

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 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

*Corresponding author

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

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

32 of risk, for example, that the virtual sales agent recommends more expen-
33 sive products, or those from manufacturers that the seller has a preferential
34 relationship with, or does not provide full information about the quality of
35 the product. Without a full understanding of the virtual sales agent’s source
36 code, the specifications of the alternative products, and the relationships be-
37 tween the sellers and manufacturers (Akerlof, 1970; Mahadevan, 2000; Lewis,
38 2011), some degree of risk and hence trust must be involved (Luhmann, 1979;
39 Grabner-Kraeuter, 2002; Kumar et al., 2020). Similarly, when a virtual as-
40 sistant gives us directions, we do not have complete information either about
41 the route planning algorithm that it is using, or about relevant environmen-
42 tal conditions such as traffic levels. Again, this means that the use of such
43 systems necessarily involves some degree of risk and hence trust.

44 This raises the question: how will people behave when interacting with
45 these kinds of intelligent agents? How will they handle the complexity of the
46 interaction? Ultimately, this question will need to be answered by empirical
47 work. However, to guide the empirical work it is necessary to generate hy-
48 potheses about how we expect people to behave. Because intelligent agents
49 exhibit goal directed behaviour, and their goals (as programmed by their
50 designers) may be in conflict with the goals of their users, evolutionary game
51 theory (EGT) (Maynard Smith, 1982; Sigmund, 2010) provides a suitable
52 formal framework for modelling the strategic interaction and understanding
53 behavioural dynamics (Shoham, 2008). This is because not only is the inter-
54 action strategic, but there is empirical evidence that people use a standard
55 set of social scripts whether they are interacting with a person or a machine in
56 a particular social situation (Nass and Moon, 2000). This suggests that pre-
57 dictions from game theoretical studies about human behaviour in traditional
58 (e-)commerce, for example (e.g. Laaksonen et al. 2009; Dahlstrom et al.
59 2014), can also be useful when the interaction is between a human and an
60 intelligent agent representing another entity (individual, firm, organisation),
61 rather than with that entity directly.

62 In light of this, we propose that the types of interaction discussed above
63 can be modelled as repeated games between the user and the agent (acting
64 to fulfil the goals of its designer). Moreover, in important cases the actions
65 available to the agent and the user correspond to “cooperate” and “defect”.
66 Cooperation between players represents both the user and agent behaving
67 honestly, reliably and transparently with each other. For example, coopera-
68 tion would be a virtual sales agent selling products that match the preferences
69 that the user has revealed in the conversation, while defection might corre-

70 spond to trying to upsell products or warranties. On the side of the user,
71 cooperation could represent continued use of the agent, which benefits the
72 seller by reducing their opportunity costs of answering customer enquiries
73 themselves. Defection would then represent refusing to use the agent and
74 instead speaking directly to a human sales advisor.

75 The folk theorem of repeated games tells us that the key to cooperative
76 outcomes, which benefit both sides, is sufficient information for the play-
77 ers to be able to condition their actions on the past actions of the other
78 player(s) (Fudenberg and Maskin, 1986). This allows for reciprocal strate-
79 gies, e.g. cooperate if the other player cooperated in the previous interaction,
80 as exemplified by the Tit-for-Tat strategy (Axelrod, 1984). However, the use
81 of reciprocal strategies necessarily carries an opportunity cost. Part of this
82 comes from devoting cognitive resources to remembering a history of past
83 actions, and processing this when deciding how to act. But in addition to
84 this, reciprocal strategies also involve *verifying* whether the observed action
85 of the other player actually was cooperative or not. In traditional face-to-face
86 interactions between humans verifying whether the other player cooperated
87 might involve, for example, checking the quality and specification of goods
88 that have been purchased, or that the correct amount of change has been
89 given. However, these costs are usually assumed to be low compared to the
90 benefit and cost of cooperation (Ho, 1996; Imhof et al., 2005a; Han, 2013),
91 and are mostly omitted in (evolutionary) game theoretic models (McNally
92 et al., 2012; Han et al., 2013b; Martinez-Vaquero et al., 2015; Garcia and
93 van Veelen, 2018; Hilbe et al., 2017; Glynatsi and Knight, 2020; Han et al.,
94 2020). But the transition to interactions over the internet increases these
95 costs (Grabner-Kraeuter, 2002), since the increased separation in space and
96 time over the course of the interaction makes verifying the action of the other
97 player more costly. The move to interacting with intelligent agents increases
98 these costs even more, since the interaction becomes less transparent to the
99 user, and artificial agents have limited capacity to explain their action com-
100 pared to humans. This issue becomes even more relevant when considering
101 hybrid societies of humans and intelligent agents (Paiva et al., 2018; Santos
102 et al., 2019).

103 It is often argued that humans use trust as a cognitive shortcut, to reduce
104 the complexity of the interaction that they need to reason about (Luhmann,
105 1979; Grabner-Kraeuter, 2002; Petruzzi et al., 2014). In this paper, we for-
106 malise this in EGT by introducing trust-based strategies in repeated games,
107 and study their evolutionary viability when competing with other strategies

108 in repeated games, in a similar fashion to [Imhof et al. \(2005a\)](#). Unlike tra-
 109 ditional Tit-for-Tat, trust-based strategies only check a co-player’s actions
 110 occasionally after a trust threshold has been reached, i.e. after their co-
 111 player has cooperated for a certain number of rounds. By doing so, they
 112 reduce the opportunity cost of verifying the action of their co-player every
 113 round. We demonstrate that trust-based strategies can be more successful
 114 than Tit-for-Tat when the opportunity cost of using a conditional strategy
 115 is non-negligible. Moreover, one may ask under what kinds of interaction
 116 or business at hand are trust-based strategies more likely to be used by the
 117 parties involved? For instance, will users trust in a chatbot to handle highly
 118 important interactions such as a multi-million dollars transition? We show
 119 that trust-based strategies are most successful when the interaction is of in-
 120 termediate importance, and the interaction is repeated over many rounds.
 121 These results provide game theoretic support for the theory that humans use
 122 trust to reduce the complexity of interactions, and suggest that people are
 123 likely to behave even more in this manner when interactions are with intelli-
 124 gent agents, since the opportunity costs of verifying the actions of intelligent
 125 agents are likely to be greater.

126 2. Models and Methods

127 2.1. Models

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

130 2.2. Interaction between Individuals

131 Interactions are modelled as a symmetric two-player Prisoner’s Dilemma
 132 game, defined by the following payoff matrix (for row player)

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

133 A player who chooses to cooperate (C) with someone who defects (D) receives
 134 the sucker’s payoff S , whereas the defecting player gains the temptation to
 135 defect, T . Mutual cooperation (resp., defection) yields the reward R (resp.,
 136 punishment P) for both players. Depending on the ordering of these four
 137 payoffs, different social dilemmas arise ([Macy and Flache, 2002](#); [Santos et al.](#),

138 2006). Namely, in this work we are concerned with the Prisoner’s Dilemma
 139 (PD), where $T > R > P > S$. In a single round, it is always best to defect,
 140 but cooperation may be rewarded if the game is repeated. In repeated PD, it
 141 is also required that mutual cooperation is preferred over an equal probability
 142 of unilateral cooperation and defection ($2R > T + S$); otherwise alternating
 143 between cooperation and defection would lead to a higher payoff than mutual
 144 cooperation. For convenience and a clear representation of results, we later
 145 mostly use the Donation game (Sigmund, 2010)—a famous special case of
 146 the PD—where $T = b$, $R = b - c$, $P = 0$, $S = -c$, satisfying that $b > c > 0$,
 147 where b and c stand respectively for “benefit” and “cost” (of cooperation).

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

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

152 In a population of N individuals interacting via a repeated (or iterated)
 153 Prisoner’s dilemma, whenever two specific strategies are present in the pop-
 154 ulation, say \mathbf{A} and \mathbf{B} , the fitness of an individual with a strategy \mathbf{A} in a
 155 population with k \mathbf{A} s and $(N - k)$ \mathbf{B} s can be written as

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

156 where $\pi_{A,A}(j)$ ($\pi_{A,B}(j)$) stands for the payoff obtained from a round j as
 157 a result of their mutual behavior of an \mathbf{A} strategist in an interaction with
 158 a \mathbf{A} (\mathbf{B}) strategist (as specified by the payoff matrix above), and r is the
 159 total number of rounds of the Prisoner’s dilemma. As usual, instead of
 160 considering a fixed number of rounds, upon completion of each round, there is
 161 a probability w that yet another round of the game will take place, resulting
 162 in an average number of $r = (1 - w)^{-1}$ rounds per interaction (Sigmund,
 163 2010). In the following, all values of Π will be computed analytically.

164 2.3. Strategies in IPD and the opportunity cost

165 The repeated (or iterated) PD is usually known as a story of tit-for-tat
 166 (TFT), which won both Axelrod’s tournaments (Axelrod, 1984; Axelrod and

167 [Hamilton, 1981](#)). *TFT* starts by cooperating, and does whatever the oppo-
168 nent did in the previous round. It will cooperate if the opponent cooperated,
169 and will defect if the opponent defected.

170 As a conditional strategy, *TFT* incurs an additional opportunity cost,
171 denoted by ϵ , compared to the unconditional strategies, namely, *ALLC* (al-
172 ways cooperate) and *ALLD* (always defect). This cost involves a cognitive
173 cost (to memorise previous interaction outcomes with co-players and make a
174 decision based on them) and moreover, a cost of revealing the actual actions
175 of co-players (cf. Introduction). The latter cost is usually ignored in previous
176 works of *IPD*, but it can be non-trivial and thus significantly influence the
177 nature of interactions. For instance, this cost is crucial to be considered in
178 the context of human-machine interactions. For example, it might be quite
179 costly and time consuming to check if one was charged the right amount
180 when pay online/by Card/on ATM/ and whether the quality of the coffee
181 produced by your coffee machine is reducing (and to what extent). This cost
182 is even greater when interacting with intelligent agents whose operation and
183 hence goals are less transparent, and which might, for example, be designed
184 to hide pertinent information from users.

185 *Trust-based strategies*

186 We consider a new trust-based strategy that is capable of switching off the
187 costly deliberation process when it trusts its co-players enough ¹. Namely,
188 this strategy starts an *IPD* interaction as a *TFT* player. When its ongoing
189 trust level towards the co-player—defined here as the difference between the
190 number of cooperative and defective moves from the co-player so far in the
191 *IPD*—reaches a certain threshold, denoted by θ , it will play *C* uncondition-
192 ally. We denote this strategy by *TUC*. *TUC* is illustrated in the Figure 1
193 representing one game between *TUC* and *TFT*.

194 Given the possibility of being exploited, but still to avoid costly deliber-
195 ation, we assume that *TUC* will check, with a probability p , the co-player’s
196 actions after switching off ². If the co-player is found out to defect, *TUC* will

¹Our 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.

²We assume that, given a sufficient cost of checking, *TUC* can always correctly find

vs TFT.png

Trust reached

Player 1	Checking	Yes	Yes	Yes	No	Yes	No	Yes	No
	TUC Action	C	C	C	C	C	C	C	C
Player 2	Action	C	C	C	C				
	TFT Checking	Yes	Yes	Yes	Yes				

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.

197 revert to its initial strategy, i.e. TFT. As a counterpart of TUC, we consider
 198 TUD that whenever the ongoing trust level reaches the threshold θ , switches
 199 to playing D unconditionally. TUD is illustrated in the Figure 2 representing
 200 one game between TUC and TUD.

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

$$\begin{matrix}
 & \mathbf{AllC} & \mathbf{AllD} & \mathbf{TFT} & \mathbf{TUC} & \mathbf{TUD} \\
 \mathbf{AllC} & R & S & R & R & \frac{\theta R + (r-\theta)S}{r} \\
 \mathbf{AllD} & T & P & \frac{T+(r-1)P}{r} & \frac{T+(r-1)P}{r} & \frac{T+(r-1)P}{r} \\
 \mathbf{TFT} & R - \varepsilon & \frac{S+(r-1)P}{r} - \varepsilon & R - \varepsilon & R - \varepsilon & \frac{\theta R + S + (r-\theta-1)P}{r} - \varepsilon \\
 \mathbf{TUC} & R - \frac{\theta\varepsilon}{r} - \frac{p(r-\theta)\varepsilon}{r} & \frac{S+(r-1)P}{r} - \varepsilon & R - \frac{\theta\varepsilon}{r} - \frac{p(r-\theta)\varepsilon}{r} & R - \frac{\theta\varepsilon}{r} - \frac{p(r-\theta)\varepsilon}{r} & \frac{\Pi_{TUC,TUD}}{r} \\
 \mathbf{TUD} & \frac{\theta R + (r-\theta)T - \theta\varepsilon}{r} & \frac{S+(r-1)P}{r} - \varepsilon & \frac{\theta R + T + (r-\theta-1)P - \theta\varepsilon}{r} & \Pi_{TUD,TUC} & \frac{\theta R + (r-\theta)P - \theta\varepsilon}{r}
 \end{matrix} \quad (2)$$

203

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

out the co-player's actions.

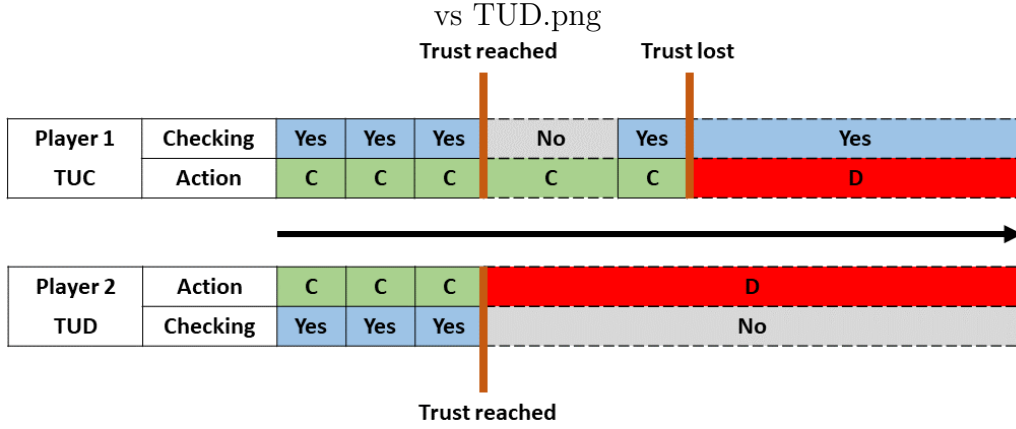


Figure 2: Diagram representing repeated interactions between a trust-based cooperator TUC and a trust-based defector TUD. First, both strategies cooperate and check the other player's action. After θ rounds (here $\theta = 3$), trust is reached for both strategies. TUC now cooperates and TUD defects.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

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

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

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

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

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

211 *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,

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

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

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

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

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

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

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

234 The fixation probability of a single mutant with a strategy A in a population
 235 of $(N-1)$ agents using B is given by (Traulsen et al., 2006; Karlin and Taylor,
 236 1975; Imhof et al., 2005b)

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

237 When considering a set $\{1, \dots, s\}$ of distinct strategies, these fixation proba-
 238 bilities determine the Markov Chain transition matrix $M = \{T_{ij}\}_{i,j=1}^s$, with
 239 $T_{ij, j \neq i} = \rho_{ji}/(s-1)$ and $T_{ii} = 1 - \sum_{j=1, j \neq i}^s T_{ij}$. The normalised eigenvec-
 240 tor of the transposed of M associated with the eigenvalue 1 provides the
 241 above described stationary distribution (Imhof et al., 2005b), which defines
 242 the relative time the population spends while adopting each of the strategies.

243 3. Results

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

254 The default values of the parameters, unless otherwise specified, are for
 255 the game payoffs $R = 1, S = -1, T = 2, P = 0$ (i.e. $b = 2$ and $c = 1$ in
 256 Donation game), the importance of the game $\gamma = 1$, the number of rounds
 257 $r = 50$, the population size $N = 100$, the intensity of selection $\beta = 0.1$, the
 258 trust threshold $\theta = 3$, the probability of checking its partner $p = 0.25$ and the
 259 opportunity cost $\epsilon = 0.25$. The analysis of the model has been implemented
 260 using the package EGTTools (Fernández Domingos, 2020).

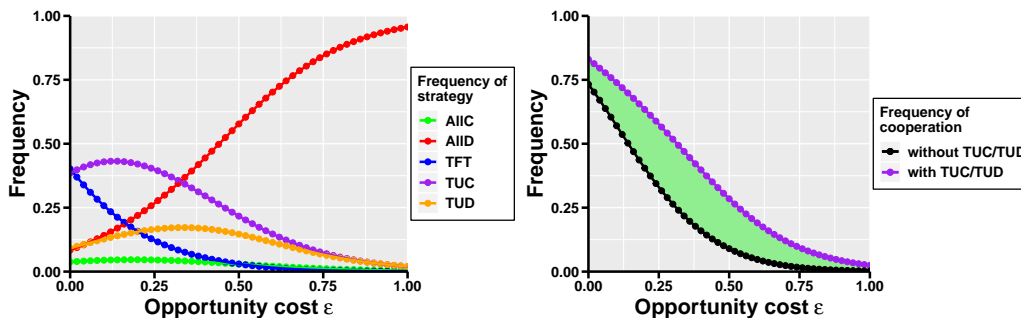


Figure 3: **Left:** Frequency of strategies as a function of the opportunity cost ϵ . **Right:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ . The difference in frequency of cooperation between the two scenario is shaded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

261 *3.1. Trust as a mechanism to reduce opportunity costs*

262 The intuitive benefit of trusting something is to limit the cost of moni-
 263 toring their actions in a long-term interaction, providing a shortcut in the
 264 decision making process. This is in line with several common definitions and
 265 theories of trust (Luhmann, 1979; Grabner-Kraeuter, 2002; Petruzzi et al.,
 266 2014). Thus, we explore first the effect of opportunity cost ϵ on the strategies
 267 employed by individuals and the resulting frequency of cooperation.

268 The left panel of Figure 3 shows that TUC is the most common strategy
 269 for a low to intermediate opportunity cost ϵ (between 0 and 0.3). When the
 270 opportunity cost ϵ is zero, both TUC and TFT are successful strategies and
 271 the population is composed of either one of them for most of the time. The
 272 success of TUC and TFT is explained by the capacity of these strategies to
 273 maintain high levels of cooperation within their homogeneous populations,
 274 while avoiding exploitation by AllD. Yet, the success of TFT is limited by
 275 the opportunity cost paid to check its partner’s actions. This is shown in the
 276 results by the population being mostly AllD when the opportunity cost ϵ is
 277 high. Compared to TFT, TUC can limit this opportunity cost by reducing
 278 its attention to its partner’s actions once trust is reached. This is why as
 279 the opportunity cost increases, the frequency of TFT plummets while TUC
 280 becomes more commonly observed.

281 The right panel of Figure 3 shows that the presence of trust-based strate-

gies 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 $\beta \rightarrow 0$), it would be 60% cooperation vs 66.6% cooperation which is not in favour of our model.

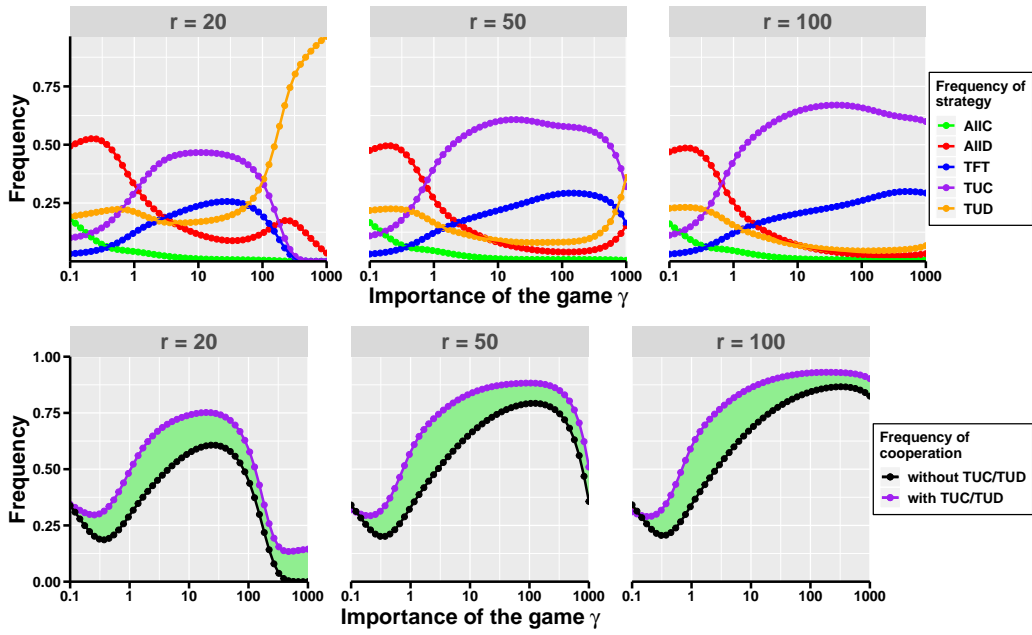


Figure 4: **Top:** Frequency of strategies as a function of the number of rounds r and importance of the game γ (logarithmic scale); **Bottom:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the number of rounds r and importance of the game γ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

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

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

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

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

343 3.3. Trustfulness and TUD

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

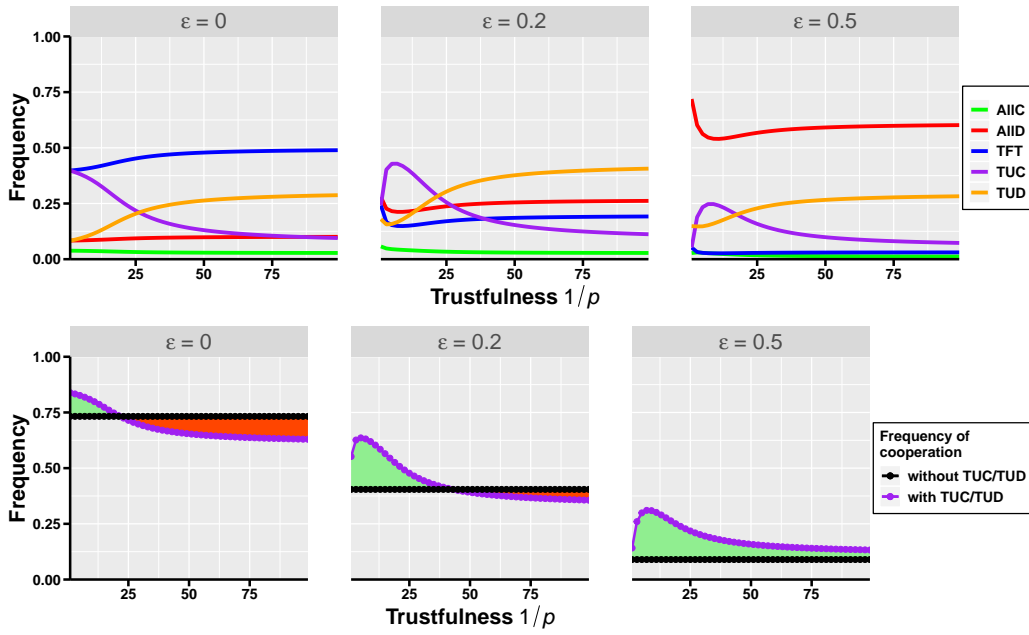


Figure 5: **Top:** Frequency of strategies as a function of the opportunity cost ϵ and trustfulness $1/p$ (average number of rounds between checking event). **Bottom:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ and trustfulness $1/p$ (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.

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

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

367 The bottom part of Figure 5 shows that the presence of trust-based strate-
368 gies increases the frequency of cooperation when the opportunity cost is mod-
369 erate or high. This cooperation improvement is the highest for the optimal
370 trustfulness at which TUC is very frequent. Again, the presence of trust-
371 based strategies can lead to an increase in the frequency of cooperation even
372 if they are not the most frequent strategy e.g. for very high opportunity costs
373 ϵ . Unlike previously, the results also show that the presence of trust-based
374 strategies can reduce the frequency of cooperation. This happens when the
375 opportunity cost ϵ is low and the trustfulness $1/p$ is high. In these con-
376 ditions, trustful and careless TUC players get exploited by TUD players,
377 which increases the frequency of TUD, making cooperation a less viable op-
378 tion (evolutionarily). In the absence of trust-based strategies, TFT is careful
379 enough to avoid this pitfall.

380 To conclude, unconditional trust is a viable strategy only in limited con-
381 ditions and how much TUC relies on trust can have significant effect on the
382 success of the strategy.

383 4. Discussion

384 Trust is a commonly observed mechanism in human interactions, and dis-
385 cussions on the role of trust are being extended to social interactions between

386 humans and intelligent machines (Andras et al., 2018). It is therefore impor-
387 tant to understand how people behave when interacting with those machines;
388 particularly, whether and when they might exhibit trust behaviour towards
389 them? Answering these questions is crucial for providing suitable designs
390 of mechanisms or infrastructures to facilitate human-intelligent machine in-
391 teractions, e.g. in engineering pro-sociality in a hybrid society of humans
392 and machines (Paiva et al., 2018). To this end, this paper provides a game
393 theoretic analysis, where we formalised trust as a behavioural strategy and
394 integrated it into an EGT model to study (i) its success in competition with
395 non-trusting strategies and (ii) its effect on the level of cooperation.

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

406 Overall, our analysis shows that trust can emerge because it reduces the
407 opportunity costs paid by individuals during interactions. It is a form of
408 cognitive shortcut that, while exposing the player to some risks, can allow
409 individuals to cooperate at lesser cost. If the pitfalls of trust have often been
410 discussed, our results underlie the importance of taking into account both
411 the benefits and the risk that the use of trust involves. Understanding the
412 balance between these two is a first step to optimise the benefits of trust in
413 intelligent machines while limiting the costs. On this line, further work could
414 expand the model to look at different forms of trust based cooperation and
415 defection strategies, how they co-evolve, and how exploitation of trust-based
416 cooperators can be avoided. It is noteworthy that our additional (numerical)
417 analyses have shown that all the observations described above (i.e. in Figures
418 3 - 5) are robust, e.g. for different values of the threshold number of rounds
419 required for trust to be established (i.e. θ) (see Appendix). Moreover, note
420 that in the current model we consider that TUC and TUD have the same
421 θ , which is the worst case scenario for the evolution of TUC (and cooper-
422 ation), since it represents the situation where TUD can perfectly recognise
423 when TUC starts trusting a cooperative co-player and therefore becomes less

424 vigilant of exploitation. More realistically, TUD might need to spend extra
425 resources to gather information about TUC (e.g. providers learn about their
426 customers' preferences and behaviours) to determine what is TUC's θ . On
427 the other hand, TUC should not easily reveal or make available their infor-
428 mation (that can be used to infer their θ), to better deal with TUD. Future
429 work should address how these aspects might change the outcome of the
430 evolutionary dynamics.

431 One of the most famous previous formalisations of trust is an experiment
432 from behavioural economics called the trust game (Berg et al., 1995). This
433 game consists of one individual receiving an endowment of money, of which
434 it must choose a certain amount (which can be zero) to send to the other
435 player. The amount sent to the other player is tripled by the experimenter (so
436 that sending money represents an investment). The other player then decides
437 what amount of this money (if any) to send back to the first player (so that
438 there is risk in the first player sending money). While the Nash equilibrium
439 is for the first player to send nothing to the other player, in experiments in-
440 dividuals usually deviate from this by sending a positive amount (Berg et al.,
441 1995). The amount that the first player sends can be understood as mea-
442 suring how much the first player trusts the second to reciprocate. However,
443 in contrast to our formalisation, the trust game measures more a willingness
444 to take risks blindly, as interactions are between anonymous individuals and
445 are played only once. By contrast, we have considered repeated interactions
446 between the same individuals, which has enabled us to look at the success of
447 strategies that build up trust over time.

448 Within the context of the trust game, it has been shown recently that
449 trust and trustworthiness cannot evolve in well-mixed and spatial networks
450 with a homogeneous structure; they can evolve only in heterogeneous net-
451 works under highly advantageous conditions (Kumar et al., 2020). Moreover,
452 within an overall grand challenge to understand the evolution of moral be-
453 haviour (beyond that of cooperation) (Capraro and Perc, 2018), the role of
454 network topology in promoting honesty has been studied in (Capraro et al.,
455 2019), extending works on honest signalling (Smith, 1991). Similarly, ly-
456 ing behaviour have recently been looked at within the context of well-mixed
457 populations (Capraro et al., 2020).

458 In line with our approach, trust enabling strategies were previously con-
459 sidered in the context of repeated games (Han et al., 2011), where trust is
460 built over time as a component of a larger decision making process, for pre-
461 diction of opponents' behaviour. Trust was also considered for enabling co-

462 operation in a one-shot prisoner’s dilemma ([Janssen, 2008](#); [McNamara et al.,](#)
463 [2009](#)), but it was assumed that players can recognise how trustworthy a co-
464 player is based on additional cues such as signalling. Our work differs from
465 these approaches in that we consider trust as a cognitive shortcut to avoid
466 deliberation and having to check the outcomes of previous interaction(s),
467 thereby limiting the opportunity cost of conditional strategies. More gener-
468 ally, the role of an opportunity cost of monitoring the action of a co-player on
469 the equilibrium level of cooperation has been studied in classic game theory
470 models (for instance, see [Lehrer and Solan \(2018\)](#)). Using an evolutionary
471 game theory approach, we complement this literature by showing that trust-
472 based strategies are likely to emerge to deal with the cost of monitoring, even
473 when players are short-sighted and only care about their immediate payoffs.
474 We show conditions under which the presence of trust-based strategies can
475 promote a high level of cooperation in comparison to the case where trust-
476 based strategies are not available, e.g. where only classic reciprocal strategies
477 such as TFT are possible.

478 In addition, trust has been used extensively in various computerised sys-
479 tems, such as in multi-agent open and distributed systems, to facilitate
480 agents’ interactions. Agents may have limited computational and storage
481 capabilities that restrict their control over interactions, and trust is used
482 to minimise uncertainty associated with the interactions, especially when
483 agents inhabit in uncertain and constantly changing environments ([Ram-
484 churn et al., 2004](#); [Falcone and Castelfranchi, 2001](#)). This is the case for vari-
485 ous applications including peer-to-peer computing, smart-grids, e-commerce,
486 etc ([Kamhoua et al., 2011](#); [Ramchurn et al., 2004](#); [Papadopoulou et al., 2001](#);
487 [Petruzzi et al., 2014](#); [Brooks et al., 2020](#)). These studies utilise trust for the
488 purpose of regulating individual and collective behaviours, formalising dif-
489 ferent aspects of trust (such as reputation and belief) ([Castelfranchi, 1997](#);
490 [Castelfranchi and Falcone, 2010](#)). Our results and approach provide novel
491 insights into the design of such computerised and hybrid systems as these
492 require trust to ensure high levels of cooperation or efficient collaboration
493 within a group or team of agents, including human-machine hybrid interac-
494 tions. For instance, our results show that the importance of the business
495 at hand (relative to the opportunity cost) needs to be taken into account
496 to ensure a desired level of cooperation. Also, the system needs to be de-
497 signed so that the opportunity cost of verifying the actions of an intelligent
498 machine is sufficiently low (relative to the benefit and cost of the game) to
499 enable a long-term trusting relationship with customers, e.g. making the

500 activities transparent either directly to the user or to expert auditors that
501 follow professional codes of ethics (Andras et al., 2018).

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

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

529 In the current work we have focused on the Prisoner's Dilemma as it repre-
530 sents the hardest (pairwise) scenario for cooperation to emerge. Many other
531 scenarios might be represented using other social dilemmas such as coordi-
532 nation or snowdrift games (Santos et al., 2006; Sigmund, 2010). Considering
533 such games where it is easier for cooperation to emerge has the potential to
534 open new windows of opportunity for long-term trust-based relationships to
535 be established. Our future work will study how trust-based strategies (as we
536 have modelled) evolve in the context of other social dilemmas. Moreover,
537 given the importance of population structures in the emergence of trust and

538 trustworthiness in the context of the trust game (Kumar et al., 2020), our
539 future work will examine how different population structures influence the
540 outcome of trustful behaviours in our context. Finally, we have assumed that
541 agents always pay the cost of checking, but an alternative is that, they might
542 choose not to check when it is too difficult or costly to do so (for example,
543 checking if an AI development company complies with safety requirements in
544 the development process within a competition to reach technology supremacy
545 (Han et al., 2020)).

546 5. Conclusion

547 We have demonstrated in this paper that evolutionary game theory pro-
548 vides a valuable framework to study trust. Social interactions often result
549 in complex dynamics with unexpected consequences, which a quantitative
550 model is able to shed light on. Our model provides formal support for the
551 theory that trust is a cognitive shortcut which people use to reduce the com-
552 plexity of their interactions. The results of the model provide new insights
553 into the questions of whether and when humans might trust intelligent ma-
554 chines, generating reasonable behavioural hypotheses that empirical studies
555 can test.

556 Acknowledgements

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

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

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

566 Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, ISBN 0-
567 465-02122-2.

568 Axelrod, R. and Hamilton, W. (1981). The evolution of cooperation. *Science*,
569 211:1390–1396.

- 570 Beldad, A., Hegner, S., and Hoppen, J. (2016). The effect of virtual sales
571 agent (VSA) gender – product gender congruence on product advice credi-
572 bility, trust in VSA and online vendor, and purchase intention. *Computers*
573 *in Human Behavior*, 60:62–72.
- 574 Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social
575 history. *Games and Economic Behavior*, 10:122–142.
- 576 Brooks, N. A., Powers, S. T., and Borg, J. M. (2020). A mechanism to
577 promote social behaviour in household load balancing. *Artificial Life Con-*
578 *ference Proceedings*, 32:95–103.
- 579 Capraro, V. and Perc, M. (2018). Grand challenges in social physics: In
580 pursuit of moral behavior. *Frontiers in Physics*, 6:107.
- 581 Capraro, V., Perc, M., and Vilone, D. (2019). The evolution of ly-
582 ing in well-mixed populations. *Journal of the Royal Society Interface*,
583 16(156):20190211.
- 584 Capraro, V., Perc, M., and Vilone, D. (2020). Lying on networks: The
585 role of structure and topology in promoting honesty. *Physical Review E*,
586 101(3):032305.
- 587 Castelfranchi, C. (1997). Modeling social action for AI agents. *IJCAI Inter-*
588 *national Joint Conference on Artificial Intelligence*, 2(c):1567–1576.
- 589 Castelfranchi, C. and Falcone, R. (2010). *Trust theory: A socio-cognitive and*
590 *computational model*, volume 18. John Wiley & Sons.
- 591 Chattaraman, V., Kwon, W.-S., and Gilbert, J. E. (2012). Virtual agents
592 in retail web sites: Benefits of simulated social interaction for older users.
593 *Computers in Human Behavior*, 28(6):2055–2066.
- 594 Chung, H., Iorga, M., Voas, J., and Lee, S. (2017). Alexa, can i trust you?
595 *Computer*, 50(9):100–104. Conference Name: Computer.
- 596 Dahlstrom, R., Nygaard, A., Kimasheva, M., and M. Ulvnes, A. (2014).
597 How to recover trust in the banking industry? A game theory approach
598 to empirical analyses of bank and corporate customer relationships. *Inter-*
599 *national Journal of Bank Marketing*, 32(4):268–278. Publisher: Emerald
600 Group Publishing Limited.

- 601 Dasgupta, P. (2000). Trust as a commodity. *Trust: Making and breaking*
602 *cooperative relations*, 4:49–72.
- 603 Duong, M. H. and Han, T. A. (2019). On equilibrium properties of the
604 replicator–mutator equation in deterministic and random games. *Dynamic*
605 *Games and Applications*, pages 1–23.
- 606 Falcone, R. and Castelfranchi, C. (2001). Social trust: A cognitive approach.
607 In *Trust and deception in virtual societies*, pages 55–90. Springer.
- 608 Fernández Domingos, E. (2020). Egttools: Toolbox for evolutionary game
609 theory. <https://github.com/Socrats/EGTTools>.
- 610 Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games
611 with discounting or with incomplete information. *Econometrica*, 54:533–
612 554.
- 613 García, J. and Traulsen, A. (2012). The structure of mutations and the
614 evolution of cooperation. *PloS one*, 7(4):e35287.
- 615 Garcia, J. and van Veelen, M. (2018). No strategy can win in the repeated
616 prisoner’s dilemma: linking game theory and computer simulations. *Frontiers in Robotics and AI*, 5:102.
- 618 Glynatsi, N. and Knight, V. (2020). Using a theory of mind to find best
619 responses to memory-one strategies. *Scientific Reports*, 17287:10.
- 620 Grabner-Kraeuter, S. (2002). The role of consumers’ trust in online-shopping.
621 *Journal of Business Ethics*, 39(1):43–50.
- 622 Han, T. A. (2013). *Intention recognition, commitment and their roles in the*
623 *evolution of cooperation: From Artificial intelligence techniques to evolu-*
624 *tionary game theory models*. Springer SAPERE, vol. 9.
- 625 Han, T. A., Moniz Pereira, L., and Santos, F. C. (2011). Intention recognition
626 promotes the emergence of cooperation. *Adaptive Behavior*, 19(4):264–279.
- 627 Han, T. A., Pereira, L. M., and Santos, F. C. (2012). The emergence of com-
628 mitments and cooperation. In *Proceedings of the 11th International Con-*
629 *ference on Autonomous Agents and Multiagent Systems (AAMAS’2012)*,
630 pages 559–566. ACM.

- 631 Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013a). Good
632 agreements make good friends. *Scientific reports*, 3:2695.
- 633 Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013b). Why is
634 it so hard to say sorry? evolution of apology with commitments in the iter-
635 ated prisoner’s dilemma. In *Proceedings of the Twenty-Third international*
636 *joint conference on Artificial Intelligence*, pages 177–183.
- 637 Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2020). To Regu-
638 late or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal*
639 *of Artificial Intelligence Research*, pre-print available at *arXiv:1907.12393*.
640 In Press.
- 641 Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., and Sigmund, K. (2007).
642 Via freedom to coercion: the emergence of costly punishment. *science*,
643 316(5833):1905–1907.
- 644 Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K., and Nowak, M. A. (2017).
645 Memory-n strategies of direct reciprocity. *Proceedings of the National*
646 *Academy of Sciences*, 114(18):4715–4720.
- 647 Ho, T.-H. (1996). Finite automata play repeated prisoner’s dilemma with
648 information processing costs. *Journal of economic dynamics and control*,
649 20(1-3):173–207.
- 650 Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2005a). Evolutionary cycles
651 of cooperation and defection. *Proceedings of the National Academy of*
652 *Sciences of the United States of America*, 102(31):10797–10800.
- 653 Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2005b). Evolutionary cycles
654 of cooperation and defection. *Proceedings of the National Academy of*
655 *Sciences of the United States of America*, 102:10797–10800.
- 656 Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2007). Tit-for-tat or win-
657 stay, lose-shift? *Journal of Theoretical Biology*, 247(3):574 – 580.
- 658 Janssen, M. A. (2008). Evolution of cooperation in a one-shot prisoner’s
659 dilemma based on recognition of trustworthy and untrustworthy agents.
660 *Journal of Economic Behavior & Organization*, 65(3-4):458–471.

- 661 Jugovac, M. and Jannach, D. (2017). Interacting with Recom-
662 menders: Overview and research directions. *ACM Transactions on Inter-*
663 *active Intelligent Systems*, 7(3):10:1–10:46.
- 664 Kamhoua, C. A., Pissinou, N., and Makki, K. (2011). Game theoretic model-
665 ing and evolution of trust in autonomous multi-hop networks: Application
666 to network security and privacy. In *2011 IEEE International Conference*
667 *on Communications (ICC)*, pages 1–6. IEEE.
- 668 Karlin, S. and Taylor, H. E. (1975). *A First Course in Stochastic Processes*.
669 Academic Press, New York.
- 670 Kumar, A., Capraro, V., and Perc, M. (2020). The evolution of trust and
671 trustworthiness. *Journal of the Royal Society Interface*, 17(169):20200491.
- 672 Laaksonen, T., Jarimo, T., and Kulmala, H. I. (2009). Cooperative strategies
673 in customer–supplier relationships: The role of interfirm trust. *Interna-*
674 *tional Journal of Production Economics*, 120(1):79–87.
- 675 Lehrer, E. and Solan, E. (2018). High frequency repeated games with costly
676 monitoring. *Theoretical Economics*, 13(1):87–113.
- 677 Lewis, G. (2011). Asymmetric information, adverse selection and on-
678 line disclosure: The case of eBay Motors. *American Economic Review*,
679 101(4):1535–1546.
- 680 Luhmann, N. (1979). *Trust and Power*. John Wiley & Sons, Chichester.
- 681 Macy, M. W. and Flache, A. (2002). Learning dynamics in social dilemmas.
682 *Proceedings of the National Academy of Sciences of the United States of*
683 *America*, 99:7229–7236.
- 684 Mahadevan, B. (2000). Business models for internet-based e-commerce: An
685 anatomy. *California Management Review*, 42(4):55–69. Publisher: SAGE
686 Publications Inc.
- 687 Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., and Lenaerts, T. (2015).
688 Apology and forgiveness evolve to resolve failures in cooperative agree-
689 ments. *Scientific reports*, 5:10639.
- 690 Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge
691 University Press, Cambridge, UK.

- 692 McNally, L., Brown, S. P., and Jackson, A. L. (2012). Cooperation and the
693 evolution of intelligence. *Proceedings of the Royal Society B: Biological*
694 *Sciences*, 279(1740):3027–3034.
- 695 McNamara, J. M., Stephens, P. A., Dall, S. R., and Houston, A. I. (2009).
696 Evolution of trust and trustworthiness: social awareness favours person-
697 ality differences. *Proceedings of the Royal Society B: Biological Sciences*,
698 276(1657):605–613.
- 699 Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses
700 to computers. *Journal of Social Issues*, 56(1):81–103.
- 701 Nowak, M. A. and Sigmund, K. (1993). A strategy of win-stay, lose-shift
702 that outperforms tit-for-tat in prisoner’s dilemma. *Nature*, 364:56–58.
- 703 Nunes, I. and Jannach, D. (2017). A systematic review and taxonomy of
704 explanations in decision support and recommender systems. *User Modeling*
705 *and User-Adapted Interaction*, 27(3):393–444.
- 706 Paiva, A., Santos, F. P., and Santos, F. C. (2018). Engineering pro-sociality
707 with autonomous agents. In *Thirty-second AAAI conference on artificial*
708 *intelligence*.
- 709 Papadopoulou, P., Andreou, A., Kanellis, P., and Martakos, D. (2001). Trust
710 and relationship building in electronic commerce. *Internet research*.
- 711 Petruzzi, P. E., Busquets, D., and Pitt, J. (2014). Experiments with social
712 capital in multi-agent systems. In Dam, H. K., Pitt, J., Xu, Y., Gover-
713 natori, G., and Ito, T., editors, *PRIMA 2014: Principles and Practice of*
714 *Multi-Agent Systems*, Lecture Notes in Computer Science, pages 18–33,
715 Cham. Springer International Publishing.
- 716 Pu, P. and Chen, L. (2007). Trust-inspiring explanation interfaces for rec-
717 ommender systems. *Knowledge-Based Systems*, 20(6):542–556.
- 718 Ramchurn, S. D., Huynh, D., and Jennings, N. R. (2004). Trust in multi-
719 agent systems. *The Knowledge Engineering Review*, 19(1):1–25.
- 720 Rand, D. G., Tarnita, C. E., Ohtsuki, H., and Nowak, M. A. (2013). Evolu-
721 tion of fairness in the one-shot anonymous ultimatum game. *Proceedings*
722 *of the National Academy of Sciences*, 110(7):2581–2586.

- 723 Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006). Evolutionary dynam-
724 ics of social dilemmas in structured heterogeneous populations. *Proceed-*
725 *ings of the National Academy of Sciences of the United States of America*,
726 103:3490–3494.
- 727 Santos, F. P., Pacheco, J. M., Paiva, A., and Santos, F. C. (2019). Evolu-
728 tion of collective fairness in hybrid populations of humans and agents. In
729 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33,
730 pages 6146–6153.
- 731 Shoham, Y. (2008). Computer science and game theory. *Communications of*
732 *the ACM*, 51(8):74–79.
- 733 Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton University Press.
- 734 Smith, J. M. (1991). Honest signalling: the philip sidney game. *Animal*
735 *Behaviour*.
- 736 Traulsen, A., Hauert, C., De Silva, H., Nowak, M. A., and Sigmund, K.
737 (2009). Exploration dynamics in evolutionary games. *Proceedings of the*
738 *National Academy of Sciences*, 106(3):709–712.
- 739 Traulsen, A., Nowak, M. A., and Pacheco, J. M. (2006). Stochastic dynamics
740 of invasion and fixation. *Phys. Rev. E*, 74:11909.
- 741 Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley
742 & Sons.
- 743 Yoo, K.-H., Gretzel, U., and Zanker, M. (2012). *Persuasive Recommender*
744 *Systems: Conceptual Background and Implications*. Springer Publishing
745 Company, Incorporated, 1st edition.

746 **Competing interests statement**

747 The authors have not competing interests.

748

749 **Author contributions**

750 **The Anh Han:** Conceptualization, Methodology, Formal Analysis, Writing
751 - Original draft preparation

752 **Cedric Perret:** Conceptualization, Methodology, Formal Analysis, Soft-
753 ware, Writing - Original draft preparation

754 **Simon T. Powers:** Conceptualization, Methodology, Writing - Original
755 draft preparation

756

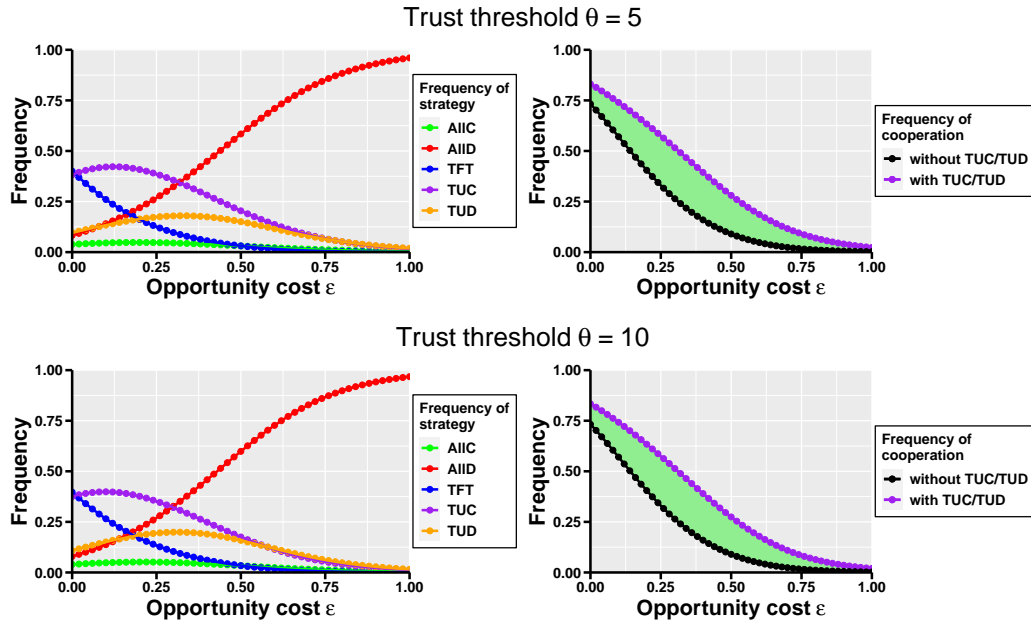


Figure 6: **Left:** Frequency of strategies as a function of the opportunity cost ϵ . **Right:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ . The difference in frequency of cooperation between the two scenario is shaded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

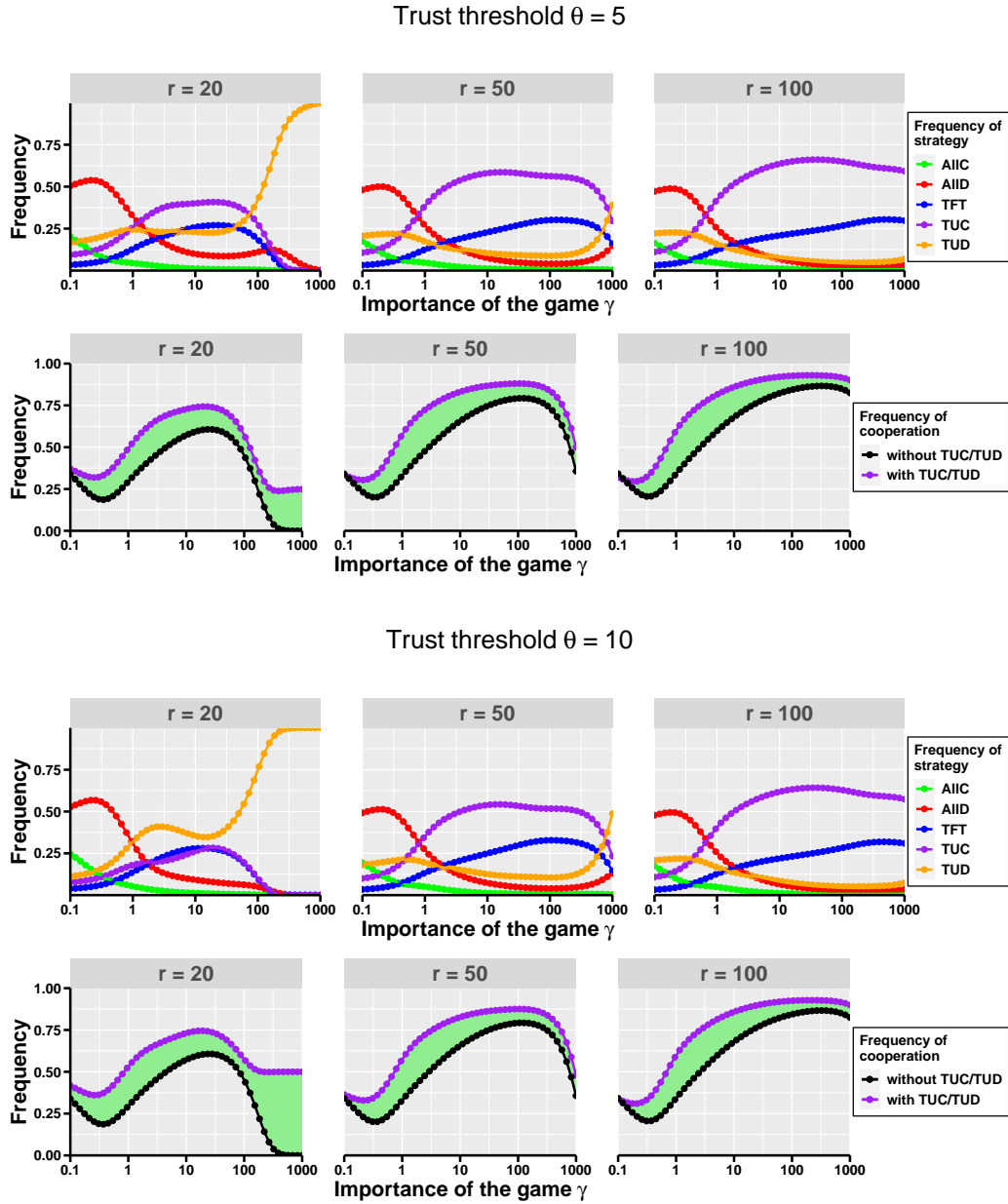


Figure 7: **Top:** Frequency of strategies as a function of the number of rounds r and importance of the game γ (logarithmic scale); **Bottom:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the number of rounds r and importance of the game γ (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: $\beta = 0.1$, population size $N = 100$, $p = 0.25$, $R = 1$, $S = -1$, $T = 2$, $P = 0$.

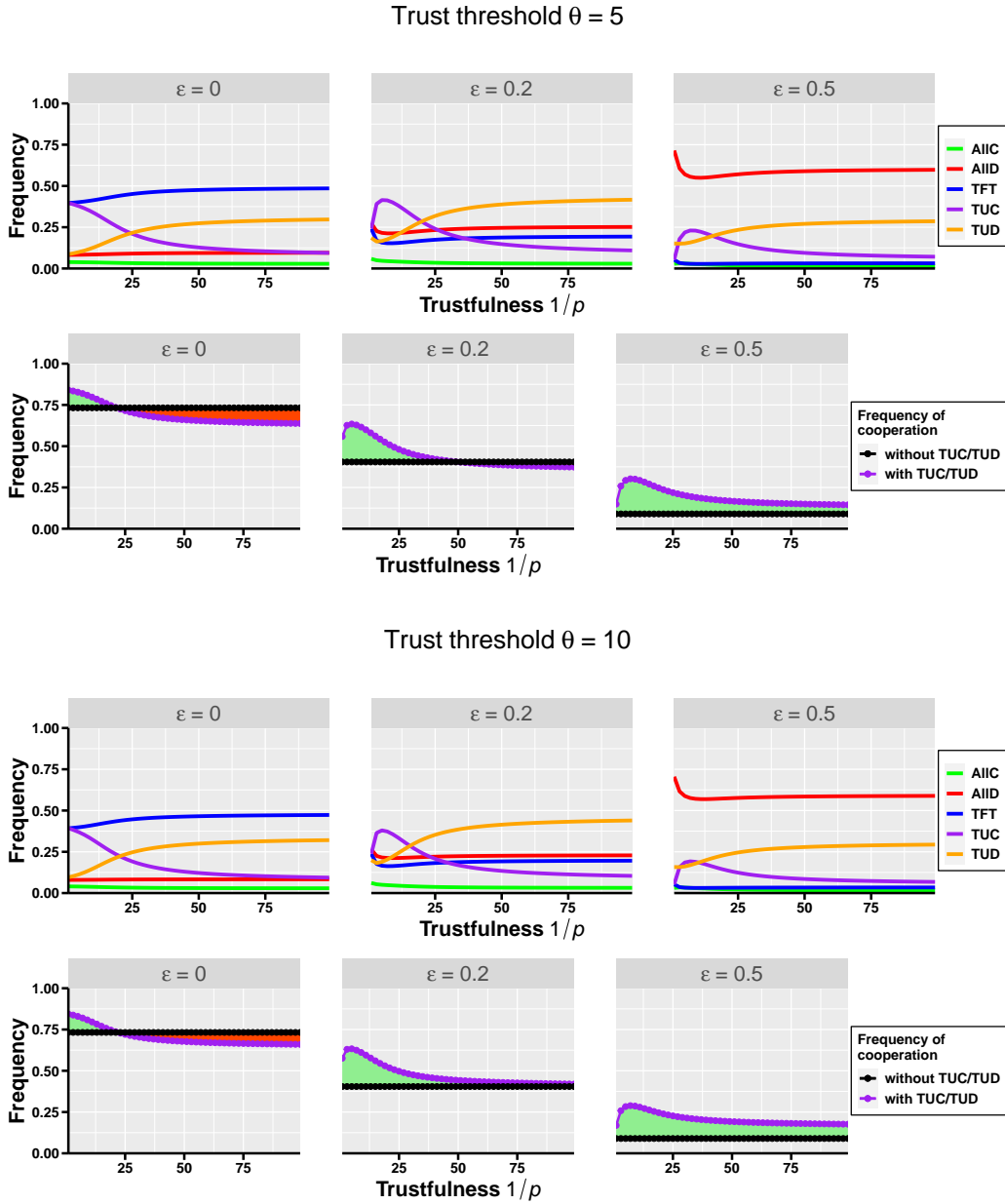


Figure 8: **Top:** Frequency of strategies as a function of the opportunity cost ϵ and trustfulness $1/p$ (average number of rounds between checking event). **Bottom:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ and trustfulness $1/p$ (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: $\beta = 0.1$, population size $N = 100$, $\gamma = 1, r = 50, R = 1, S = -1, T = 2, P = 0$.

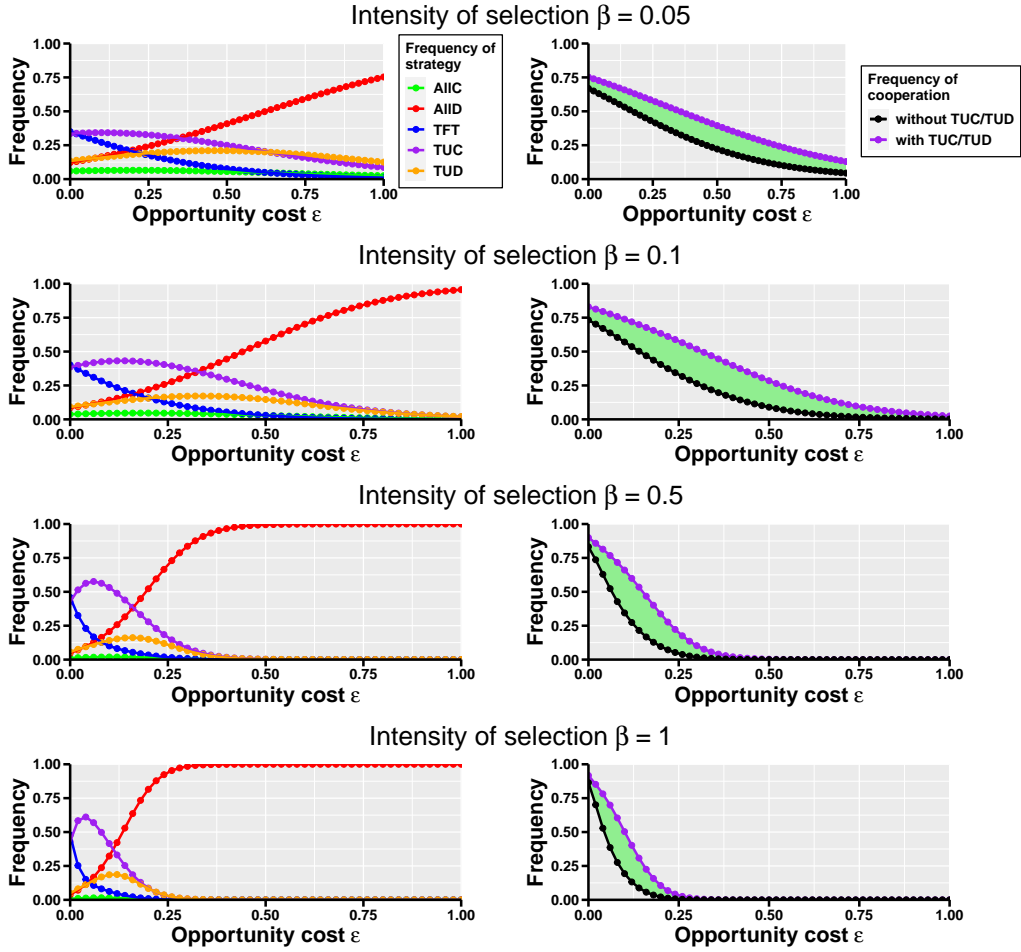


Figure 9: **Left:** Frequency of strategies as a function of the opportunity cost ϵ . **Right:** Frequency of cooperation in absence or presence of trust-based strategies TUC and TUD, as a function of the opportunity cost ϵ . The difference in frequency of cooperation between the two scenario is shaded 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: population size $N = 100$, $\gamma = 1$, $r = 50$, $p = 0.25$, $R = 1$, $S = -1$, $T = 2$, $P = 0$.