# DECEPTIVE AUTONOMOUS AGENTS

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

## ABSTRACT

The development of Artificial Intelligence (AI) tools that can be used for forging data along with recent events revolving around the problem of fake news propagation indicate new and critical potential threats to intelligence analysis, defence, security, and, by extension, to modern society in general. One such threat that we can derive from the further advancement of AI is the emergence of malicious autonomous artificial agents that could develop their own reasons and strategies to act dishonestly. In order to be able to prevent or mitigate the malicious behaviour of deceptive artificial and autonomous agents, we must first understand how they might be designed, modelled, or engineered. This paper describes the work in [Panisson et al., 2018a], [Sarkadi et al., 2019b], and [Sarkadi et al., 2019a] in which we have modelled and studied how artificial agents that deceive and detect deception can be engineered and evaluated.

***K*eywords** Deceptive Machines · Multi-Agent Systems · Dishonest Behaviour · Complex Reasoning

## 1 Introduction

What is *deception*? A simple definition may be the following: the act of trying to make an entity (whether that entity is a single person or an organisation) hold a belief about something to be true, when in fact that something is false. What follows from this simple definition are other "definitions" that instead of actually defining deception, they describe the ways in which one can deceive, or, to be more specific, the methods and strategies one can use to deceive. Some of these methods can be identified as the following: falsification or negation (usually referred to in as lying in popular language), concealment (the act of hiding the relevant information), equivocation (the act of stating half-truths) [Mahon, 2016]. By combining these three types of actions (or strategies to a certain extent), one can achieve more complex ones (ruses) such as, but not limited to: simulation (imitation of a situation or process from which one can develop strategies or theories), dissimulation (when one acts in such a way as to conceal one's intentions), propaganda (the propagation of misleading information for political purposes), bluffing (when one acts as if he means to do something but he does not actually mean to do that something), misrepresentation (presenting false information as being true to someone in order to make that someone believe something), exaggerations (presenting information as being better or worse than it actually is), understatement (presenting information as being less important than it actually is), camouflage (the act of blending information in such a way as to not be found), disguise (the act of changing appearance in order to conceal one's identity) and theatricalities (displaying something in a dramatic way).

Recent advances in Artificial Intelligence (AI) along with recent events revolving around the problem of fake news indicate new and critical potential threats to intelligence analysis, defence, security, and, by extension, to modern society in general. One such threat that we can derive from the development of AI is the emergence of malicious autonomous artificial agents that could develop their own reasons and strategies to act dishonestly [Sarkadi, 2018].

This paper summarises, contextualises and evaluates as a single overarching approach three novel bodies of work in multi-agent systems (MAS) presented in [Panisson et al., 2018a], [Sarkadi et al., 2019b], and [Sarkadi et al., 2019a]. These works present original contributions towards the engineering of reasoning and behaviour mechanisms for the modelling of deceptive autonomous agents.

## 2 Motivation

Deceptive machines can be traced back to the early days of computer science. They first appear as concepts in Alan Turing's *Imitation Game*:

> [. . . ] it is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either ' X is A and Y is B ' or ' X is B and Y is A'. The interrogator is allowed to put questions to A and B thus: C: Will X please tell me the length of his or her hair? Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be 'My hair is shingled, and the longest strands are about nine inches long [Turing, 1950].

Player A (the man) has to deceive player C (the judge) that he is a woman (that is actually player B) by feeding player C information in such a way that resembles the way in which a woman would. Further on, Turing proposed that the role of player A can be played by a machine and the role of player B can be played by a human. Thus, the objective of the machine (player A) would be to deceive a human judge into thinking that it player A is a human.

The Imitation Game is arguably limited when it comes to the underlying dynamics of deception. In day to day life, humans engage in much more complex interactions compared to the ones in Turing's game. Thus, an interplay between different types of information, not just linguistic, emerges. Such information can be verbal, non-verbal, and contextual.

Studying deception requires one to look inside systems known as black boxes, which in the case of deceptive interactions are the minds of the deceivers and their targets [Sarkadi, 2018]. Reducing the study to the inputs and outputs of these black boxes will most likely tell us nothing about things such as hidden intentions or goals. For example, imagine if, for whatever reason, you wish to deceive your friend into thinking that you like tea, but in reality you like coffee and hate tea. You could drink tea every day in front of your friend. Your friend will start to think (rationally) that you like tea. The only data your friend has about you is the behaviour you exhibit, which is that you drink tea. Thus, the most likely explanation for you drinking tea that your friend can come up with is that you like tea. One can generate deceptive data to influence the results of a data-driven analysis. However, if your friend had access to your deceptive intentions, then your friend would have interpreted your behaviour entirely different and would have gotten the bigger picture.

Two main paradigms seem to be emerging within the AI community: (i) a *model-driven* paradigm and (ii) a *data-driven* paradigm. The model-driven paradigm stands for building an AI that reasons using models which contain beliefs and knowledge about the world and about other agents, in order to interpret evidence (data) and to act according to these models. The data-driven paradigm stands for building AIs that reason based on available evidence (data) without using such models. A data-driven approach seems counter-intuitive to use in the study of deception, because such an approach would limit itself to the analysis of agents' observable behaviour.

To analyse deception from an AI perspective one must refer to beliefs and knowledge, and to include things such as goals, intentions, or desires. Security and Intelligence analysts often confront themselves with the problem of reading the intentions of and accessing the knowledge of their potential adversaries. This problem imposes cognitive limitations on building strategies to deter or counter malicious activities [Heuer, 1999]. Therefore, we have adopted a model-driven approach to study deceptive interactions between agents and see what might emerge from these interactions given multiple setups and scenarios. The approach we present here was designed in such a way that it can be even used for the future study and evaluation of historical and intelligence analysis cases of deception.

## 3 Theoretical Background

This section presents the theoretical background necessary to understand the contributions and technicalities of the models of deception presented later in the paper. The background includes specialised literature from Communication Theory, Cognitive Science, and MAS.

### 3.1 Information Manipulation Theory 2 & Interpersonal Deception Theory

Communication Theory provides two main theories of deception, namely Information Manipulation Theory 2 (IMT2) [McCornack et al., 2014] and Interpersonal Deception Theory (IDT) [Buller and Burgoon, 1996]. The two theories cover different aspects of deception. While IMT2 focuses on the mechanisms behind discourse production, IDT focuses on socio-cognitive aspects of interpersonal interaction that can influence speech production, but is not limited to speech production. A very trivial distinction is that IMT2 addresses the problem of what do we actually say when we want to deceive, while IDT addresses the question of how we say it.

IMT2 focuses on how agents manipulate information to deceive. In particular, IMT2 makes reference to the Mannheim School's psychological models of speech-act production [Herrmann, 2012], implying that information manipulation is related to two main reasoning processes that determine speech production: (i) *Pars Pro Toto*, which means 'parts for the whole' and refers to the process of selecting only the necessary information from a certain context that is sufficient for conveying the entire meaning implied by the speech act; and (ii) *Totum Ex Parte*, which means 'the whole from the parts' and refers to the process used to infer the entire meaning implied by a speech act, given the limited information received through the speech act and the information that is implicit in that situation/context.

IDT focuses on the social factors of deception. IDT argues that there exists a set of social constraints that influence the ability of agents to deceive and detect deception. The most important social constraints are 1) the *trust* between agents, which determines whether an agent believes in the information provided by another agent or chooses to believe the opposite; 2) the *communicative skill* of the agents that determine how skilled are the agents at deceiving and detecting deception; 3) the *cognitive load* of the agents that determines how much information can agents handle in order to succeed in deceptive interactions; the greater the cognitive load, the higher the risk of agents getting caught due to the unintended leaking of information.

### 3.2 Agent Communication Languages

Agent communication languages have been developed based on the speech act theory [Searle, 1969]. Speech act theory is concerned with the role of language as actions. In speech act theory, a speech act is composed by (i) a locution, which represents the physical utterance; (ii) an illocution, which provides the speaker intentions to the hearer; and (iii) the perlocution, which describes the actions that occur as result of the illocution. For example, "I order you to shut the door" is a locution with a illocution of a command to shut the door, and the perlocution may be that the header shuts the door. Thus, an illocution is considered to have two parts, the illocutionary force and a proposition (content). The illocutionary force describes the speech act used, e.g., assertive, directives, commisives, declaratives, expressives.

Among the agent communication languages which emerged from the speech act theory, FIPA-ACL [FIPA, 2008] and KQML [Finin et al., 1994] are the best known. Knowledge Query and Manipulation Language (KQML) was designed to support interaction among intelligent software agents, describing the message format and message-handling protocol to support run-time agent communication. In this work, for practical reasons, we choose KQML, which is the standard communication language in Jason Platform [Bordini et al., 2007], the multi-agent platform we chose to implement this work.

### 3.3 Agent Oriented Programming Languages

Agent-Oriented Programming Languages (AOPLs) can be considered the implementations of agent communication languages. There are many benefits to model ToM in AOPLs, not only to improve the development of MAS, but also to investigate different humans/machines' attitudes towards simulation. Among the many AOPL and platforms, such as Jason, Jadex, Jack, AgentFactory, 2APL, GOAL, Golog, and MetateM, as discussed in [Bordini et al., 2009], we chose the Jason platform [Bordini et al., 2007] for our work. Jason extends the AgentSpeak language, an abstract logic-based AOPL introduced by Rao [Rao, 1996], which is one of the best-known languages inspired by the *belief-desire-intention* BDI architecture. Besides specifying agents with well-defined mental attitudes based on the BDI architecture, the Jason platform has some other features that are particularly interesting for our work, for example, strong negation, belief annotations, and (customisable) speech-act based communication. Strong negation helps the modelling of uncertainty, allowing the representation of things that the agent: (i) believes to be true; (ii) believes to be false; and (iii) is ignorant about. Also, Jason automatically generates annotations for all the beliefs in the agents' belief base about the source from where the belief was obtained (which can be from sensing the environment, communication with other agents, or a mental note created by the agent itself). Also, annotations in Jason can be easily extended to include other meta-information, for example, trust and time as used in [22, 25]. Another interesting feature of Jason is the communication between agents, which is done through a predefined (internal) action. There are a number of performatives allowing rich communication between agents in Jason, as explained in detail in [4]. Further, new performatives can be easily defined (or redefined) in order to give special meaning to them , which is an essential for representing nested reasoning.

### 3.4 Theory of Mind

A Theory of Mind (ToM) of the target is necessary for successful deception according to [Frith and Wolpert, 2004]. ToM is the ability of humans to ascribe elements such as beliefs, desires, and intentions, and relations between these elements to other human agents. In other words, it is the ability to form mental models of other agents. One version of ToM is the Theory-Theory of Mind (henceforth TT). TT can be described as a theory based approach of assigning states to other agents. While some argue TT is nothing else but folk psychology, others say that it is a more scientific way of mind-reading [Gopnik et al., 2004]. Another version is Simulation Theory of Mind (henceforth ST). Adopting Goldman's description of it, Barlassina and Gordon explain it as 'process-driven rather than theory-driven' [Barlassina and Gordon, 1997]. Thus, ST emphasises the process of putting oneself into another's shoes. TT argues for a hypothesis testing method of model extraction, whereas ST argues for a simulation based method for model selection.

[Isaac and Bridewell, 2017] also argue that ToM is crucial for machines to be able to deceive and detect deception. How could a machine be able to reason successfully about the beliefs of other agents if it does not have some knowledge and understanding of its targets' minds? An important stepping-stone towards the modelling and implementation of agents in MAS that use ToM to deceive was the modelling of agents that are able to model the minds of other agents. The introduction of the formal semantics for ToM in [Panisson et al., 2018b] as well as the modelling of uncertainty in [Sarkadi et al., 2018] using the same agent semantics have showed how agents can acquire, update, simulate and use models of other agents' minds to reach shared beliefs and to improve communication and decision making between themselves. The model of uncertainty present in [Sarkadi et al., 2018] implies the existence of two important factors in agent communication, namely the uncertainty of the communication channel and the levels of trust between agents.

## 4 The Dialogue Games Approach

Our modelling approach starts from the following definition of deception within artificial intelligence, and more specifically in agent-based systems: The process in which an agent engages in order to make another agent (or itself) believe something to be false (true) that itself believes to be true (false). In order for deception to take place, several requirements must be met. These requirements are: 1) Intention (an agent must have the intention to deceive); 2) A target (there must be a target that the same agent, that has got the intention to deceive, wants to deceive); 3) A model of the opponent's mind.

We have used *belief-desire-intention* BDI-like architectures to model the cognitive properties of the agents that play these dialogue games [McBurney and Parsons, 2009]. Giving BDI agents a communication protocol along with a reasoning mechanism enables them to think pragmatically about their beliefs, their desires, and their intentions in order to perform speech acts [Rao et al., 1995]. In argumentation, these speech acts can represent arguments, as well as argument chains and argument systems. Apart form performing speech acts, our agents are able to reason about their opponent's mind.

The three models that we have developed in [Panisson et al., 2018a], [Sarkadi et al., 2019a], and [Sarkadi et al., 2019b] present different desirable properties that are useful for the study of machine deception (See Table 1). Below we describe each of these properties:

1. **Explainability** should be a crucial property of argumentation-based models of deception. We should be able to evaluate deceptive mind games and say whether deception takes places and if it does we need to explain why and under which conditions it does. An explainable model should be able to inform us if deception can be prevented or mitigated in different contexts.
2. **Unintended Deception** happens when the Deceiver does not attempt deception, but the consequences of its communicative acts result in its potential target to be deceived. It is important for models of deception to be able to represent such unintended consequences as they are critical for accountability. We need to be able to tell if an agent that has the ability to deceive should be held responsible for its actions or not.
3. **Uncertainty** in communication should be considered when modelling deception. This is especially important for modelling an agent that estimates its likelihood of success, as well as modelling agents with different degrees of trust in each other. While most of the times trust should be a default attitude towards others [Levine, 2014], in cases of potential deception this is not the case.
4. **Storytelling** is the ability of an agent to communicate arguments in such a way as to describe to another agent a meaningful chain of events. The ability to build narratives is an emerging topic in AI. Deceptive agents can use this ability to their own benefit, e.g. deliver a fictitious story that compels a jury into absolving them of a crime. Therefore, it should be desirable for models of deception to consider or represent such mechanisms.
5. **Deception Detection** is desirable to be represented in a model. While some models represent and explain why deception is successful, they do not represent how and why deception might be detected. It is also important to distinguish between an agent that has the ability to detect deception and one that has the tendency to believe or not what a Deceiver is communicating, which is the case in some models. Representing deception

detection could be also useful in showing how a Deceiver might act knowing that its target is able to detect its deceptive intents, as well as how its target might detect them.

6. **Implementation** is to be desired, but not necessary for modelling deception, or any other social phenomenon. However, demonstrating an implementation of the model helps others to use it for studying different MAS setups and scenarios of social interactions. Implementation also improves the **transparency** of a model, increasing the model's accessibility through its code.

| **Model Properties** | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| [Panisson et al., 2018a] | ✓ | - | - | - | - | ✓ |
| [Sarkadi et al., 2019a] | ✓ | - | - | ✓ | ✓ | - |
| [Sarkadi et al., 2019b] | ✓ | ✓ | ✓ | - | - | ✓ |

Table 1: Comparison of our models in terms of their respective desirable properties for the study of machine deception.

### 4.1 Agents that Lie, Bullshit, and Deceive

In our first work on deception, presented in [Panisson et al., 2018a], we model lies, bullshit and deception in an AOPL named Jason [3], which is based on the BDI (Belief-Desire-Intention) architecture. Modelling these dishonest attitudes in MAS allows us to simulate agent interactions in order to understand how agents might behave if they have reasons to adopt these dishonest behaviours. Understanding such behaviours also allows us to identify and deal with such phenomena, as proposed by [7]. Even though the AI community has investigated computational models of lies [31], "bullshit", and deception [6], to the best of our knowledge, our work is one of the first attempts to model these types of agent attitudes in the practical context of an AOPL. AOPLs offer an attractive way of improving the research of dishonest agent behaviour through simulations of agent interactions with explicit representation of relevant mental states. Our study has two main contributions: (i) A comparative model of lies, bullshit, and deception in an AOPL based on the BDI architecture, which allows us to define and simulate these dishonest behaviours. (ii) Making the respective model practical, by implementing an illustrative scenario to show how an agent called car dealer is able to deceive other agents called buyers in buying a car3. In this scenario, the car dealer also tells lies and bullshit in order to make the buyers believe a car is suitable for them, when in fact it is not. The dishonest agent could behave according to the following definitions of lying, bullshitting and deception:

**Definition 1 (Lying)** *The dishonest behaviour of an agent $Ag_i$ to tell another agent $Ag_j$ that $\neg\psi$ is the case, when in fact $Ag_i$ knows that $\psi$ is the case.*

**Definition 2 (Bullshit)** *The dishonest behaviour of an agent $Ag_i$ to tell another agent $Ag_j$ that $\psi$ is the case, when in fact $Ag_i$ does not know if $\psi$ is the case.*

**Definition 3 (Deception)** *The* ***intended*** *dishonest behaviour of an agent $Ag_i$ to tell another agent $Ag_j$ that $\psi$ is the case, when in fact $Ag_i$ knows that $\neg\psi$ is the case, in order to make $Ag_j$ conclude that $\varphi$ given that $Ag_i$ knows that $Ag_j$ knows that $\psi \rightarrow \varphi$ and $Ag_i$ also knows that $Ag_j$ is rational.*

### 4.2 Agents that Deceive using Theory of Mind

Previously, in [Panisson et al., 2018a], we have implemented a dishonest agent in an AOPL that can lie, bullshit and deceive. However, that agent works under the assumption of complete certainty and does not engage in mental simulation to determine the optimal deceptive action. Thus, the model in [Panisson et al., 2018a] does not take into consideration factors that might influence the uncertainty of deceptive outcomes. Therefore, we have further improved that work by taking into consideration the uncertainty behind of deceptive behaviour.

Hence, in our second work on deception [Sarkadi et al., 2019b], we have presented a high-level approach for modelling deception using Theory of Mind in MAS that integrates components of two major theories of deception, namely IDT and IMT2. In this model, we have adopted the following definition of deception: *The intention of a deceptive agent to make another interrogator agent believe something is true that the deceiver believes is false, with the aim of achieving an ulterior goal or desire.*

Our aim was to increase the understanding of how future machines might be able to deceive others by building a mechanism that is able to represent the psychological dynamics between agents under some constraints inspired by the two main theories of deception. The architectures and reasoning mechanisms of the agents in this model allow them to execute *Pars Pro Toto* and *Totum Ex Parte* under social constraints such as trust, levels of communicative skill, and different degrees of certainty in their ToM of their opponent. Besides formalising and evaluating the agent model using BDI-like architectures, we have also successfully implemented the model in Jason, the BDI based AOPL we used in [Panisson et al., 2018a], describing all the steps of the implementation. This shows good synergy between formal specification and implementation while adopting the approach presented in [Panisson et al., 2018b] and [Sarkadi et al., 2018]. Our model in [Sarkadi et al., 2019b] applies the approach in [Panisson et al., 2018b] and [Sarkadi et al., 2018] to give (i) the interrogator the ability to ask for the information it desires based on its partial knowledge of the deceiver's beliefs in order to reach a state of shared beliefs and (ii) to the deceiver the ability to simulate its target's mind.

Furthermore, in order to offer the possibility of extending the model so that it can serve various domains for the study of deception, we have proposed four agent profiles which influence the execution of different behaviours by considering the likelihood of trust and deception between agents. We have also evaluated all the possible outcomes of interaction between these profiles, showing the contexts from which deception emerges.

**Deceiver Profiles:**

- **Reckless** $Ag$ will attempt deception even if the estimated success of deception is low. A reckless deceiver does not care that another agent might misinterpret the reckless deceiver's actions.
- **Cautious** $Ag$ will only attempt deception if the estimated success of deception is high. This means that a cautious deceiver thinks that is wiser to be honest, than to attempt deception and be caught.

**Interrogator Profiles:**

- **Credulous** $Ag_i$ will mostly believe what another $Ag_j$ is saying even if trust is low. A credulous interrogator is an agent that usually does not have a default reason to distrusts others.
- **Sceptical** $Ag_i$ will tend to distrust another $Ag_j$ even if trust is high. A sceptical interrogator believes that there is always a good reason to distrust others.

The most significant result of our model indicates that some agent dynamics can result in cases of *unintended* deception. According to our analysis of the model this means that sceptical attitudes of agents can be detrimental in contexts of deception. This is crucial to take into account in the modelling, design and application of AI in the areas of agreement, cooperation and social interaction. These are areas in which agent attitudes towards trust play a significant role in the outcomes of agent interactions such that deceptive agents might be able to exploit either intentionally or unintentionally.

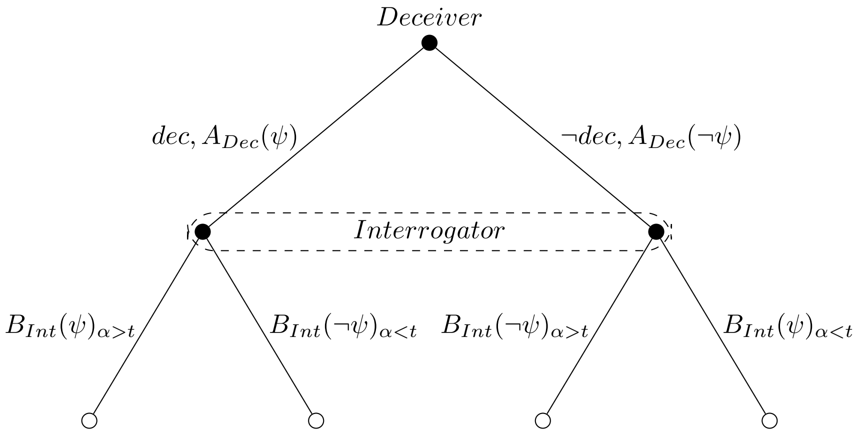

$Deceiver$
$dec, A_{Dec}(\psi)$
$\neg dec, A_{Dec}(\neg\psi)$
$Interrogator$
$B_{Int}(\psi)_{\alpha>t}$
$B_{Int}(\neg\psi)_{\alpha<t}$
$B_{Int}(\neg\psi)_{\alpha>t}$
$B_{Int}(\psi)_{\alpha<t}$
$(dec, B_{Int}(\varphi))$
$(fail, \neg B_{Int}(\varphi))$
$(\neg dec, \neg B_{Int}(\varphi))$
$(fail, B_{Int}(\varphi))$

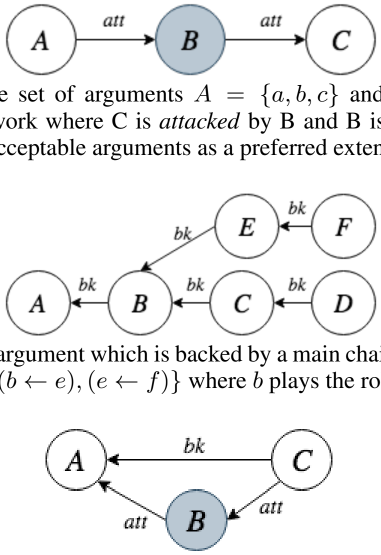
Figure 1: An extensive-form representation of deceptive and non-deceptive plays of Deceiver and all of their possible outcomes. The most right-hand branch represents **unintended deception**; *fail* represents the failure of the intended attempt ($dec$ for intend to deceive or $\neg dec$ for not intend to deceive). $B_{Int}(\varphi)$ represents the Interrogator's belief that $\varphi$ is the case. The social parameter $\alpha$ represents the trust in the communicated message $A_{Dec}(\psi)$ and $t$ represents a trust threshold. If $\alpha > t$, then the message is believed. See the model described in [Sarkadi et al., 2019b] for more details.

### 4.3 Agents that Deceive through Storytelling

Another way to design deceptive machines is to enable them to tell stories and to adapt their stories according to what they think their audience might be more likely to believe. Therefore, in our third work on deception, presented in [Sarkadi et al., 2019a], we explore how a deceptive machine could use stories to deceive an interrogator. We have represented this problem as a dialogue-based argumentation game between two players (deceiver and interrogator) that have partial models of their opponent's mind and we have explained what role stories can play in such a game and how to define them in this context.

We have introduced the idea of stories as complex arguments in a dialogue game between a deceptive storytelling machine and an interrogator. The argumentation-game model allows the two players to develop their own complex arguments using models of their opponent's mind. The players can use these complex arguments as moves in the game. As the game progresses, the players can adapt to the responses of their opponent. The responses can either back the arguments that were previously provided, or they can attack those arguments. Both the Deceiver and the Interrogator use the same reasoning mechanisms to win the dialogue game.

Figure 2: Arguments and complex arguments for storytelling in [Sarkadi et al., 2019a]



(a) Argument system $S = \langle A, R \rangle$ with the set of arguments $A = \{a, b, c\}$ and the relations between the arguments $R = (a, b), (b, c)$. The figure represents a framework where C is *attacked* by B and B is *attacked* by A. We consider that A *defends* C from B. From this we can infer the set of acceptable arguments as a preferred extension $E = \{a, c\}$.

(b) A complex argument where $a$ is the main argument which is backed by a main chain of arguments consisting of $\{(a \leftarrow b), (b \leftarrow c), (c \leftarrow d)\}$ and a sub-chain of arguments $\{(b \leftarrow e), (e \leftarrow f)\}$ where $b$ plays the role of the main argument of that sub-chain.

(c) A complex argument where a main argument $a$ is *attacked* $(b, a)$ by an argument $b$ and *backed* by another argument $(a \leftarrow c)$ which also *defends* it from its attacker $(c, b)$.

The long-term aim of this paper is to emphasise the role of story-telling machines in dialogue games and to also open up future research paths in the area of trust and story-enabled explainable AI in order to understand how it is possible to detect deception by machines, and how to mitigate or ameliorate such deceptive activities. To the best of our knowledge, there is no similar research approach that addressed this particular problem.

## 5 Related Work

The works on machine deception closest to our approach is [Clark, 2010], where the author defines a theoretical machine that uses Theory of Mind (ToM) to formulate illusory sophistic arguments. The machine is represented by an argument scheme. The lying machine feeds information to an audience by exploiting known reasoning fallacies which individuals engage in. The philosophical approach of developing the lying machine has been successfully evaluated using psychological studies of users. In contrast, our approach consists in the design, implementation and evaluation of a MAS that is based on solid theories of deceptive communication. Therefore, our approach is conceptually and methodologically different from [Clark, 2010] and also offers a method to analyse the deceptive machines inside the model, independently of user studies.

Other works on deception in the area of AI that have to be mentioned: in [Christian and Young, 2004] the authors define a model of a heuristic plan search for finding a deception plan. [Jones, 2015] models self-deception using epistemic logic; [Sakama and Caminada, 2010] and [Sakama, 2015] define multiple types of deception using a modal logic of belief and action; [Caminada, 2009] describes the difference between lies, bullshit and deception; and [Lambert, 1987] builds a cognitive model of deception based on human-computer interaction, which specifies how the computer agent's strategies of deception should be improved by the agent's programmer after being defeated by a human in a *battleships* game. Another interesting approach was explored in [Smith et al., 2016], where the authors look at the cognitive aspects of conjuring tricks and describe them through an analysis of their logical form considering it as a contradiction between an expected state and a believed state.

Given that deception is a form of belief manipulation, we mention [Hunter et al., 2017], where the authors describe and implement a model of belief manipulation using propositional public announcements. Their mechanism is similar to ours in the sense that it finds a public announcement $\phi$ that together with a knowledge base $K$ of an agent $A_i$ will make the agent believe a goal $\psi_i$ while being consistent with $K$. However, this model mainly focuses on unidirectionally finding a public announcement for multiple agents and is not able to represent nested beliefs. In contrast, our model focuses on the interactions between two agents where one agent is the target of the other's attempt at belief manipulation. Our model in [Sarkadi et al., 2019b] not only (i) represents nested beliefs, but as a result of ToM modelling, the agents are able to (ii) perform nested reasoning and simulate the other agent's nested beliefs in order to find an announcement that will make the other agent infer a desired belief (iii) while taking into account the likelihood of the announcement's success at manipulation.

## 6 Conclusion

This paper summarises and presents as a single approach three recent and original interrelated bodies of work that have significantly contributed to the understanding of machine deception. These works focus on how to engineer reasoning and interaction mechanisms in order to model, implement, and evaluate deceptive autonomous agents. The first work is the one in [Panisson et al., 2018a] in which we have designed and implemented an agent that can exhibit three types of dishonest behaviour, namely lying, bullshitting, and deception. Building on the first one, in the second body of work [Sarkadi et al., 2019b], we have designed and implemented more complex deceptive reasoning and behaviour mechanisms by integrating Information Manipulation Theory 2 [McCornack et al., 2014] and Interpersonal Deception Theory [Buller and Burgoon, 1996]. In the third body of work, we have defined an argumentation dialogue game model that allows two players to dynamically generate stories and interrogation techniques for deception and deception detection.

## References

[Barlassina and Gordon, 1997] Barlassina, L. and Gordon, R. M. (1997). Folk psychology as mental simulation.

[Bordini et al., 2009] Bordini, R. H., Dastani, M., Dix, J., and Seghrouchni, A. E. F. (2009). *Multi-Agent Programming*. Springer.

[Bordini et al., 2007] Bordini, R. H., Hübner, J. F., and Wooldridge, M. (2007). *Programming multi-agent systems in AgentSpeak using Jason*, volume 8. John Wiley & Sons.

[Buller and Burgoon, 1996] Buller, D. B. and Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory*, 6(3):203–242.

[Caminada, 2009] Caminada, M. (2009). Truth, lies and bullshit; distinguishing classes of dishonesty. In *In: Social Simulation Workshop at the International Joint Conference on Artificial Intelligence (SS@ IJCAI*. Citeseer.

[Christian and Young, 2004] Christian, D. and Young, R. M. (2004). Strategic deception in agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 218–226. IEEE Computer Society.

[Clark, 2010] Clark, M. H. (2010). *Cognitive illusions and the lying machine: a blueprint for sophistic mendacity*. PhD thesis, Rensselaer Polytechnic Institute.

[Finin et al., 1994] Finin, T., Fritzson, R., McKay, D., and McEntire, R. (1994). KQML as an agent communication language. In *Proceedings of the Third International Conference on Information and Knowledge Management*, pages 456–463. ACM.

[FIPA, 2008] FIPA, T. (2008). Fipa communicative act library specification. *Foundation for Intelligent Physical Agents, http://www.fipa.org/specs/fipa00037/SC00037J.html (15.02.2018).*

[Frith and Wolpert, 2004] Frith, C. D. and Wolpert, D. (2004). *The Neuroscience of Social Interaction: Decoding, Influencing, and Imitating the Actions of Others*. Oxford University Press UK.

[Gopnik et al., 2004] Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3.

[Herrmann, 2012] Herrmann, T. (2012). *Speech and situation: A psychological conception of situated speaking*. Springer Science & Business Media.

[Heuer, 1999] Heuer, R. J. (1999). *Psychology of intelligence analysis*. Jeffrey Frank Jones.

[Hunter et al., 2017] Hunter, A., Schwarzentruber, F., Rennes, E., Bruz, F., and Tsang, E. (2017). Belief manipulation through propositional announcements. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1109–1115. AAAI Press.

[Isaac and Bridewell, 2017] Isaac, A. and Bridewell, W. (2017). *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press.

[Jones, 2015] Jones, A. J. (2015). On The Logic of Self-deception. *South American Journal of Logic*, 1:387–400.

[Lambert, 1987] Lambert, D. (1987). A cognitive model for exposition of human deception and counterdeception. Technical report, DTIC Document.

[Levine, 2014] Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, 33(4):378–392.

[Mahon, 2016] Mahon, J. E. (2016). The definition of lying and deception. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

[McBurney and Parsons, 2009] McBurney, P. and Parsons, S. (2009). Dialogue games for agent argumentation. In Simari, G. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 261–280. Springer US.

[McCornack et al., 2014] McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., and Zhu, X. (2014). Information manipulation theory 2: A propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, 33(4):348–377.

[Panisson et al., 2018a] Panisson, A. R., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. H. (2018a). Lies, bullshit, and deception in agent-oriented programming languages. In *20th International Trust Workshop*, pages 50–61, Stockholm, Sweden.

[Panisson et al., 2018b] Panisson, A. R., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. H. (2018b). On the formal semantics of theory of mind in agent communication. In *6th International Conference on Agreement Technologies*, Bergen, Norway.

[Rao, 1996] Rao, A. S. (1996). AgentSpeak (L): BDI agents speak out in a logical computable language. In *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 42–55. Springer.

[Rao et al., 1995] Rao, A. S., Georgeff, M. P., et al. (1995). BDI agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319.

[Sakama, 2015] Sakama, C. (2015). A formal account of deception. In *2015 AAAI Fall Symposium Series*.

[Sakama and Caminada, 2010] Sakama, C. and Caminada, M. (2010). The many faces of deception. *Proceedings of the Thirty Years of Nonmonotonic Reasoning (NonMon@ 30)*.

[Sarkadi, 2018] Sarkadi, S. (2018). Deception. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5781–5782, Stockholm, Sweden.

[Sarkadi et al., 2019a] Sarkadi, S., McBurney, P., and Parsons, S. (2019a). Deceptive storytelling in artificial dialogue games. In *Proceedings of the AAAI 2019 Spring Symposium Series on Story-Enabled Intelligence*.

[Sarkadi et al., 2018] Sarkadi, S., Panisson, A. R., Bordini, R. H., McBurney, P., and Parsons, S. (2018). Towards an approach for modelling uncertain theory of mind in multi-agent systems. In *6th International Conference on Agreement Technologies*, Bergen, Norway.

[Sarkadi et al., 2019b] Sarkadi, S., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., and Chapman, M. D. (2019b). Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302.

[Searle, 1969] Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

[Smith et al., 2016] Smith, W., Dignum, F., and Sonenberg, L. (2016). The construction of impossibility: a logic-based analysis of conjuring tricks. *Frontiers in psychology*, 7:748.

[Turing, 1950] Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):433–460.