The Triangles of Dishonesty: Modelling the Evolution of Lies, Bullshit, and Deception in Agent Societies

Ștefan Sarkadi
King's College London
London, United Kingdom
stefan.sarkadi@kcl.ac.uk

Peter R. Lewis
Ontario Tech University
Oshawa, Canada
peter.lewis@ontariotechu.ca

ABSTRACT

Misinformation and disinformation in agent societies can be spread due to the adoption of dishonest communication. Recently, this phenomenon has been exacerbated by advances in AI technologies. One way to understand dishonest communication is to model it from an agent-oriented perspective. In this paper we model dishonesty games considering the existing literature on lies, bullshit, and deception, three prevalent but distinct forms of dishonesty. We use an evolutionary agent-based replicator model to simulate dishonesty games and show the differences between the three types of dishonest communication under two different sets of assumptions: agents are either self-interested (payoff maximizers) or competitive (relative payoff maximizers). We show that: (i) truth-telling is not stable in the face of lying, but that interrogation helps drive truth-telling in the self-interested case but not the competitive case; (ii) that in the competitive case, agents stop bullshitting and start truth-telling, but this is not stable; (iii) that deception can only dominate in the competitive case, and that truth-telling is a saddle point in which agents realise deception can provide better payoffs.

KEYWORDS

Deception; Dishonesty; Evolution; Agent Societies; Simulation

ACM Reference Format:
Ștefan Sarkadi and Peter R. Lewis. 2024. The Triangles of Dishonesty: Modelling the Evolution of Lies, Bullshit, and Deception in Agent Societies. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024,* IFAAMAS, 9 pages.

1 INTRODUCTION & BACKGROUND

The literature in philosophy and AI identifies three main types of dishonest communication, namely lying, bullshitting[1], and deception [10, 14, 17]. Each type of dishonesty is produced by an agent under different epistemological conditions. In this paper, we aim to model lying, bullshitting, and deception, based on the three definitions given by [14], who modelled the them in agent-oriented programming languages (AOPLs) based on the definitions of lying and bullshit by [7], namely:

Lying - *The dishonest behaviour of an agent A to tell another agent B that $\neg\phi$ is the case, when in fact A believes that ϕ is the case.*

Bullshit - *The dishonest behaviour of an agent A to tell another agent B that ϕ is the case, when in fact A does not know if ϕ is the case.*

Both lying and bullshitting can be intentional or unintentional. When an agent intends that the target believes the untruthful statement it communicates to be true, lying happens, whereas when an agent intends the target to believe that what it communicates to be true irrespective of the actual truth-value of the communicated statement, then bullshitting happens.

Deception - *The intentional process of an agent A to make another agent B believe something is true (false) that A believes is false (true), with the aim of achieving an ulterior goal.*

The shrouded master of the triangle of dishonesty, deception, is ultimately the most complex, sophisticated and most difficult to detect if well performed. Deception is more fine-grained w.r.t. intentionality. When an agent attempts deception from a practical reasoning perspective, this is always an intended process and can comprise of both lying and truth-telling [12]. However, under certain circumstances, the target might be so skeptical or biased that deception happens even if the deceiver does not intend it to [20]. Or, in the case of deception in nature, it is an intrinsic evolutionary trait of a species, e.g. green fungi that attract insects for reproductive purposes [25]. Others argue that deception in animals can actually indicate higher-cognition in some species [23]. Furthermore, it has been theorised that both deception and deception detection have evolved during natural selection, and that both are crucial drivers of complex adaptation in agents on multiple levels, namely physiological, social, and cognitive [2]. Yet, our understanding of deception and its role in the evolution of cognition is still in its infancy [3].

From an evolutionary perspective in Artificial Intelligence (AI), [21, 22] have studied deception in hybrid self-organising agent societies, comprised of agents that learned through imitation and exploration, by modelling knowledge sharing as public goods games and then later showed how an arms race in Theory of Mind is triggered by the presence of deception [18]. A different approach was taken by [8], who have studied the evolution of multiple types of deceptive strategies using an iterated prisoner's dilemma game. Additionally, [6] have studied the evolution of preferences in deceptive signalling games using the bounded confidence model, and [5] developed a model to specify coordinated deception strategies in adaptive software architectures.

In this paper we take a new approach and adopt the three types of dishonesty and apply evolutionary agent-based modelling with replicator dynamics, similarly to the approach by [16], in order to study the effects of the 2-agent interaction outcomes on agent

[1]We understand that the term 'bullshit' might be offensive in certain contexts. However, here it is used to refer to a particular form of dishonesty previously studied and well defined, using this term, in the literature [7, 15].

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.

Table 1: Parameters

P	Value	
		Game parameters
α	$[0, 1]$	value of information
β	$[0, 1]$	false information factor
ϵ	$[0, 1]$	penalty of exposure
ρ	$[0, 1]$	bullshitting factor/reputation gain
γ	$[0, 1]$	investigation factor
θ	$[0, 1]$	deception factor
		System parameters
G	10	generations
R	100	replication steps simulated in each generation

societies where agents can choose between one of the dishonest strategies and two other cooperative strategies of communication.

2 MODELLING DISHONESTY GAMES

In agent societies, these three types of dishonest behaviour are being used alongside cooperative and punishment behaviours. Dishonesty might give communicative agents some evolutionary advantages over others, as it happens in nature according to [25]. Each of the following games are designed to check how a population of agents evolves using replicator dynamics. In each game there are three strategies present, one strategy being either of the three dishonest strategies, namely lying (L), bullshitting (B), or deceiving (D). The other two cooperative strategies are always truth-telling (T), which is the strategy of communicating truthfully when sharing information, and investigation (I), which represents the strategy of communicating truthfully and fact-checking the agents that use dishonest strategies.

Every game is a 2-player game. This means that the agent-agent interactions always happen in pairs, i.e. payoffs are a 3x3 matrix that represent the payoffs the agents get when they interact with another agent of the same or different strategy.

We assume that information is not spread through the population, i.e., agents do not re-share information that they learnt from other agents, but only draw on their own prior expertise when sharing information. This allows us to ignore the possibility that incorrect information received from another agent may be passed on unwittingly by cooperators. However, an extension of this model to explore the spread of misinformation could prove interesting. We make two sets of assumptions regarding the games.

Assumptions
We study and compare the evolutionary dynamics under two different assumptions:
(i) **Self-Interested Assumption:** Agents only care about the payoffs they receive – this forms our set of base games;
(ii) **Competitive Assumption:** Agents care if they do better than others, irrespective of the absolute magnitude of their own payoffs – this leads to game variants which we denote with an X.

Before we describe the payoffs agents get when they play dishonesty games, i.e. choose dishonest strategies, or meet other agents that use dishonest strategies, we first describe the payoffs they get when they choose to play cooperative strategies when they meet other agents that also use cooperative strategies in the context of multi-agent communication, i.e. when Truth-Tellers and Investigators meet other Truth-Tellers and Investigators.

Truth-Teller-meets-Truth-Teller (TT). When a Truth-Teller meets another Truth-Teller they exchange new information in a truthful manner. The value of the truthful information is represented by α. Hence, each Truth-Teller gains the value of new information α. Under the competitive assumption, each Truth-Teller gains 0, because the agent it interacts with it gains the same payoff.

Truth-Teller-meets-Investigator (TI). When a Truth-Teller meets an Investigator they exchange new information in a truthful manner. Hence, the Truth-Teller gains the value of new information α. When an Investigator meets a Truth-Teller they exchange information in a truthful manner. Hence, The Investigator gains $\alpha - \gamma$, the value of the new information minus the cost associated with fact-checking that information. In the competitive case, the Truth-Teller gains γ, because even if it communicates truthfully with the Investigator, the Investigator still has to pay γ in order to fact-check the exchanged information. Under the competitive assumption, the Investigator gains $-\gamma$, because compared to the truth-teller, it still needs to fact-check information.

Investigator-meets-Investigator (II). When an Investigator meets another Investigator, they both exchange truthful information and also fact-check each other. This type of information exchange could represent an ideal evidence-based interrogation dialogue, as described by [26], where both agents are cooperatively and only partially information-seeking (because they can access the received information somewhere else to fact-check it). Hence, each Investigator gains the same payoff as if it meets a Truth-Teller minus the cost of fact-checking $\alpha - \gamma$. Under the competitive assumption, the Investigators gain 0.

2.1 The Liar's Game (LTI)

In this game, agents can choose to lie (L) or tell the truth (T) based on their beliefs about the state of the world, or investigate the state of the world, tell the truth, fact-check what others say and punish others if they intentionally spread lies (I). - see Tables 2,3. Possible outcomes from interactions:

Liar-meets-Truth-Teller (LT). The Liar gains the information value α, provided by the truth-teller. The Truth-Teller gains the value of the information, α, but some or all of this information is false. The cost of this falsity is β, so it gains $\alpha - \beta$. For example, if all of the information exchanged is false, $\beta = \alpha$ and there is no value to the information gained. If $\beta < \alpha$ then some residual value remains in the information. If $\beta > \alpha$ then the cost of receiving false information is worse than receiving no information at all. Under the competitive assumption, the Liar gains β because it is the cost imposed by the Liar on the Truth-Teller. Under the competitive assumption, the Truth-Teller gains $-\beta$.

Liar-meets-Liar (LL). Each Liar gains the information value α, but some or all of this information is false, at a cost of β, so they each gain $\alpha - \beta$. Under the competitive assumption, the Liar gains 0 because the Liars cancel each other out.

Liar-meets-Investigator (LI). Liar gains α if it meets the Investigator. If we also consider the penalty for exposure ϵ, then Liar gains $\alpha - \epsilon$. When an Investigator meets a Liar, it gains the information value and pays the cost of fact-checking $\alpha - \gamma$.

Under the competitive assumption the Investigator gains $-\gamma$, while the Liar gains γ, as this is the cost the Liar imposes on the

Investigator. If we take into account the penalty for exposure under the competitive assumption, then the Liar gains $\gamma - \epsilon$ and the Investigator gains $\epsilon - \gamma$, as the Investigator receives as a reward the cost it imposes on the Liar.

Table 2: Liar's Game

	no ϵ			with ϵ		
	L	T	I	L	T	I
L	$\alpha - \beta$	α	α	$\alpha - \beta$	α	$\alpha - \epsilon$
T	$\alpha - \beta$	α	α	$\alpha - \beta$	α	α
I	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$

Table 3: Liar's Game X

	no ϵ			with ϵ		
	L	T	I	L	T	I
L	0	β	γ	0	β	$\gamma - \epsilon$
T	$-\beta$	0	γ	$-\beta$	0	γ
I	$-\gamma$	$-\gamma$	0	$\epsilon - \gamma$	$-\gamma$	0

2.2 The Bullshitter's Game (BTI)

In this game, agents can choose to bullshit by being ignorant about the state of the world (B), tell the truth without being ignorant of the state of the world (T), or tell the truth and fact-check what others say about the state of the world (I). - see Tables 4,5. Possible outcomes from interactions:

Bullshitter-meets-Truth-Teller (BT). When a Bullshitter meets a Truth-Teller, it gains both the information value α and gains reputation ρ. When a Truth-Teller meets a Bullshitter, it gains $\alpha - \beta$ for the same reason as when meeting a Liar. Under the competitive assumption, the Bullshitter gains reputation as well as the cost of false information it imposes on the Truth-Teller $\rho + \beta$, whereas the Truth-Teller only gains $-\beta$.

Bullshitter-meets-Bullshitter (BB). When a Bullshitter meets another Bullshitter, it gains the value of the information that is affected by the other's false information. However, unlike the Liar, the Bullshitter also gains reputation by interacting with another dishonest agent and thrives on this reputation gain because it cares that others perceive it sharing information. Hence the Bullshitter gains $\alpha - \beta + \rho$. Under the competitive assumption, both Bullshitters gain 0.

Bullshitter-meets-Investigator (BI). When a Bullshitter meets an Investigator, it gains $\alpha - \rho$ because it receives the information value, but it also loses reputation. If we also consider the penalty for exposure, then the Bullshitter gains $\alpha - \rho - \epsilon$. When an Investigator meets a Bullshitter, the Investigator gains the information value and pays the cost of fact-checking $\alpha - \gamma$. Under the competitive assumption, the Investigator gains $\rho - \gamma$. If we consider the penalty of exposure under the competitive assumption, then the Investigator gains $\rho + \epsilon - \gamma$ by also receiving as a reward the additional cost of exposure imposed on the Bullshitter. Under the competitive assumption, the Bullshitter gains $\gamma - \rho$ because it receives as reward the cost imposed on the Interrogator, while also losing reputation. If we also consider the penalty for exposure under the competitive assumption, then the Bullshitter gains $\gamma - \rho - \epsilon$.

Table 4: Bullshitter's Game

	no ϵ			with ϵ		
	B	T	I	B	T	I
B	$\alpha - \beta + \rho$	$\alpha + \rho$	$\alpha - \rho$	$\alpha - \beta + \rho$	$\alpha + \rho$	$\alpha - \rho - \epsilon$
T	$\alpha - \beta$	α	α	$\alpha - \beta$	α	α
I	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$	$\alpha - \gamma$

Table 5: Bullshitter's Game X

	no ϵ			with ϵ		
	B	T	I	B	T	I
B	0	$\rho + \beta$	$\gamma - \rho$	0	$\rho + \beta$	$\gamma - \rho - \epsilon$
T	$-\beta$	0	γ	$-\beta$	0	γ
I	$\rho - \gamma$	$-\gamma$	0	$\rho + \epsilon - \gamma$	$-\gamma$	0

2.3 The Deceiver's Game (DTI)

In this game, agents can choose to deceive others (not necessarily about the state of the world), tell the truth, or tell the truth and investigate whether others are being truthful or deceptive.

What is distinct from the previous games here, is that the deceiver has an ulterior goal that it aims to achieve. Liars for instance, only aim to successfully report a falsehood which can be easily fact-checked by the investigator. - see Tables 6, 7. Possible outcomes from interactions:

Deceiver-meets-Truth-Teller (DT). When a Deceiver meets a Truth-Teller, it gains $\alpha - \theta$, because it gains the information value and pays the cognitive cost for deceiving. The Truth-Teller gains $\alpha - \beta$ when it meets a Deceiver because it gains the information value, but this information value is affected by the false information factor. Under the competitive assumption, the Truth-Teller gains $-\beta + \theta$ because it also imposes the cost of performing deception on the Deceiver. Under the competitive assumption, the Deceiver gains $-\theta + \beta$ because it receives as reward the cost of false information imposed on the Truth-Teller.

Deceiver-meets-Deceiver (DD). When a Deceiver meets another Deceiver, it gains $\alpha - \beta - \theta$ which means it gains the information value that is influenced by the false information factor and pays the cost for deception. Under the competitive assumption both Deceivers gain 0.

Deceiver-meets-Investigator (DI). When a Deceiver meets an Investigator, their interactions resembles an interrogation based dialogue as described by [20], where the Deceiver aims to outsmart and cause the Interrogator to have a false belief, whereas the Investigator aims to outsmart the Deceiver, find out the truth, and expose the deception. Hence, the Deceiver gains $\alpha - \theta - \epsilon(\gamma - \theta)$ because it gains the information value, pays the cognitive cost for deceiving, but it also penalty for being exposed ϵ. Notice, however, that in the Deceiver's Game, the penalty for being exposed is discounted by the difference between the investigation factor, which in this case we interpret as resources, cognitive or otherwise, dedicated

for deception detection by the Investigator, and the deception factor, which in this case we interpret as the resources, cognitive or otherwise, allocated by the Deceiver to deceive the Investigator in order not to be exposed. Hence this penalty for exposure might actually become a reward for not being exposed if the deception is done properly. When an Investigator meets a Deceiver, it gains $\alpha - \gamma + \epsilon(\gamma - \theta)$ because it gains the information value and pays the cost for fact-checking, but it also gains the reward or pays the cost for trying to expose the Deceiver. Notice that this reward or cost is based on the penalty for exposure ϵ, which is discounted by the difference between the resources, cognitive or otherwise, allocated by the Investigator versus those allocated by the Deceiver. Under the competitive assumption the Deceiver gains $\gamma - \theta - \epsilon(\gamma - \theta)$. Under the competitive assumption, the Investigator gains $-\gamma + \theta + \epsilon(\gamma - \theta)$.

Table 6: Deceiver's Game with penalty for exposure ϵ.

	D	T	I
D	$\alpha - \beta - \theta$	$\alpha - \theta$	$\alpha - \theta - \epsilon(\gamma - \theta)$
T	$\alpha - \beta$	α	α
I	$\alpha - \gamma + \epsilon(\gamma - \theta)$	$\alpha - \gamma$	$\alpha - \gamma$

Table 7: Deceiver's Game X with penalty for exposure ϵ.

	D	T	I
D	0	$-\theta + \beta$	$+\gamma - \theta - \epsilon(\gamma - \theta)$
T	$-\beta + \theta$	0	γ
I	$-\gamma + \theta + \epsilon(\gamma - \theta)$	$-\gamma$	0

3 RESULTS

The results described in this section have been produced using the parameters described in Table 1. We used an ABM replicator to run the simulations for 10 experimental setups, each corresponding to (i) one of the dishonesty games, namely LTI, BTI, and DTI, together with (ii) the condition of whether the game is played under the competitive assumption, and (iii) for the LTI and BTI games whether the penalty for exposure ϵ is considered. We used the egtplot library by [13] to plot the replicator dynamics as a 2D simplex, where each corner of the triangle represents one of the strategy in the games. For exploring the parameter effects we have varied each parameter one-at-a-time, except for the preset values of α, β, and γ specific to each game type, and the number of generations G and replication steps T. Each simulation starts with 100 initial conditions that represent the state of the population w.r.t. the strategy distributions on the simplex grid. When exploring parameter effects, such as those of ϵ, ρ, and θ we chose the values $[0.1, 0.5, 1]$ for varying the parameters.

To read the figures, one must keep the following in mind: starting conditions of populations of agents are represented in grid-form on a 2D simplex; the top corner represents the Investigator strategy; the bottom-right corner represents the Truth-Teller strategy; the bottom-left corner represents the Dishonest strategy, which can be either Liar, Bullshitter or Deceiver, depending on the game that is modelled, namely LTI, BTI, or DTI; dashed lines which indicate that every point on an edge or line is an equilibrium; white circles indicate an unstable equilibrium; black circles indicate a stable equilibrium; grey circles indicate a saddle point; the colours of the speed bars indicate the speed at which the population of agents changes, and the arrows indicate the direction towards which the population evolves in the respective figures.

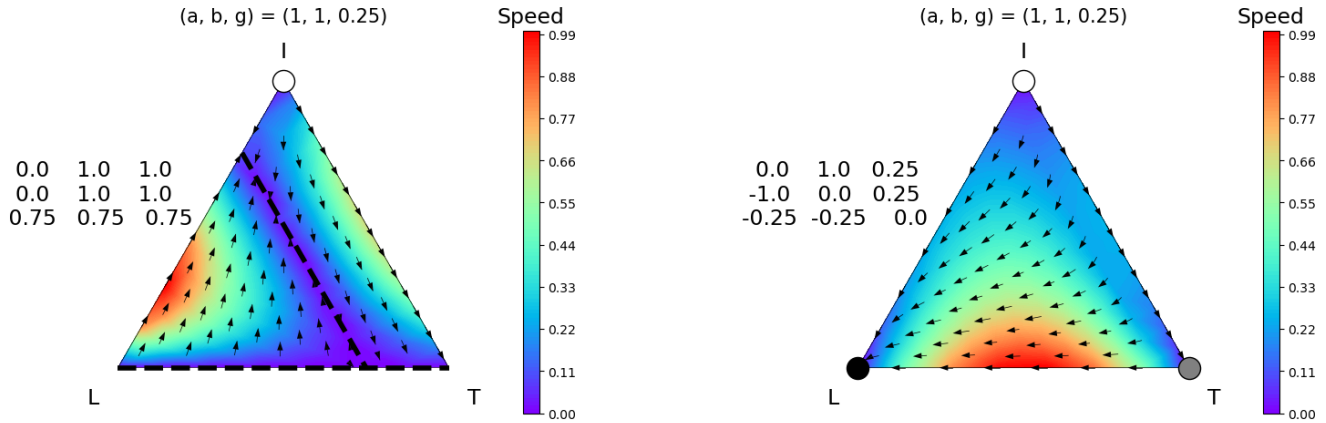
Fig. 1 shows us the results for the Liar's Game (Fig. 1a) and Liar's Game X (Fig. 1b) - that is the game played under the competitive assumption. In these two games, $\alpha = \beta = 1$ and $\gamma = 0.25$ with no penalty for exposure taken into account. Fig. 2 shows us the games when considering the penalty for exposure ϵ in the Liar's Game in Fig. 2a, and the results for Liar's Game X in Fig. 2b when playing the game under the competitive assumption.

The parameter values for α and γ remain the same in the Bullshitter's Game. However β is changed to $\beta = 0.5$ in order to model the effect of bullshitting, where the false information has an effect on the value of information, but it might very well be that the bullshitter agent actually makes a truthful statement, but the bullshitter does not know it. Fig. 3a shows us the results for the Bullshitter's Game and Fig. 3b the results for Bullshitter's Game X along with the effect of the reputation gain ρ for the Bullshitter. As before, these results do not take into account the penalty for exposure ϵ. The effect of ϵ on the results can be seen in Fig. 4a and in Fig. 4b under the competitive assumption.

The results for the Deceiver's Game can be seen in Fig. 5a and for Deceiver's Game X in Fig. 5b, which is played, as the other X games, under the competitive assumption. Similarly to the Bullshitter's Game, the false information factor is set to $\beta = 0.5$ because deception can be done through a combination of lying and truth-telling. Moreover, the Deceiver's game does not have a different condition for the penalty for exposure ϵ, as the game of deception is always played taking this factor into account. Additionally, the results also show the effect of the factor of deception θ.

4 DISCUSSION

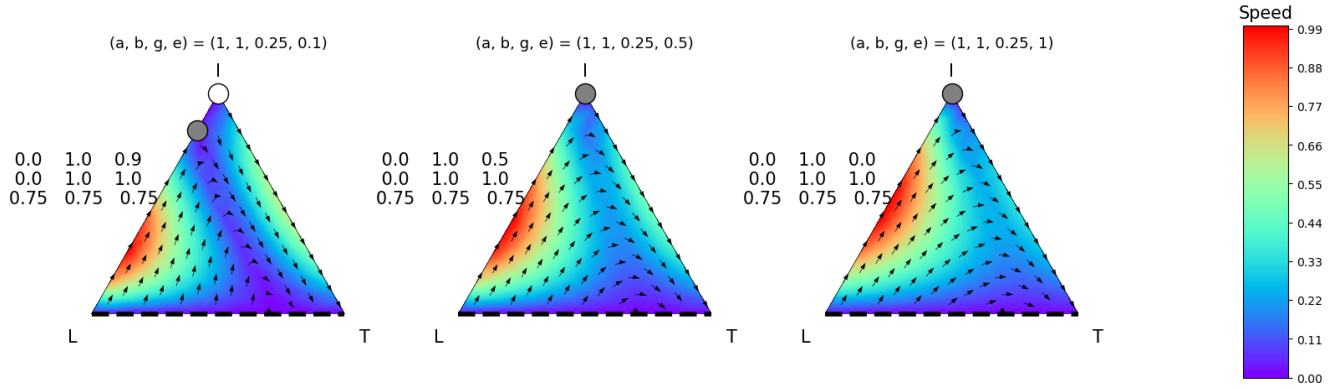
So what do the results tell us about the self-organisation of agent societies under the evolutionary pressures of dishonesty? For starters, truth-telling is not stable under evolutionary pressure in the Liar's Game. Neutral drift will lead to lying. But then investigation will root out the lying, so long as $\gamma < \beta$. Once investigation is stable, truth-telling can take hold once more. This can be seen by looking at the dashed lines given by the stable equilibrium in the Liar's Game which tell us that the agent society will evolve into a population of investigative, truth-telling, and liar agents where there will be only a small proportion of liars (Fig. 1a). This result reflects the findings by [24] whose empirical findings suggest the prevalence of a few prolific liars. Otherwise, under the competitive assumption, where agents care to do better than others, lying is actually a stable equilibrium, providing agents with a strategy that gives them advantage over others (Fig. 1b). The same can be observed in the case of deceptive agents under the competitive assumption when $\theta = 0.1$ and the penalty for being exposed is $\epsilon \leq 0.5$ (Fig. 5b). This evolutionary advantage can be observed in both social context in human history [27], as well as in nature where deception need not be intentional or purely cognitive, as per the description by [25] of



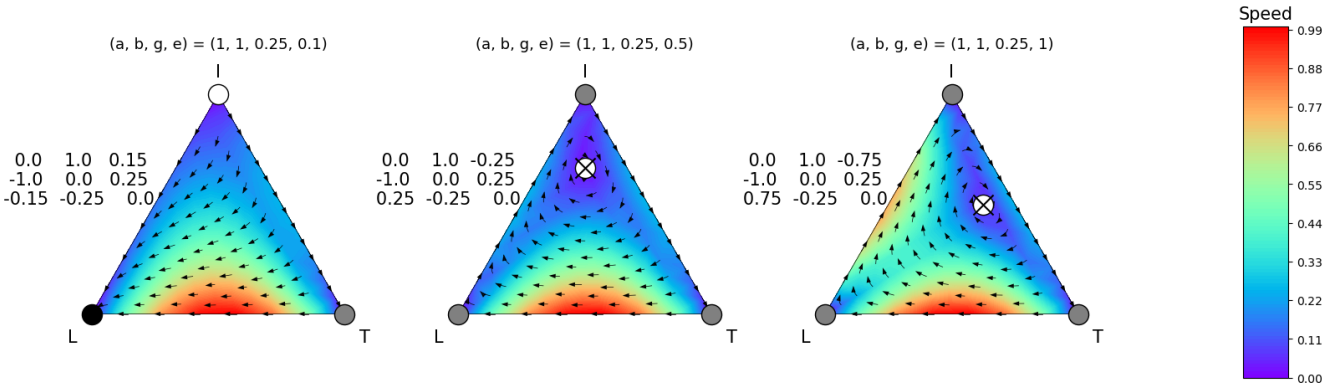
(a) Liar's Game.

(b) Liar's Game X.

Figure 1: Liar's Game under the two different assumptions without considering the penalty for exposure ϵ. Paramaters and their respective values are listed in the round brackets, where a is α, b is β, g is γ. We can observe that truth-telling is not stable in the face of lying, but that interrogation helps drive truth-telling, unless agents care to do better than others and lie.

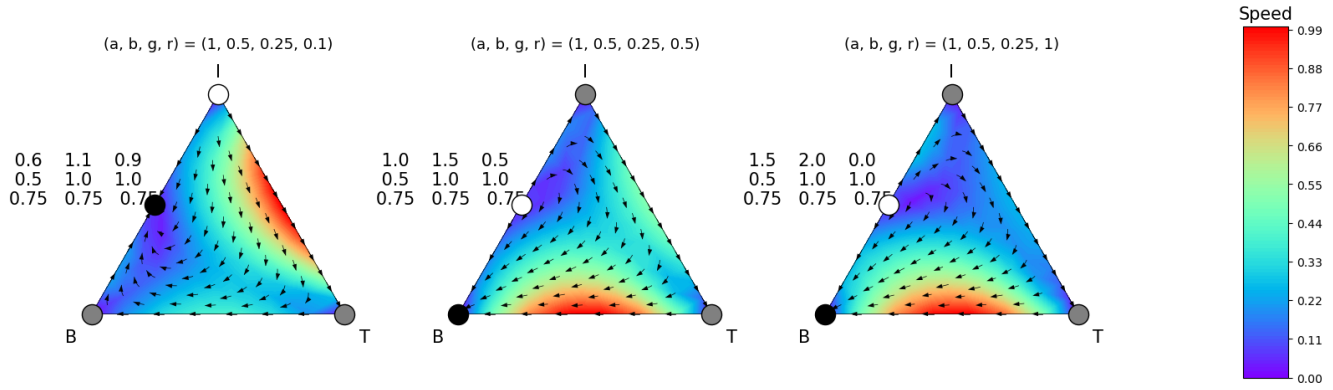


(a) Liar's Game with ϵ.

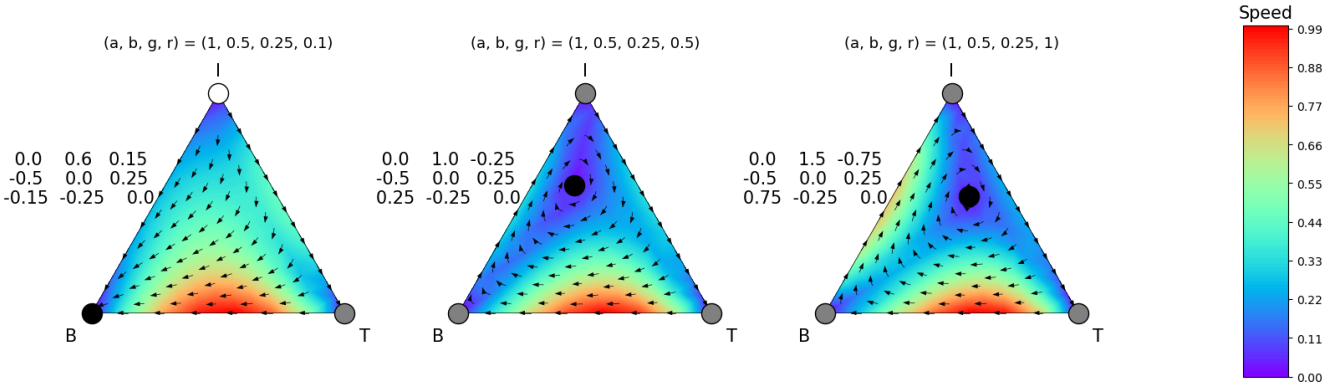


(b) Liar's Game X with ϵ.

Figure 2: Liar's Game under the two different assumptions considering the penalty for exposure ϵ and varying its numerical values. Paramaters and their respective values are listed in the round brackets, where a is α, b is β, g is γ, and e is ϵ. We can observe that the penalty for exposure drives agents towards truth-telling and towards truth-telling and investigation under the competitive assumption - unless the penalty is very low.



(a) Bullshitter's Game.



(b) Bullshitter's Game X.

Figure 3: Bullshitter's Game under the two different without considering the penalty for exposure. Paramaters and their respective values are listed in the round brackets, where a is α, b is β, g is γ, and r is ρ. We can observe that if agents care about doing better than others and $\epsilon \geq 0.5$, then they prefer not to always bullshit and are driven towards cooperative communication.

the evolutionary advantages of deceptive traits given to different animal and plant in different survival and reproductive contexts.

Under the competitive assumption in the Liar's Game, when the penalty for exposure is low $\epsilon = 0.1$, lying is dominant as a stable equilibrium. However, if the penalty for exposure is greater than the investigation factor (i.e., the cost of fact checking), e.g. when $\epsilon \geq 0.5$, then we can also observe a non-stable cycle forming which drifts the population towards investigation and truth-telling (Fig. 2b).

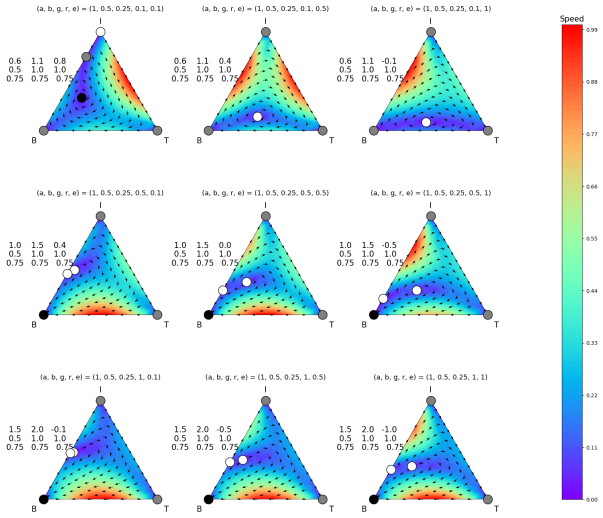
In the case of the Bullshitter's Game, we can observe that reputation gain drives the stability of bullshitting, e.g. when $\rho \geq 0.5$ (Fig. 3a). Even when the penalty for exposure is considered, bullshit is still dominant when $\rho \geq 0.5$ (Fig. 4a). However, in the X game, the penalty for exposure actually plays a significant role, driving the population in a cycle of stable equilibrium prevalent with Investigators and Cooperators for $\epsilon = 1$ (Fig. 4b).

Perhaps the most interesting result is that shown in the Deceiver's Game, where the presence of deception seems to promote the stability of truth-telling driven by investigation as a saddle-point (Fig. 5a). This is intuitive, as deception is more difficult and complex to perform than either lying or bullshitting. However, as mentioned before, deception in Game X does provide evolutionary advantages when the cost for deception is low (Fig. 5b). This can happen, for instance, when the deceivers are able to reason about the minds of others, which would force investigators to respond by out-smarting deceivers to increase the cost of deception. This kind of effect would eventually lead to a mentalisation arms race, as shown in [18].

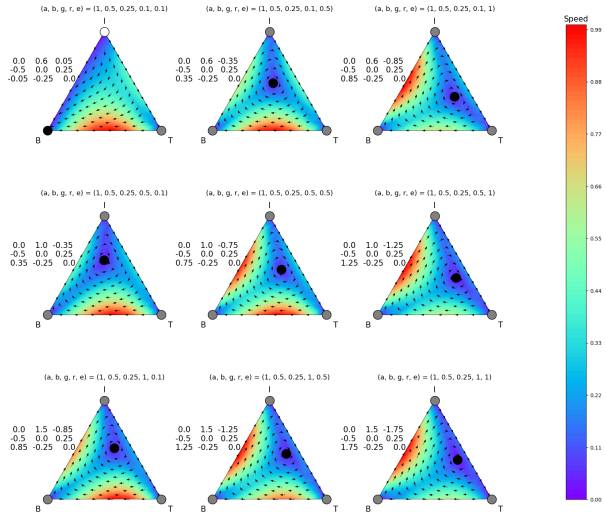
5 CONCLUSIONS

In this paper we have presented three dishonesty games based on the definitions of lying, bullshitting and deception by [14], and used evolutionary agent-based simulation approach similar to the one described by [16] to study their dynamics in agent societies. Our results[2] show important differences in the evolutionary dynamics of agent societies under the pressures imposed on them by the three different forms of dishonesty. The main differences that we have observed mainly emerge by either (i) **Self-Interest**, namely assuming that agents care about the payoffs they receive, or (ii)

[2]**Supplementary material** with code for reproducing the results and animated plots - https://osf.io/4eg35/.



(a) Bullshitter's Game with ϵ.



(b) Bullshitter's Game X with ϵ.

Figure 4: Bullshitter's Game under the two different assumptions considering the penalty for exposure ϵ and varying its numerical values. Parameters and their respective values are listed in the round brackets, where a is α, b is β, g is γ, r is ρ and e is ϵ.

Competitiveness, i.e. assuming that agents care if they perform better than others, irrespective of their own payoffs.

As with any ABM approach, there are limitations. One main limitation is that we do not explicitly model and represent the computations of socio-cognitive factors such as trust, and how cognitive load, communication and investigation skill change with the population as others, see [21]. We plan to address this in the future by modelling truth-telling, investigation, lying, bullshitting, and deception in 5-strategy n-player games, considering how these parameters are dynamically computed inside the game model. A further area for future exploration is the case when agents believe that another agent is attempting to deceive them, when this is in fact not the case. Such situations can arise, for example, in explainable AI [1].

Given the advancement of AI, our societies will increasingly have to deal with other 'kinds' of agents, each with their own sets and degrees of deceptive traits and capabilities [11]. It might very well be that the emergence of complex and autonomous artificial agents capable of deception and deception detection will create new evolutionary pressures on our species and on our socio-technical societies where humans and machines will interact physically, socially, and culturally.

By having a better understanding of dishonesty form an evolutionary perspective, we have the potential to better adapt to newer AI technologies and ensure that their widespread adoption is done in a manner where people are informed about the risks of and propensity for dishonest behaviour. This would also enable us to return the evolutionary pressures on deceptive AI technologies themselves, potentially managing dishonest AI in a systemic way. Yet, the recent literature in evolutionary ABMs and our breaking down of dishonest strategies with replicator dynamics is just a small step forward, and we still have much to understand about the human-AI ecosystems to build a comprehensive socio-cognitive computational theory of both trust and deception, as proposed a long time ago in the agents and multi-agent systems community by Castelfranchi and Tan [4].

Most importantly we should aim to re-focus the mainstream AI community's attention from purely behavioural aspects of AI to richer, more complex socio-cognitive aspects, that include metacognitive properties such as reflection [9] and Theory of Mind [18]. Finally, as members of increasingly hybrid societies, we must continuously reflect on the meaning of the term 'deceptive AI', and how it relates to our human condition as part of socio-technical ecosystems [19].

ACKNOWLEDGMENTS

This project was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Research Fellowship program.

This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

We would like to give special thanks to the Reviewers for taking the time and effort necessary to review this manuscript.

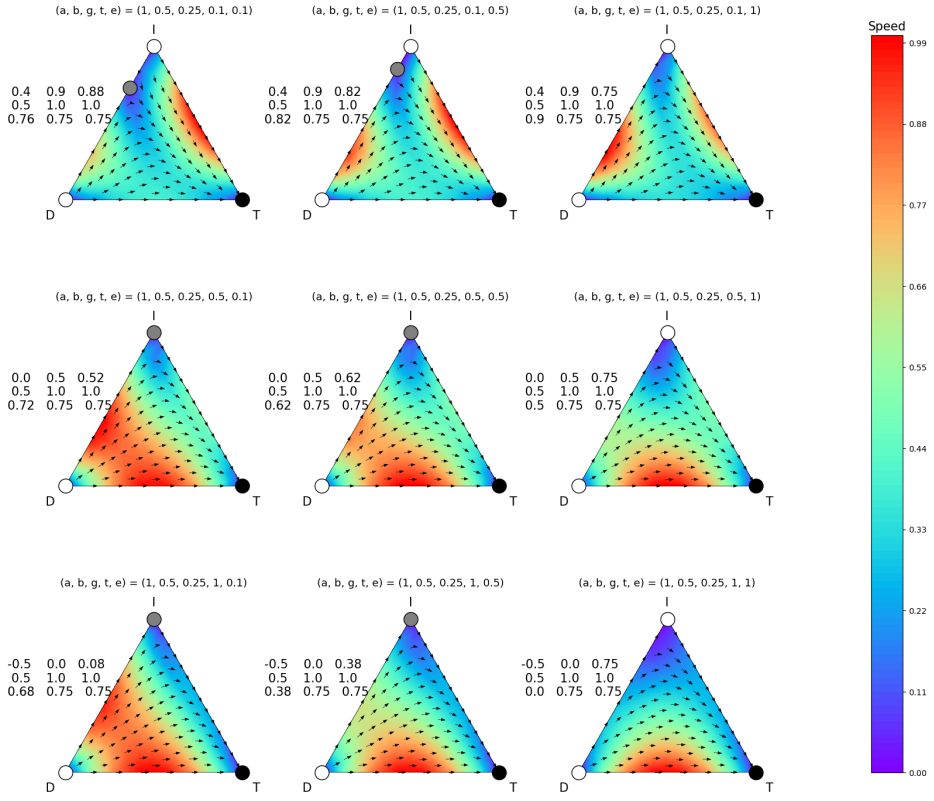
REFERENCES

[1] Zahra Atf and Peter R. Lewis. 2023. Human Centricity in the Relationship Between Explainability and Trust in AI. *Technology and Society Magazine* 42, 4 (2023), 66–76.

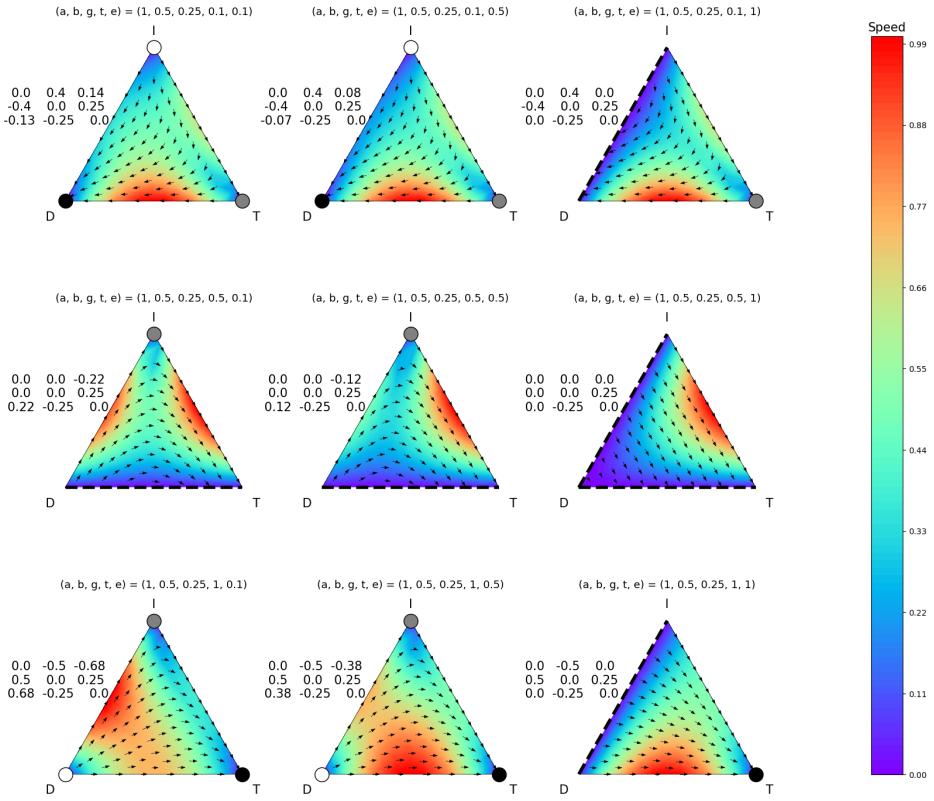
[2] Charles F Bond and Michael Robinson. 1988. The evolution of deception. *Journal of nonverbal behavior* 12 (1988), 295–307.

[3] Sarah F Brosnan and Redouan Bshary. 2010. Cooperation and deception: from evolution to mechanisms. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 1553 (2010), 2593–2598.

[4] Cristiano Castelfranchi and Yao-Hua Tan. 2001. *Trust and deception in virtual societies.* Springer.



(a) Deceiver's Game.



(b) Deceiver's Game X.

Figure 5: Deceiver's Game under the two different assumptions which include the penalty for exposure ϵ, and deception factor θ, whose numerical values are varied. Parameters and their respective values are listed in the round brackets, where a is α, b is β, g is γ, t is θ and e is ϵ. We can observe that deception can only dominate under the competitive assumption, where we also observe that truth-telling is a saddle point in which agents realise deception can provide better payoffs.

[5] Cristiano De Faveri and Ana Moreira. 2016. Designing adaptive deception strategies. In *2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. IEEE, 77–84.

[6] Yucheng Dong, Yuxiang Fan, Haiming Liang, Francisco Chiclana, and Enrique Herrera-Viedma. 2019. Preference evolution with deceptive interactions and heterogeneous trust in bounded confidence model: A simulation analysis. *Knowledge-Based Systems* 175 (2019), 87–95.

[7] Harry G Frankfurt. 2005. *On bullshit*. Princeton University Press.

[8] Carlo Kopp, Kevin B Korb, and Bruce I Mills. 2018. Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to" fake news". *PloS one* 13, 11 (2018), e0207383.

[9] Peter R. Lewis and Ştefan Sarkadi. 2024. Reflective Artificial Intelligence. *Minds and Machines* (2024). In Press.

[10] JE Mahon. 2014. History of Deception: 1950 to the Present. *Encyclopedia of Deception* (2014), 618–619.

[11] Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. 2021. Characterising Deception in AI: A Survey. In *Deceptive AI*, Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney (Eds.). Springer International Publishing, Cham, 3–16.

[12] Steven A McCornack, Kelly Morrison, Jihyun Esther Paik, Amy M Wisner, and Xun Zhu. 2014. Information manipulation theory 2: a propositional theory of deceptive discourse production. *Journal of Language and Social Psychology* 33, 4 (2014), 348–377.

[13] Inom Mirzaev, Drew FK Williamson, and Jacob G Scott. 2018. egtplot: A Python Package for 3-Strategy Evolutionary Games. *bioRxiv* (2018), 300004.

[14] Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. 2018. Lies, Bullshit, and Deception in Agent-Oriented Programming Languages. In *Proceedings of the 20th International TRUST Workshop @ IJCAI/AAMAS/ECAI/ICML*. CEUR Workshop Proceedings, Stockholm, Sweden, 50–61.

[15] Gordon Pennycook, James Allan Cheyne, Nathaniel Barr, Derek J Koehler, and Jonathan A Fugelsang. 2015. On the reception and detection of pseudo-profound bullshit. *Judgment and Decision making* 10, 6 (2015), 549–563.

[16] Steve Phelps, Kai Cai, Peter McBurney, Jinzhong Niu, Simon Parsons, and Elizabeth Sklar. 2008. Auctions, evolution, and multi-agent learning. In *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning: 5th, 6th, and 7th European Symposium, ALAMAS 2005-2007 on Adaptive and Learning Agents and Multi-Agent Systems, Revised Selected Papers*. Springer, 188–210.

[17] Chiaki Sakama, Martin Caminada, and Andreas Herzig. 2010. A logical account of lying. In *European Workshop on Logics in Artificial Intelligence*. Springer, 286–299.

[18] Stefan Sarkadi. 2023. An Arms Race in Theory-of-Mind: Deception Drives the Emergence of Higher-level Theory-of-Mind in Agent Societies. In *4th IEEE International Conference on Autonomic Computing and Self-Organizing Systems ACSOS 2023*. IEEE Computer Society.

[19] Ştefan Sarkadi. 2023. Deceptive AI and Society. *IEEE Technology and society magazine* 42, 4 (2023), 77–86.

[20] Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, and Martin D. Chapman. 2019. Modelling Deception using Theory of Mind in Multi-Agent Systems. *AI Communications* 32, 4 (2019), 287–302.

[21] Ştefan Sarkadi, Alex Rutherford, Peter McBurney, Simon Parsons, and Iyad Rahwan. 2021. The evolution of deception. *Royal Society Open Science* 8, 9 (2021), 201032.

[22] Ştefan Sarkadi. 2024. Self-Governing Hybrid Societies and Deception. *ACM Transactions on Autonomous and Adaptive Systems* (2024).

[23] William A Searcy and Stephen Nowicki. 2010. *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems: Reliability and Deception in Signaling Systems*. Princeton University Press.

[24] Kim B Serota and Timothy R Levine. 2015. A few prolific liars: Variation in the prevalence of lying. *Journal of Language and Social Psychology* 34, 2 (2015), 138–157.

[25] Martin Stevens. 2016. *Cheats and deceits: how animals and plants exploit and mislead*. Oxford University Press.

[26] Douglas Walton. 2003. The interrogation as a type of dialogue. *Journal of Pragmatics* 35, 12 (2003), 1771–1802.

[27] Barton Whaley. 1982. Toward a general theory of deception. *The Journal of Strategic Studies* 5, 1 (1982), 178–192.