# An Arms Race in Theory-of-Mind: Deception Drives the Emergence of Higher-level Theory-of-Mind in Agent Societies

Ștefan Sarkadi
*Dept. of Informatics*
*King's College London*
London, UK
stefan.sarkadi@kcl.ac.uk

***Abstract*—It has been argued that using Theory-of-Mind (ToM) is a fundamental cognitive ability that underpins effective deceptive behaviour. However, we still do not have a clear understanding of the effect ToM has on deception under evolutionary pressure in self-organising multi-agent systems (MAS). To better understand it, in this paper we introduce the Evolutionary Deception Game (EDG) to model deception and investigation of deception in societies of agents with different levels of ToM. We show that the impact of higher levels of ToM is different in deceptive agents compared to agents that investigate deception. We also show the creation of a dominance cycle, i.e., an arms race in ToM where investigators need to be one level higher to dominate deceivers. Additionally, we provide new insight into the effect that important societal factors in MAS such as social learning and population size have on the dominance of strategies in the EDG. We conclude that ToM is both the solution to agent deception, as well as its defence.**

***Index Terms*—deception, deceptive AI, dishonesty, evolution, arms race, Theory of Mind**

## I. Introduction & Motivation

More than 20 years after Castelfranchi's wager [7] that machines will necessarily deceive us and each other, recent advances in AI technology emphasise more than ever the importance of research around deceptive AI systems [35].

A crucial question to understand deception in hybrid societies, where humans and machines interact, from a socio-cognitive perspective is what kind of cognitive capabilities give agents an evolutionary advantage in deception and deception detection. From an AI and multi-agent systems (MAS) perspective, deception has been defined as — *the intentional process of an agent, the Deceiver, to make another agent, the Target, believe something is true (false) that the Deceiver believes is false (true), with the aim of achieving an ulterior goal or desire* — by Sarkadi [36], who has argued that to model and understand deception from an agent-oriented perspective, we must model AI agents that are able to deceive deliberately, but in order to do so, we need to enable them to reason about the minds of their targets, i.e. enable them to use Theory-of-Mind [36].

Theory-of-Mind (ToM) *'refers to the cognitive capacity to attribute mental states to self and others. Other names for the same capacity include "commonsense psychology", "naïve psychology", "folk psychology", "mindreading " and "mentalizing"'* [15]. ToM has different 'flavours' coming from Psychology and Cognitive Science, most popular ones being TT (Theory-Theory) that represents a top-down reasoning about the mental models of others, and ST (Simulation Theory) that is a process which simulates an agent 'as if' being in another's shoes. More recently, TT and ST have converged into hybrid ToM, which describes some sort of process responsible for the internal modelling of a practical reasoning about the mental states of others [14]. In AI, similar models to the hybrid ToM have been implemented as operational semantics for MAS communication [30], as a parameter that changes the level of recursivity in strategic reasoning of agents [9], and as a combination of consequence engines and behaviour simulators in robots [43].

The Philosophical literature also emphasises that deliberate deception is enabled by the evolution of ToM [13]. This perspective is also backed by developmental Psychology, where deception is highly correlated with the ability to reason about other minds, particularly with the ability to attribute false beliefs. Deception plays a major role in testing the development of ToM in children, e.g. in the Sally-Anne Test [16]–[18], [20]. It has also been argued in Social Psychology that deception is a particular application of ToM [26]. Perhaps the most interesting claim from an evolutionary perspective about the relation between deception and ToM has been made by Dennett [11] in a commentary to [42], who described the likelihood of a cognitive *'arms race'* driven by deception and counter-deception where creative intelligent agents respond to the evolutionary pressures of their society by adapting in order to outsmart each other. This is also evidenced in the management of communicative lies, where humans engage in at least two levels of belief recursion to maintain them [12].

As argued in [40], AI agents can be designed to have even more sophisticated Theory-of-Mind than the two levels that humans usually require to maintain lies and detect lies. In AI and MAS there is valuable work at the intersection of ToM and deception [29], [33], [37], but so far, this relation has not been explored from an evolutionary agent-based perspective. To explore this relation, we adopted an evolutionary game

theoretic approach to model deception games as public goods games (PGGs) in agent societies where we treat ToM similarly to [9], as a parameter that changes the relation between the levels reasoning in competing strategies. Additionally, similarly to [34], we integrate socio-cognitive factors described in Truth-Default Theory (TDT) [26], Information-Manipulation Theory 2 (IMT2) [28], and Interpersonal Deception Theory (IDT) [5]. This type of approach has been previously used to study the evolution of deception in agent societies and to find mechanisms for re-establishing cooperation [34]. Moreover, this type of approach studies the evolutionary stability and dominance of strategies that can be adopted by agents of a fixed population and we also use it in this paper for showing dominance cycles of strategies [39].

We consider the following main strategies for playing PGGs, namely: the **Cooperator** that uses and maintains public knowledge by sharing truthful information; the **Defector** that exploits the public knowledge without sharing any type of information for maintaining the public knowledge; the **Deceiver** that exploits the public knowledge and pretends to be a cooperator by 'maintaining' it through sharing inaccurate, misleading or false information. Finally, the **Investigator** that not only contributes with its own truthful information to the public knowledge, but also puts effort into punishing Defectors and interrogating and punishing Deceivers. Thus, in this context the Deceiver not only targets the other agents who use the public knowledge, but must also deceive the Investigator that is aiming to uncover the deception. These strategies are elaborated in Section II-D.

The research question we aim to answer is: *Does deception drive the emergence of ToM in self-organising MAS? And/or the other way round?* To begin answering our question, we first introduce the Evolutionary Deception Game (EDG). Then, we describe the setup and analyse the results from four experiments. We start with a baseline EDG model to show the dominance of cooperation when agents do not have the ability to use ToM as an explicit mechanism to gain evolutionary advantages. Afterwards, we introduce a ToM mechanism for agents to reason about the other agents' deception and investigation strategies in the EDG model, and then we proceed to alternatively increase the levels of ToM for Deceivers and Investigators. This approach allows us to check if a dominance cycle is created between Deception and Investigation when the levels of ToM are increased.

## II. The Evolutionary Deception Game

The aim in this paper is to introduce the modelling of ToM in a an evolutionary framework based on social learning theory [3]. We know from [34] that if agents are able to select a strategy that represents the decentralised interrogation and punishment of free-riding, then this makes cooperation dominant in self-organising agent societies under the condition of strong social learning. What makes our contributions novel compared to [33] and [34] is that our results address the scaling up of the effects of both deception and ToM in agent societies, to not only show the effects of ToM inside the minds of individual agents, but also on large self-organising agent populations.

We define the evolutionary deception game (EDG) as an tuple whose parameters are defined in Table I. The EDG is played with the following set of strategies $S = \{C, D, Dec, I, L\}$, where D is defection, C is cooperation, Dec is deception, I is investigative cooperation (Investigators), and L is non-participation (Loners).

For our EDG model, each strategy, except for Non-Participation, falls into one of the meta-strategies of playing PGGs, namely Cooperation and Free-Riding. The **cooperation** meta-strategy, which requires an agent to make a contribution to the social good, includes Cooperation (C) and Investigation (I). The **free-riding** meta-strategy, which requires an agent to not contribute anything to the social good while enjoying the benefits of the social good, includes Defection (D) and Deception (Dec). Hence, for each strategy $i \in S \setminus L$, an EDG payout is received:

$$Payout = c \times r \times \frac{N - N_{FR} - N_L - 1}{N - N_L - 1} \tag{1}$$

In Eq.1 where an EDG is played by a fixed population with $N$ agents, $N_{FR}$ represents the total number of Free-Riders, and $N_L$ represents the total number of Loners (Non-Participants). This payout is consistent with the previous evolutionary models based on PGGs [1], [21], [31], [34], [38].

### A. Evolutionary Dynamics

Algorithm 1 provides an overview of the evolutionary deception game (EDG). Each round $t$ of the game corresponds to a public-goods game being played over time $T$. According to the composition of the population $k_t = [k_{t,C}, k_{t,D}, k_{t,Dec}, k_{t,I}, k_{t,L}]$ at time $t$, i.e. how many agents play each strategy $i$, a payoff $\Pi_{t,i}$ is computed and dynamics of *mutation* and *imitation* may take place.

**input :** $k_0, s, \mu, T$
**output:** $k_T$
*t = 1*;
**do**
  $\Pi_t =$ `playRound(`$k_t$`)`
  **if** `random(0,1)` $< \mu$ **then**
    $k_{t+1} \leftarrow$ `mutate(`$k_t$`)`;
  **else**
    $k_{t+1} \leftarrow$ `imitate(`$k_t, \Pi_t, s$`)`;
  *t ++*
**while** $t < T$;
**return** $k_T$

**Algorithm 1:** The Evolutionary Deception Game EDG, where each round corresponds to $t$.

*a) Mutation:* In each round $t$ of the EDG, an agent is randomly selected to randomly change its strategy according to the *mutation rate* $\mu$.

| Parameter | Value | Definition |
|---|---|---|
| | | **Game parameters** |
| $S \neq \emptyset$ | | a non-empty set of strategies |
| $N$ | 100 | population size |
| $N_i$ | $\leq N$ | the number of agents using strategy $i$ |
| $M$ | 5 | the number of agents selected to play the EDG at a given iteration from a population of $N$ |
| $T$ | $10^5$ | number of iterations of an EDG |
| $c$ | 1 | contribution to the EDG |
| $r$ | 3 | a multiplication factor that acts as an incentive for contributing $c$ |
| $s \geq 0$ | $10^3$ | social learning (imitation) strength |
| $\mu \geq 0$ | 0.001 | exploration rate (inclination of agents to randomly adopt another strategy) |
| $\sigma$ | 0.3 | non-participation payoff |
| | | **Cooperation and Defection** |
| $\beta$ | 1 | cost of cooperating (also investigation reward) |
| $\epsilon$ | 0.7 | the cost of defecting |
| | | **Deception** |
| $\Gamma$ | 0.8 | punishment for deception |
| $\alpha$ | 1 | worthiness of deception |
| $\pi \in [0, 1]$ | 0.1 | worthiness of deception discount |
| $\psi \in [0, 1]$ | 0.5 | the skill of Deceivers |
| $\gamma \in [0, 1]$ | $1 - \psi$ | the leakage |
| $DecToM$ | $1 \vee 2$ | the level of ToM of Deceivers |
| | | **Investigation** |
| $\lambda$ | 0.7 | the cost of punishing Defectors |
| $cost_\Gamma$ | 0.5 | the cost of punishing Deceivers |
| $\zeta$ | 0.5 | the cost of interrogating agents |
| $\delta \in [0, 1]$ | 0.1 | investigation reward discount |
| $\phi \in [0, 1]$ | 0.5 | the skill of Investigators |
| $\theta \in [0, 1]$ | $1 - \phi$ | cognitive load of Investigators |
| $IToM$ | $1 \vee 2$ | the level of ToM of Investigators |

TABLE I: EDG parameters. The game parameter values were taken from [1], [34], as well as the parameters for costs for punishment, defection and cooperation. The deception and investigation parameters were taken from [34], except for newly added parameters such as the worthiness of deception and its discount factor, the investigation reward and its discount, and the levels of ToM for both deception and investigation. $ITOM$ and $DecToM$ vary with each experiment. Similarly, $s$ and $N$ are varied in the sensitivity analysis.

*b) Imitation:* In each round $t$ of the EDG when mutation does not occur, *social learning* happens. In this case, two agents are randomly selected and one adopts the strategy of the other one according to the social learning strength parameter $s$ and the payoffs of the two agents' strategies. In particular, given two sampled agents $a$ and $b$ and the difference between their payoffs $\Delta = \Pi_{t,a} - \Pi_{t,b}$, if a number randomly selected between 0 and 1 is lower than $f(s, \Delta)$, then $a$ adopts $b$'s strategy, otherwise vice-versa. According to the tradition in evolutionary games (e.g., see [38], we define $f(s, \Delta)$ as follows:

$$f(s, \Delta) = \frac{1}{1 + e^{-s\Delta}}$$

If $s$ is relatively high, then the learning capabilities of the agents are stronger and it is more likely that the agent that receives the lowest payoff adopts the strategy of the better off agent; if $s$ is low, then the learning is weaker and an agent may adopt a worse strategy. This process simulates Bandura's description of social learning dynamics [3].

### B. Deception

In this paper we explore a value-based [31] (as opposed to content-based) model of how ToM affects Deception and Investigation strategies in a MAS. In a content-based model, an explicit algorithm would be used to represent step-by-step how ToM is used by an agent for deception, as in [33], whereas in a value-based model, the focus is on how a specific trait of the ToM, the level of ToM in our case, affects the evolutionary value of the agent's reasoning and/or behaviour in a MAS.

We know that ToM helps Deceivers reason about the minds of other agents in order to estimate whether the information they communicate will lead the others to infer the false beliefs the Deceivers desire them to have. According to IDT, IMT2, and TDT, (i) the amount of information involved in the process of mentalisation of other agents' minds, (ii) the socio-cognitive factors such as the cognitive load arising from mentalising, (iii) the context-based truth bias determined by the trust between agents of a society, and (iv) the Deceiver's communicative skill in managing information leakage will affect the Deceiver's success.

To compute the Deceiver's cognitive load ($cogload$) we take into account how many agents need to be deceived ($N - N_L$ - everyone except for non-participants) minus 1 (the agent performing the deception), the levels of trust between agents, where high levels of trust make it easier for the Deceiver to deceive other agents as they would be more trusting, and the skill of the Deceiver $\psi$. According to IDT, $cogload$ increases when the number of agents that need to be deceived increases, and decreases when trust and skill increase [4].

$$cogload = (N - N_L - 1) \times (1 - trust) \times (1 - \psi)$$

We consider trust between agents to be given by the proportion of Cooperative agents that will not free-ride:

$$trust = \frac{N - N_D - N_{Dec}}{N}$$

The risk of Deceivers getting caught is influenced by the prevalence of Investigators $N_I$, multiplied by the amount of information the Deceivers leak $\gamma$, and the severity of punishment Deceivers would receive if caught.

$$\gamma_{risk} = N_I \times (\gamma + \Gamma)$$

However, the ToM process becomes recursive for Deceivers when they need to face Investigators, that are skeptical agents whose aim is to catch the Deceivers, and perhaps even punish them for attempting deception. To succeed, the Deceiver has to know what the Investigator knows about what the Deceiver knows, such that the Deceiver avoids being caught when providing misleading information. To model this battle between levels of recursivity we introduce $DecToM$ and $ITOM$ to represent the level of recursion for the Deceiver's and, respectively, the Investigator's mentalisation processes of what the Deceiver is thinking. Hence, the cost of deception is computed by considering the processes above given the difference between the level of ToM of the Investigator and the level of ToM of the Deceiver.

$$Cost_{Dec} = (cogload + \gamma_{risk})^{ITOM} - (cogload + \gamma_{risk})^{DecToM}$$

The reward of a Deceiver is given by the worthiness of deception $\alpha$, that is how much the Deceiver thinks its dishonest actions would be beneficial in the long-run. This worthiness of deception is discounted by $\pi$.

$$R_{Dec} = \pi \times \alpha$$

### C. Investigation

For the Investigator, ToM will affect the cost of interrogating potential Deceivers (this includes Cooperators - remember that Deceivers pretend to be Cooperators), as well as reduce the cost of punishing Defectors by being able to outsmart them.

The cost of punishing Defectors $\lambda$ is proportional to the number of Defectors in the population $N_D$. However, the Investigators are able to reason about the minds of Defectors and prepare in advance for their free-riding behaviour. Hence, the greater their ToM level, the lower the cost of punishing Defectors.

$$cost_P = \lambda \times N_D^{\frac{1}{ITOM}}$$

Catching and punishing Deceivers might prove to be a significantly harder task for Investigators, as it requires extensive peer-to-peer interrogation. However, the leakage of the Deceivers $\gamma$ (the amount of information Deceivers give away about their deception) influences how likely is for Investigators to pay the cost of punishing Deceivers. To make it even more difficult, Investigators must not just interrogate Deceivers, but also Cooperators $N_C + N_{Dec}$ and pay the cost of interrogation $\zeta$ for each interrogated agent - remember that Deceivers pretend to be Cooperators. The cost of interrogating and punishing Deceivers is also influenced by the Investigator's interrogation skill $\phi$, which determines the cognitive load of the Investigator.

$$cost_{Int} = \theta\left(\gamma \times cost_\Gamma \times N_{Dec} + \zeta(N_C + N_{Dec})\right))$$

When Investigators face Deceivers capable of mentalising their own investigation process, then the levels of recursion in ToM become active. To succeed in deception detection, the Investigator has to know what the Deceiver knows about what the Investigator knows, such that the Investigator prevents the Deceiver from misleading the Investigator about providing misleading information to others. Same as before, $DecToM$ and $IToM$ represent the level of recursion for the Deceiver's and, respectively, the Investigator's mentalisation processes, but in this case, of what the Investigator is thinking. Hence, the cost of Investigation $Cost_I$ is the sum between the cost of punishing Defectors $cost_P$ and the the cost of interrogating Deceivers accounting for the differences in the levels of ToM for Deceivers and Investigators.

$$Cost_I = \left(cost_{Int}^{DecToM} - cost_{Int}^{IToM}\right) + cost_P$$

Finally, the reward for Investigation $R_I$ is given either by the fixed reward $\beta$ (a fixed amount) discounted with $\delta$ if there are no Investigators in the population, or by multiplying $\beta$ with the number of Defectors and Deceivers in the population that need to be interrogated and/or punished, and divided by the number of Investigators in the population.

$$R_I = \begin{cases} \beta \times \frac{N_D + N_{Dec}}{N_I} & \text{if } N_I > 0 \\ \beta \times \delta & \text{otherwise} \end{cases}$$

### D. Strategy Payoffs

A fundamental assumption we make is that the EDG is a type of voluntary public-goods game (PGG). Hence, to compute the payoffs for the strategies involved in the EDG, we need to take into account the probability that all other $M-1$ sampled individuals are Loners, namely:

$$P_\sigma = \frac{N_L!(N-M)!}{(N_L - M + 1)!(N-1)!} = \frac{\binom{N_L}{M-1}}{\binom{N-1}{M-1}} \quad (2)$$

where $N_L$ is the number of Loners in the population.

*a) Non-Participation ($L$) Payoff:* The role of Loners is twofold. Their main role as Loners is to represent agents who choose not to participate in the EDG. Second, the presence of Loners also gives the other strategies the chance to invade the population, e.g. to ensure neutral drift (see [21]).

$$\Pi_L = \sigma \quad (3)$$

*b) Cooperation ($C$) Payoff:* The Cooperator's role is to pay the contribution to the public good to receive the EDG payout (to cooperate) and pay a fee for maintaining the existence of Investigators in the population.

$$\Pi_C = P_\sigma \times \sigma + (1 - P_\sigma) \times (Payout - c) - \beta \frac{M-1}{N-1} \quad (4)$$

*c) Defection ($D$) Payoff:* The Defectors's role is to avoid paying the contribution to the public good but receive the EDG payout (to free ride) and pay a penalty if it is found and punished by Investigators.

$$\Pi_D = P_\sigma \times \sigma + (1 - P_\sigma) \times Payout - \epsilon \times N_I \frac{M-1}{N-1} \quad (5)$$

*d) Deception ($Dec$) Payoff:* The role of the Deceiver is to avoid paying the contribution to the public good (to free-ride) while receiving the EDG payout by pretending it is a Cooperator, and to avoid being caught and punished by out-reasoning the Investigators.

$$\Pi_{Dec} = P_\sigma \times \sigma + (1 - P_\sigma) \times Payout + (R_{Dec} - Cost_{Dec}) \frac{M-1}{N-1} \quad (6)$$

*e) Investigation ($I$) Payoff:* The role of the Investigator is to pay the contribution to the public good to receive the EDG payout (to cooperate), while catching and punishing Defectors and investigating and punishing Deceivers by out-reasoning them.

$$\Pi_I = P_\sigma \times \sigma + (1 - P_\sigma) \times (Payout - c) + (R_I - Cost_I) \frac{M-1}{N-1} \quad (7)$$

## III. Experimental Study

To answer our RQ, we formulated a set of four hypotheses, namely H.Base, H1, H2, and H3, and designed four experimental setups in order to test each of the hypotheses. To do this, we run extensive individual-based simulations and report the long-run avg. frequencies. For statistical significance, in each setup we run a sample size of 100 individual runs. Bars represent Means of Long-run Avg. frequencies and error bars represent $\pm 1$ SD from the mean. **Each simulation run** starts from all Defectors ($D$) as in related works [1], [34], [38]. To compare the success of strategies (the means of the long-run avg. frequencies) of each EDG we performed One-way ANOVA and pairwise t-tests between $I$, $Dec$ and every other strategy. Results are summarised in Table II.

Further, we vary social learning strength $s$ and population size $N$ to show their effects on the selection of strategies. The plots for the sensitivity analysis show the lines representing the avg. frequencies of the strategies for each parameter range, and the shaded areas $\pm 1$ standard deviation from the mean. Again, for statistical significance, sample sizes of 100 individual runs for each parameter range were used - See Figures 6, 7, 8, 9, 10, 11.

### A. Baseline (No ToM)

**H.Base** With no ToM, Investigators dominate Deceivers, Defectors, Cooperators, and Loners, and stabilise cooperation.

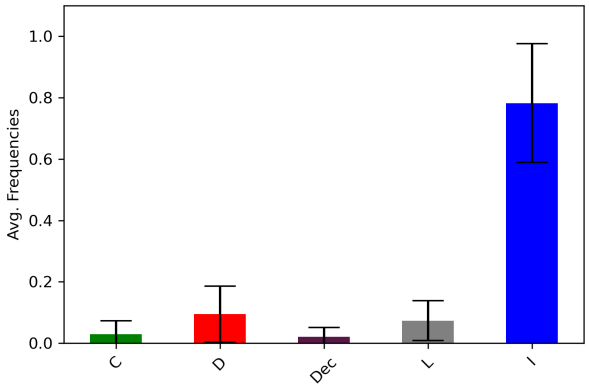



Fig. 1: H.Base - Barplot for baseline setup where $ITOM = 0$ and $DecToM = 0$. Means of long-run avg. frequencies [Strategy (Mean, SD)]: C (0.028; 0.044), D (0.094; 0.091), Dec (0.02; 0.029), **I (0.78; 0.19)**, L (0.073; 0.064).

We start our experiments by designing and testing a baseline model without ToM similar to the one from [34], but with different assumptions, namely: (i) there is a discounted reward mechanism $R_I$ of the Investigator, and that (ii) Deceivers should consider the existence of other Deceivers in the population when they reason about trust.

*a) Setup:* ToM Levels: $ITOM =$ NO, $DecToM =$ NO. Parameters from Table I.

*b) Result:* H.Base **confirmed**. The long-run avg. frequency of $I$ is greater than all other strategies, thus Investigation is evolutionary stable. One-way ANOVA and t-test show that differences in the avg. frequencies between agents are statistically significant, with all p-values $\leq 0.001$, and F-statistic $= 1010.060$. Barplots, Means and SDs for strategy frequencies are reported in Figure 1.

*c) Discussion:* This bench-marking experiment shows that deception cannot become evolutionary stable where agent societies can organise themselves to detect and punish deception in a decentralised manner. This results is consistent with the findings in [34] under strong social learning. By changing the trust model, we show that the same holds under the assumption that social learning is limited, e.g. agents can make mistakes when imitating the strategy with the higher payoff.

### B. Introducing ToM

**H1** If agents use the same order of ToM, this offers an evolutionary advantage to Deceivers in the face of Investigators.

Having shown in the previous setup that without ToM, deception is kept at bay by evolutionary stable Investigators (a cooperative strategy), we now introduce a model that takes into account the ability of agents to model other minds. In this setup, we aim to test whether this ability has an effect on the evolutionary stability of cooperation or deception. Both deceivers and the newly introduced investigators can reason about other agents' minds in order to deceive or investigate.

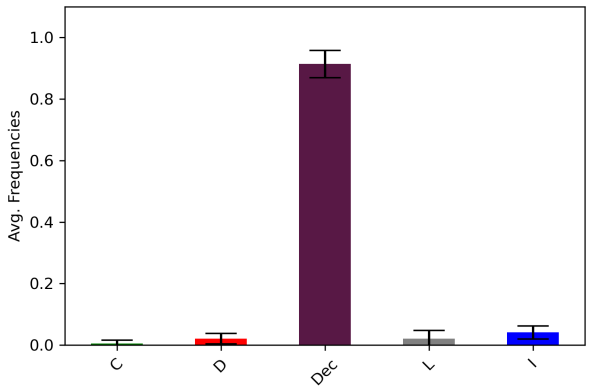



Fig. 2: H1 - Barplot when $ITOM = 1$ and $DecToM = 1$. Means of long-run avg. frequencies [Strategy (Mean, SD)]: C (0.004; 0.011), D (0.02; 0.01), **Dec (0.91; 0.04)**, I (0.04; 0.02), L (0.02; 0.02).

*a) Setup:* ToM Levels: $ITOM = 1$, $DecToM = 1$. Parameters from Table I.

*b) Result:* H1 **confirmed**. The long-run avg. frequency of $Dec$ is greater than all other strategies, thus Deception is evolutionary stable. One-way ANOVA and t-test show that differences in the avg. frequencies between agents are statistically significant, with all p-values $\leq 0.001$, and F-statistic $= 22604.0589$. Barplots, Means and SDs for strategy frequencies are reported in Figure 2.

*c) Discussion:* The ability to use ToM makes deception a dominant strategy. Given the results of the previous experiment for H.Base and the confirmation of H1 we can conclude that ToM makes deception a dominant strategy. Compared to the baseline setup, this experiment shows a clear evolutionary advantage for deceivers when they are able to use ToM. This is consistent with findings in previous works where ToM determines the success of deception in agent communication [33].

### C. Smarter Investigators

**H2** - If Investigators have a higher-order ToM than Deceivers, this stabilises cooperation.

In the previous experiment we showed that by introducing the ability to reason about other minds (ToM) for deceivers, Deception becomes the evolutionary stable strategy despite introducing the same capability for investigators. In this experiment we test what happens if we increase the order of ToM for investigators, e.g. we 'arm' them with a second order ToM, while keeping a first order ToM for deceivers.

*a) Setup:* ToM Levels: $ITOM = 2$, $DecToM = 1$. Parameters from Table I.

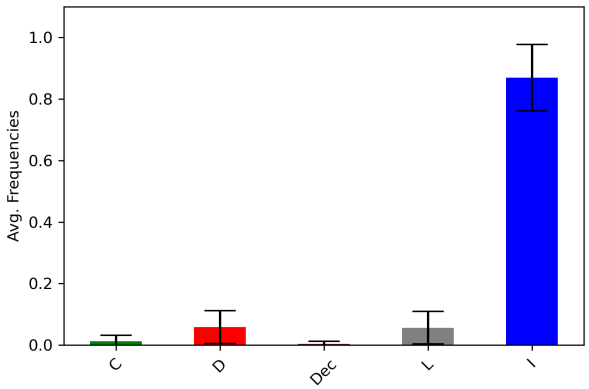



Fig. 3: H2 - Barplot when $ITоM = 2$ and $DecToM = 1$. Means of long-run avg. frequencies [Strategy (Mean, SD)]: C (0.012; 0.019), D (0.058;0.053), Dec (0.004; 0.007), **I (0.869; 0.10)**, L (0.05; 0.05).

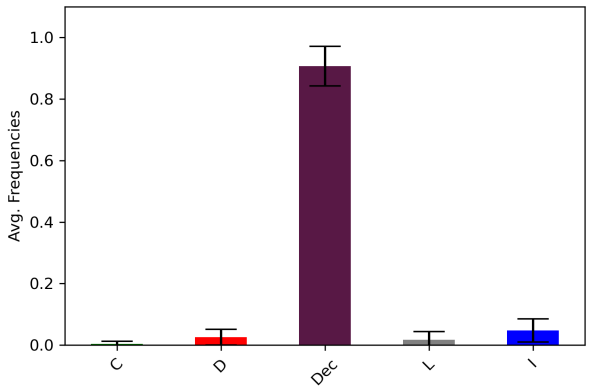



Fig. 4: H3 - Barplot when $ITоM = 2$ and $DecToM = 2$. Means of long-run avg. frequencies [Strategy (Mean, SD)]: C (0.003; 0.008), D (0.025; 0.025), **Dec (0.90; 0.06)**, I (0.047; 0.037), L (0.017; 0.026).

*b) Result:* H2 **confirmed**. The long-run avg. frequency of $I$ is greater than all other strategies, thus Investigation is evolutionary stable. One-way ANOVA and t-test show that differences in the avg. frequencies between agents are statistically significant, with all p-values $\leq 0.001$, and F-statistic $= 3949.402$. Barplots, Means and SDs for strategy frequencies are reported in Figure 3.

*c) Discussion:* Higher levels of ToM for investigators make investigative cooperation the dominant strategy. By confirming H2, our results show that in order to stabilise cooperation in agent societies, investigators need to have a higher order ToM than their deceptive counterparts. This means investigators must be able to out-think and out-reason deceivers in order to successfully catch and punish them. In other words, having higher orders of ToM is a solution for establishing and maintaining cooperation in self-organising agent societies.

### D. Arms Race

**H3** - If Deceivers match the higher-order ToM of Investigators, this stabilises free-riding.

We have showed that by arming investigators with a higher order ToM, Investigation $I$ becomes the evolutionary stable strategy. But what happens when deceivers respond by levelling up? Do they destabilise cooperation? To test this, we increased the deceivers' order of ToM to match the one of investigators, i.e., both now have second order ToM.

*a) Setup:* ToM Levels: $ITоM = 2$, $DecToM = 2$. Parameters from Table I.

*b) Result:* H3 **confirmed**. The long-run avg. frequency of $Dec$ is greater than all other strategies, thus Deception is evolutionary stable. One-way ANOVA and t-test show that differences in the avg. frequencies between agents are statistically significant, with all p-values $\leq 0.001$, and F-statistic $= 11206.436$. Barplots, Means and SDs for strategy frequencies are reported in Figure 4.

*c) Discussion:* An arms race in ToM happens where the investigators need to be one level higher to dominate deceivers. The confirmation of H.B, H1, H2, and H3 indicate that an arms race cycle is created when deceiver and investigator agents respond to each other by increasing their ToM levels. This is represented by a dominance cycle between the different setups where $I \rightarrow Dec \rightarrow I \rightarrow Dec \rightarrow ....$ This could be likely in scenarios where psychological operations and defense from disinformation happen in agent societies. One must note that the agents performing the deception will always have an advantage in the arms race, as they only need to be at the same meta-reasoning (ToM) level as the defenders, whereas the defenders (investigators) always have the harder job to detect deceptive information, expose the truth and convince their societies that disinformation is happening.

### E. Effects of Societal Factors

Cognitive factors such as social learning $s$, that is the ability of agents to identify the better strategies and learn or adapt them from others, and the size of the population $N$ can influence the emergence and stability of strategies. These societal factors create different evolutionary pressures on agents with different strategies. For instance, by varying $s$ and keeping $N = 100$, we can test how the learning ability of agents affects the adoption of strategies, e.g. how smart do agents need to be in order to select the dominant strategy. Additionally, by varying $N$ and keeping $s = 10^3$, we can explore if and when does a strategy lose or gain stability if an agent society grows, e.g. we test the strategies' scalability and sustainability. By varying the two parameters we can see how they influence the long-run avg. frequency of the modelled strategies.

*a) Result:* We report only the most significant effects for $s$ and $N$ w.r.t. strategies in the following setups (i) $ITоM = 1$ and $DecToM = 1$ (ii) $ITоM = 2$ and $DecToM = 1$. Ranges for $s \in [0, 10^3]$ and $N \in [100, 1000]$. For (i) the greater the

values for $s$, the higher the avg. frequencies of $Dec$ - see Figure 6. On the other hand, for (ii), increases in $s$ reflect an increase in the avg. frequencies for $I$ - see Figure 7. For both (i) and (ii) a value around 100 for $s$ stabilises Deception and Investigation, respectively. In terms of the sustainability of strategies in (i) and (ii) we can observe how increases in $N$ causes both $Dec$ and $I$ to become less stable and not dominant. However, $I$ does recover for larger population sizes $N \rightarrow 1000$ - see Figure 10. In both setups, Defection ($D$) gains stability in the long-run - see Figures 9 and 11.

*b) Discussion:* Apart from the clear evolutionary advantages of ToM in EDGs, it seems that both social learning and population size can have strong effects on evolutionary outcomes. We have seen in this experiment that the higher the social learning (imitation) capabilities of agents are in a population, the more likely it is for them to adopt the best strategy. In the case of (i), social learning drives the evolution of deception, while in (ii), it drives the evolution of investigative cooperation (contributing to the public good, while interrogating and punishing free-riders in a decentralised manner using models of other agents' minds). Then, we have seen that in larger populations, neither deceivers nor investigators keep their evolutionary stability. It seems that in both (i) and (ii), agents consider Deception and Investigation too demanding from socio-cognitive perspective and start Defecting. This is intuitive, as the more agents are present in a system, the more cognitive resources are necessary to either deceive or detect deception. For instance, ToM fails to scale to larger populations as agents' sampling of the world is less likely to accurately reflect the population as a whole.

## IV. Related Work

Both deception and ToM have been previously explored in agent interactions and modelled in MAS. For example, the effects of recursive ToM has been studied in agent negotiation games by [9] [8], [9]. [41] explored the effect of ToM in common resource-sharing games by extending standard approaches from opponent modelling to model ToM [41].

When it comes to the effects of ToM on deception in MAS, [33] modelled ToM in agent reasoning and communication to show its effects on deception when socio-cognitive factors from Communication Theory are present [33]. However, the respective approach to study the effects of ToM did not consider the evolutionary pressures in MAS, or of very large agent populations. On the other hand, deception has been modelled separately in evolutionary games, but very rarely by considering the socio-cognitive factors such as the ones identified by solid and relevant theories such as Truth-Default Theory [26], Interpersonal Deception Theory [4], or Information Manipulation Theory [28]. One example is the design of deceptive games for testing AI agent performance in computer games [2]. Another is defining deceptive strategies in evolutionary minority games [19]. Evolutionary models of civil violence have also been shown to present deceptive dynamics [32]. A preliminary evolutionary framework for analysis of uncertainty, vulnerability, and deception has also been attempted, but, so far, it only treats deception at a behavioural level [27].

The introduction of socio-cognitive factors such as trust and the assumption that deception is a phenomenon where the agents involved actually play different games has been modelled using hyper-game theory [24], [25]. From a system level and socio-cognitive perspective, information theoretic models of deception nicely break the costs and benefits of multiple strategies for deceiving, but lack the joint effects ToM and the socio-cognitive factors have on the strategy selection by agents [23]. Interestingly, however, socio-cognitive factors have been accounted for when modelling of self-deception as a form of internal deliberation that provides an evolutionary advantage [6], and more recently in [22].

The closest work to ours in the evolutionary modelling of deception in agent societies is that of [34], who explore how deception breaks cooperation in hybrid societies, which are societies where humans and machines interact, and how cooperation can be re-established [34]. The respective work actually takes into account socio-cognitive factors from Communication Theory, but it fails to explicitly model and account for recursively increasing levels of ToM and how factors from IMT2 such as cognitive load influence the use of ToM. Another difference is that the cost of deception in [34] was the sum of the cognitive load ($cogload$) and the risk of getting caught ($\gamma_{risk}$), whereas in our work the cost function also relies on the differences of ToM levels between deceivers and investigators. Other uses of evolutionary PGGs in agent societies have been to study the relation between cooperation and social learning [38], and the relation between cooperation and corruption [1], both in conjunction with mechanisms for punishing free-riding behaviour as in [38].

## V. Conclusion

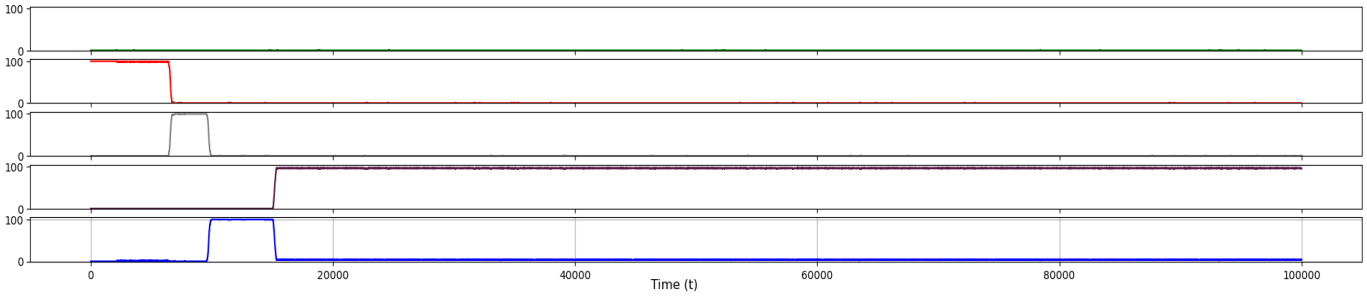
In this paper we presented the Evolutionary Deception Game (EDG), an evolutionary socio-cognitive model for studying the relation between deception and Theory-of-Mind (ToM) in self-organising agent societies. Our findings are novel, but consistent with other findings in the literature regarding the effects of ToM in multi-agent interactions [10], namely that higher levels of ToM give an advantage to agents over others and that it is crucial for deception as causing false beliefs in the minds of other agents [33]. In our case, we show that this is an evolutionary advantage for catching deceptive agents.

In conclusion, the **answer to our main RQ** is that the relation between ToM and deception creates an evolutionary cycle. Under the assumptions of our model and the selected parameters we show how ToM drives the evolutionary stability of deception, but that its emergence and evolutionary stability can cause the emergence of higher-order ToM in cooperative agents. In turn, these dynamics create an arms race cycle in ToM, i.e. in mentalizing capabilities, as previously speculated in [11].

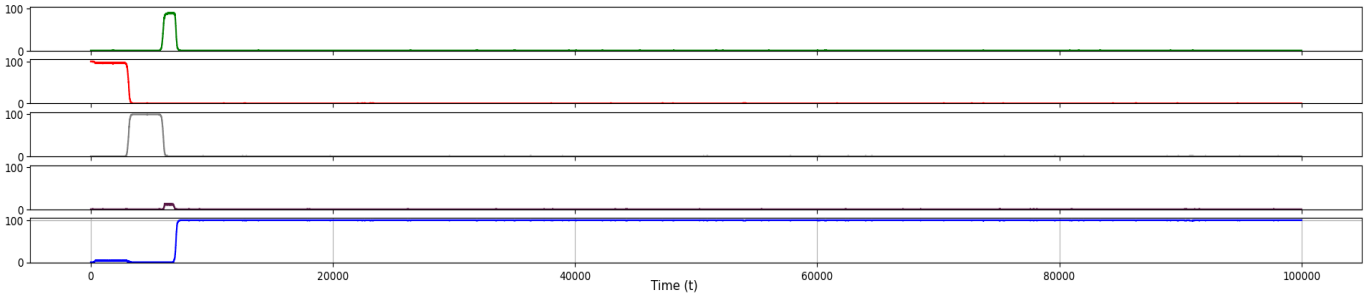
As argued in [40], AI agents can be designed to have even more sophisticated Theory-of-Mind than the two levels

| **H.Base** Benchmark (NO ToM) | **H1** DecToM 1 vs IToM 1 | **H2** DecToM 1 vs IToM 2 | **H3** DecToM 2 vs IToM 2 |
|---|---|---|---|
| **Strategy (Mn; SD)** | **Strategy (Mn; SD)** | **Strategy (Mn; SD)** | **Strategy (Mn; SD)** |
| C (0.028; 0.044) | C (0.004; 0.011) | C (0.012; 0.019) | C (0.003; 0.008) |
| D (0.094; 0.091) | D (0.02; 0.01) | D (0.058; 0.053) | D (0.025; 0.025) |
| Dec (0.02; 0.029) | **Dec (0.91; 0.04)** | Dec (0.004; 0.007) | **Dec (0.90; 0.06)** |
| **I (0.78; 0.19)** | I (0.04; 0.02) | **I (0.869; 0.10)** | I (0.047; 0.037) |
| L (0.073; 0.064) | L (0.02; 0.02) | L (0.05; 0.05) | L (0.017; 0.026) |

TABLE II: Results Table: Means and SD for all strategies in all experimental setups. Dominant strategies are highlighted in bold for each setup.




(a) H1 - $ITOM = 1$ and $DecToM = 1$.




(b) H2 - $ITOM = 2$ and $DecToM = 1$.

Fig. 5: Typical independent runs for $10^5$ iterations to showcase dynamics for H1 and H2. Y-axis represents number of agents that use a certain strategy at a specific iteration. X-axis represents the number of iterations. We can observe how Deception ($Dec$ - burgundy line) becomes evolutionary stable in 5a, whereas Investigation ($I$ - blue line) becomes evolutionary stable in 5b.

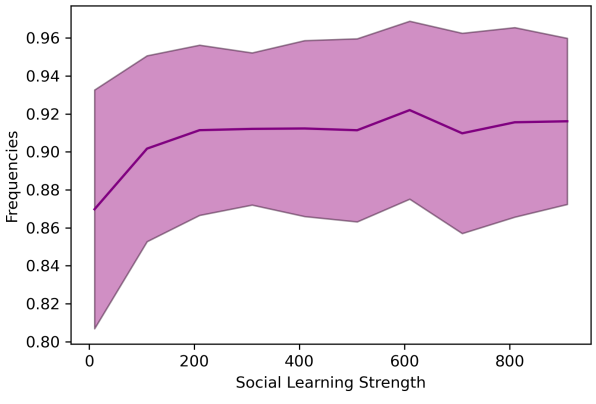



Fig. 6: **H1:** Effect of $s$ on $Dec$ for $ITOM = 1$ and $DecToM = 1$.

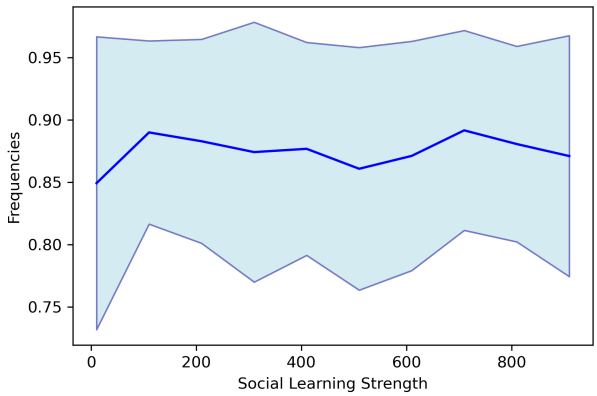



Fig. 7: **H2:**Effect of $s$ on $I$ for $ITOM = 2$ and $DecToM = 1$.

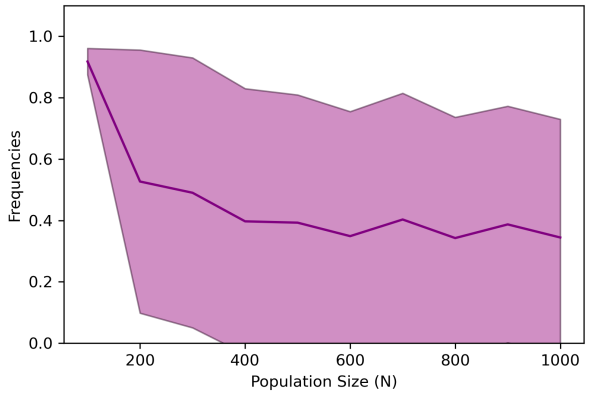



Fig. 8: **H1:** Effect of $N$ on $Dec$ for $ITоM = 1$ and $DecToM = 1$.

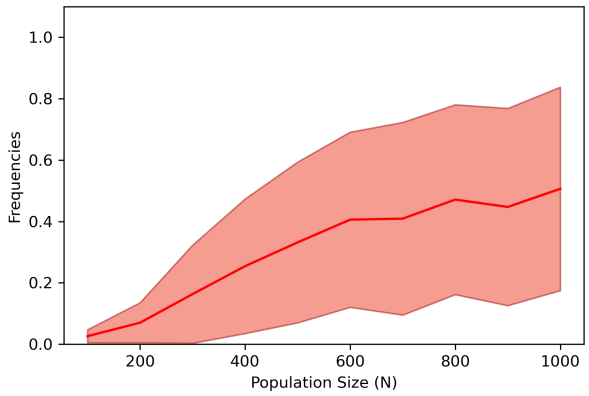



Fig. 9: **H1:** Effect of $N$ on $D$ for $ITоM = 1$ and $DecToM = 1$.

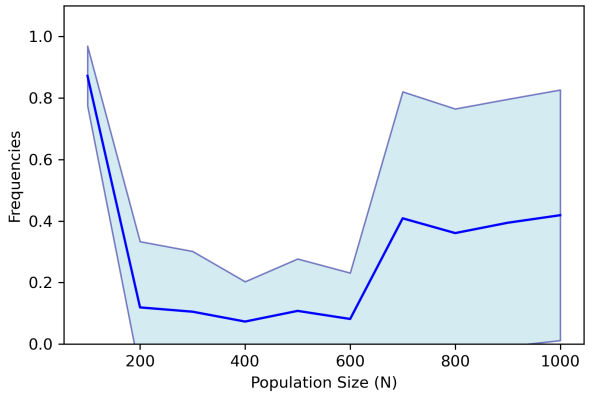



Fig. 10: **H2:** Effect of $N$ on $I$ for $ITоM = 2$ and $DecToM = 1$.

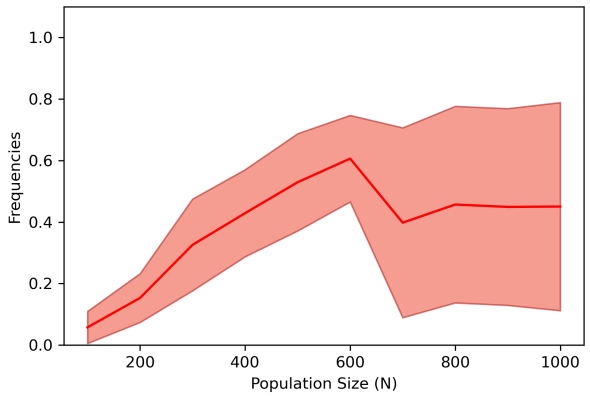



Fig. 11: **H2:** Effect of $N$ on $D$ for $ITоM = 2$ and $DecToM = 1$.

that humans usually require to maintain lies and detect lies. The presence of such AI agents in hybrid society poses new avenues of research and questions about human-agent interactions and the way our future societies will self-organise regarding cooperation, trust and knowledge exchange. Of course, as with any agent-based model, our approach has limitations, but it is a first step in understanding the relation between deception and levels of Theory-of-Mind in self-organising multi-agent systems. As a next step in future work, we aim to explore how social structures (networks) affect the dynamics of the EDG w.r.t. levels of ToM.

## Supplementary Material

Contains (i) data and code for running the simulations and reproducing the statistical analysis and (ii) appendix file with full statistical results and additional figures for showing the effect of $T = 10^5$ vs $T = 10^4$. Supplementary material has been uploaded to the Open Science Framework (OSF) can be found at https://osf.io/8gpzf/?view_only=da1f86d74d4e45cdaa8a30870e77d0b0

## 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.

I would like to thank Reviewers for taking the time and effort necessary to review the manuscript. I would also like to thank Peter R. Lewis, Francesca Mosca, and Alison. R. Panisson for helpful discussions on the topics of AI, evolution and Theory of Mind.

## Ethical Statement

There are no ethical issues. Experimental data was generated solely from simulations.

## References


[1] Sherief Abdallah, Rasha Sayed, Iyad Rahwan, Brad L LeVeck, Manuel Cebrian, Alex Rutherford, and James H Fowler. Corruption drives the emergence of civil society. *Journal of the Royal Society Interface*, 11(93):20131044, 2014.

[2] Damien Anderson, Matthew Stephenson, Julian Togelius, Christoph Salge, John Levine, and Jochen Renz. Deceptive games. In *International Conference on the Applications of Evolutionary Computation*, pages 376–391. Springer, 2018.

[3] Albert Bandura and Richard H Walters. *Social learning theory*, volume 1. Englewood cliffs Prentice Hall, 1977.

[4] David B. Buller and Judee K. Burgoon. Interpersonal Deception Theory. *Communication Theory*, 6(3):203–242, aug 1996.

[5] Judee K Burgoon, David B Buller, Kory Floyd, and Joseph Grandpre. Deceptive realities: Sender, receiver, and observer perspectives in deceptive conversations. *Communication Research*, 23(6):724–748, 1996.

[6] Christopher C Byrne and Jeffrey A Kurland. Self-deception in an evolutionary game. *Journal of Theoretical Biology*, 212(4):457–480, 2001.

[7] Cristiano Castelfranchi. Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, 2(2):113–119, 2000.

[8] Harmen de Weerd, Eveline Broers, and Rineke Verbrugge. Savvy software agents can encourage the use of second-order theory of mind by negotiators. In *CogSci*. Citeseer, 2015.

[9] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 31(2):250–287, 2017.

[10] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, 36(2):1–33, 2022.

[11] Daniel Dennett. Why creative intelligence is hard to find. *Behavioral and Brain Sciences*, 11(2):253–253, 1988.

[12] Liesbeth Flobbe, Rineke Verbrugge, Petra Hendriks, and Irene Krämer Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17:417–442, 2008.

[13] Peter Gärdenfors. Slicing the theory of mind. *Danish yearbook for philosophy*, 36:7–34, 2001.

[14] Alvin I Goldman et al. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, 2006.

[15] Alvin I Goldman et al. Theory of mind. *The Oxford handbook of philosophy of cognitive science*, pages 402–424, 2012.

[16] Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. 1992.

[17] Alison Gopnik and Henry M Wellman. Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085, 2012.

[18] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[19] Garrison W Greenwood. Deceptive strategies for the evolutionary minority game. In *2009 IEEE Symposium on Computational Intelligence and Games*, pages 25–31. IEEE, 2009.

[20] Suzanne Hala, Michael Chandler, and Anna S Fritz. Fledgling theories of mind: Deception as a marker of three-year-olds' understanding of false belief. *Child development*, 62(1):83–97, 1991.

[21] Christoph Hauert, Silvia De Monte, Josef Hofbauer, and Karl Sigmund. Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570):1129–1132, 2002.

[22] Jonathan Y Ito, David V Pynadath, and Stacy C Marsella. Modeling self-deception within a decision-theoretic framework. *Autonomous Agents and Multi-Agent Systems*, 20(1):3–13, 2010.

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

[24] Nicholas S Kovach and Gary B Lamont. Trust and deception in hypergame theory. In *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, pages 262–268. IEEE, 2019.

[25] Nicholas S Kovach, Alan S Gibson, and Gary B Lamont. Hypergame theory: a model for conflict, misperception, and deception. *Game Theory*, 2015, 2015.

[26] Timothy R Levine. *Duped: Truth-default theory and the social science of lying and deception*. University Alabama Press, 2019.

[27] Zhanshan Sam Ma. Towards an extended evolutionary game theory with survival analysis and agreement algorithms for modeling uncertainty, vulnerability, and deception. In *International Conference on Artificial Intelligence and Computational Intelligence*, pages 608–618. Springer, 2009.

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

[29] Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. Lies, bullshit, and deception in agent-oriented programming languages. In *Proceedings of the 20th International TRUST Workshop @ IJCAI/AAMAS/ECAI/ICML*, pages 50–61, Stockholm, Sweden, 2018. CEUR Workshop Proceedings.

[30] Alison R. Panisson, Stefan Sarkadi, Peter McBurney, Simon Parsons, and Rafael H. Bordini. On the formal semantics of theory of mind in agent communication. In *Proceedings of the 6th International Conference on Agreement Technologies*, pages 18–32, Bergen, Norway, 2018. Springer.

[31] Simon T Powers, Anikó Ekárt, and Peter R Lewis. Modelling enduring institutions: The complementarity of evolutionary and agent-based approaches. *Cognitive Systems Research*, 52:67–81, 2018.

[32] Han-Yang Quek, Kay Chen Tan, and Hussein A Abbass. Evolutionary game theoretic approach for modeling civil violence. *IEEE Transactions on Evolutionary Computation*, 13(4):780–800, 2009.

[33] Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, and Martin D. Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302, 2019.

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

[35] Stefan Sarkadi, Ben Wright, Peta Masters, and Peter McBurney (Eds.). *DeceptiveAI*, volume 1296. Springer, 2021.

[36] Stefan Sarkadi. *Deception*. PhD thesis, King's College London, 2021.

[37] Maayan Shvo, Toryn Q Klassen, and Sheila A McIlraith. Towards the role of theory of mind in explanation. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 75–93. Springer, 2020.

[38] Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861, 2010.

[39] Attila Szolnoki, Mauro Mobilia, Luo-Luo Jiang, Bartosz Szczesny, Alastair M Rucklidge, and Matjaž Perc. Cyclic dominance in evolutionary games: a review. *Journal of the Royal Society Interface*, 11(100):20140735, 2014.

[40] Hans van Ditmarsch, Petra Hendriks, and Rineke Verbrugge. Editors' review and introduction: Lying in logic, language, and cognition. *Topics in Cognitive Science*, 12(2):466-84, 2020.

[41] Friedrich Burkhard Von Der Osten, Michael Kirley, and Tim Miller. The minds of many: Opponent modeling in a stochastic game. In *IJCAI*, pages 3845–3851, 2017.

[42] Andrew Whiten and Richard W Byrne. Tactical deception in primates. *Behavioral and brain sciences*, 11(2):233–244, 1988.

[43] Alan FT Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference towards autonomous robotic systems*, pages 85–96. Springer, 2014.