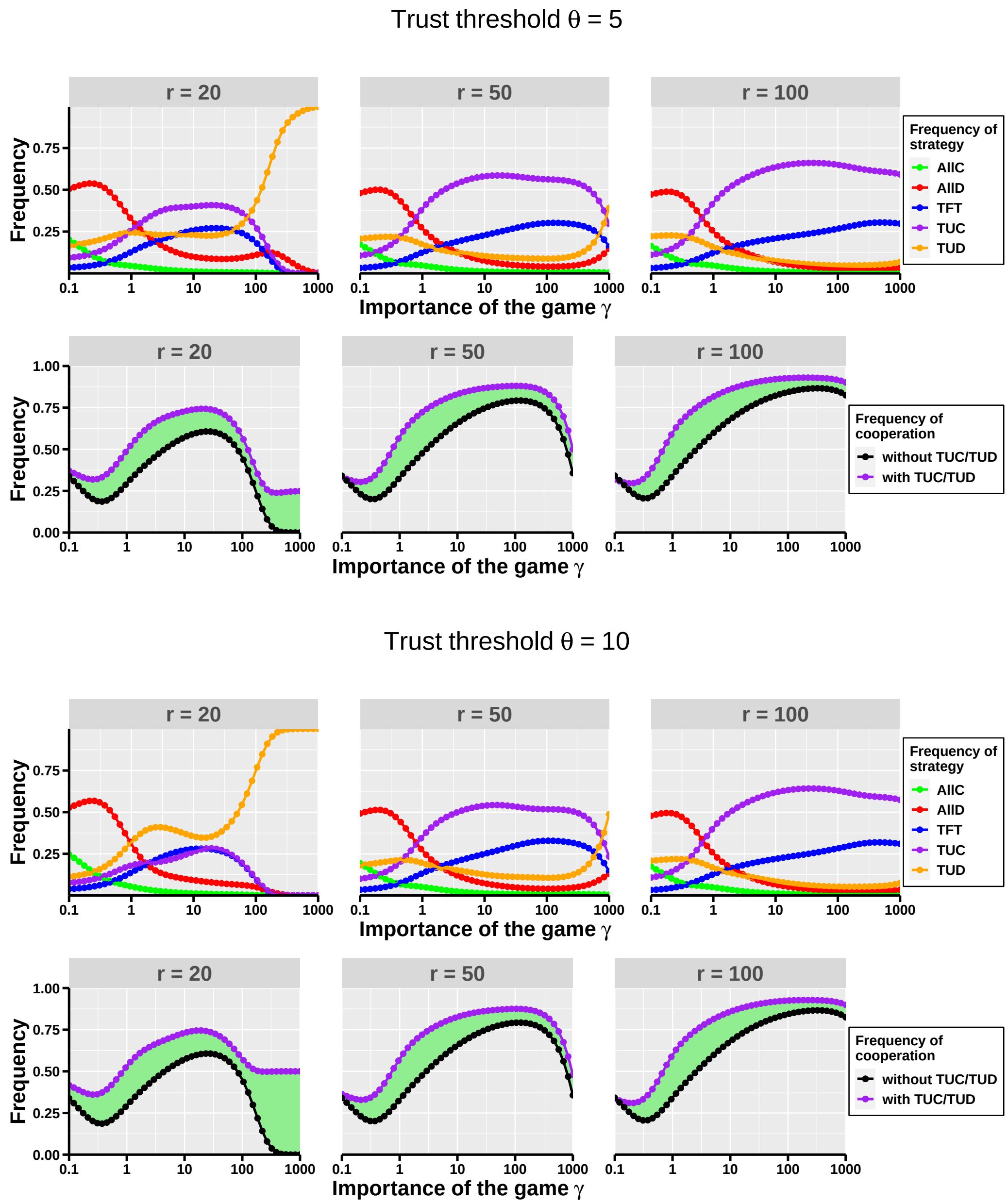

Figure 7: **Top: Frequency of strategies as a function of the number of rounds r and importance of the game γ (logarithmic scale); Bottom: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the number of rounds r and importance of the game γ (logarithmic scale).** For clarity, the difference in frequency of cooperation is shaded in green when positive and red when negative. Each results are presented for different trust threshold $\theta = 5$ and $\theta = 10$. Parameters: $\beta = 0.1$, $N = 100$, $p = 0.25$, $R = 1, S = -1, T = 2, P = 0$.

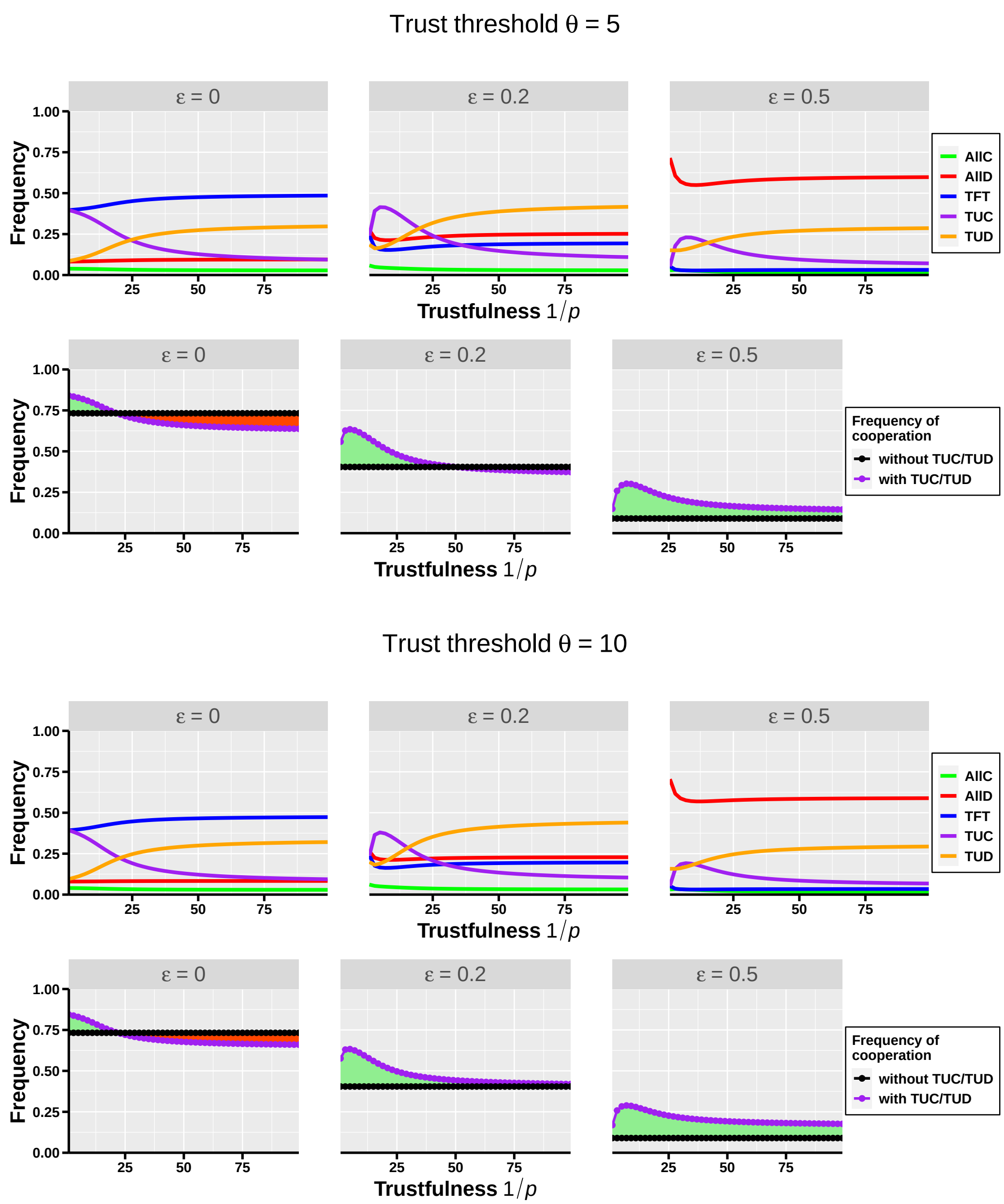

Figure 8: **Top: Frequency of strategies as a function of the opportunity cost ϵ and trustfulness $1/p$ (average number of rounds between checking event). Bottom: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ and trustfulness $1/p$ (average number of rounds between checking event)**. For clarity, the difference in frequency of cooperation is shaded in green when positive and red when negative. Each results are presented for different trust threshold $\theta = 5$ and $\theta = 10$. Parameters: $\beta = 0.1$, $N = 100$, $\gamma = 1$,$r = 50$, $R = 1, S = -1, T = 2, P = 0$.

Appendix B: Results for different intensity of selection β

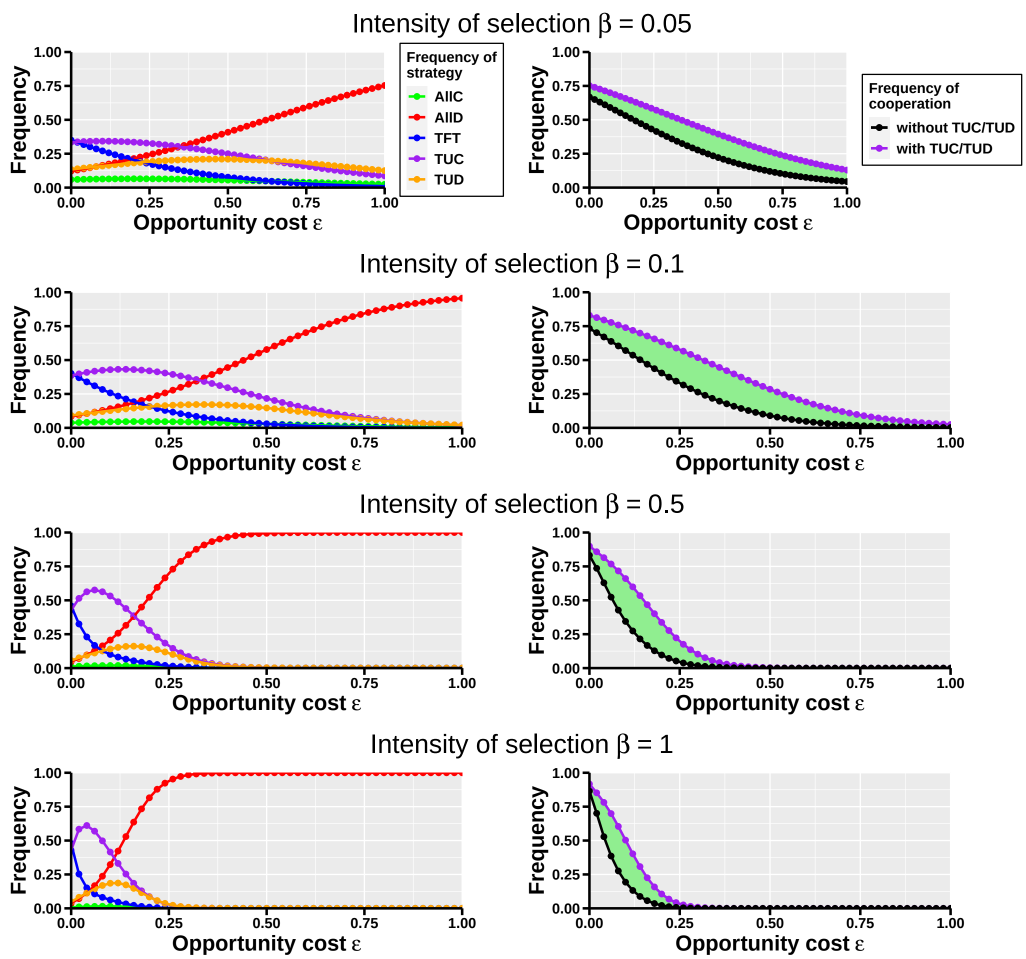

Figure 9: **Left: Frequency of strategies as a function of the opportunity cost ϵ. Right: Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ.** The difference in frequency of cooperation between the two scenario is shaded in green when positive and red when negative. Each results are presented for different intensity of selection, from top to bottom $\beta = 0.05$, $\beta = 0.1$ and $\beta = 0.5$. Parameters: $N = 100$, $\gamma = 1$, $r = 50$, $p = 0.25$, $R = 1, S = -1, T = 2, P = 0$.