The final version of this paper was accepted for publication in AI COMMUNICATIONS Journal Special Issue on Agreement Technologies, Ed. Marin Lujak.

MODELLING DECEPTION USING THEORY OF MIND IN MULTI-AGENT SYSTEMS

Ștefan Sarkadi
Department of Informatics
King's College London
London, WC2 4BG
stefan.sarkadi@kcl.ac.uk

Alison R. Panisson
Center for Technological Development
Federal University of Pelotas
Pelotas - RS, Brazil
alisonpanisson@gmail.com

Rafael H. Bordini
School of Technology
PUCRS
Porto Alegre - RS, Brazil
rafael.bordini@pucrs.br

Peter McBurney
Department of Informatics
King's College London
London, WC2 4BG
peter.mcburney@kcl.ac.uk

Simon Parsons
Department of Informatics
King's College London
London, WC2 4BG
simon.parsons@kcl.ac.uk

Martin Chapman
Department of Informatics
King's College London
London, WC2 4BG
martin.chapman@kcl.ac.uk

June 12, 2019

ABSTRACT

Agreement, cooperation and trust would be straightforward if deception did not ever occur in communicative interactions. Humans have deceived one another since the species began. Do machines deceive one another or indeed humans? If they do, how may we detect this? To detect machine deception, arguably requires a model of how machines may deceive, and how such deception may be identified. *Theory of Mind* (ToM) provides the opportunity to create intelligent machines that are able to model the minds of other agents. The future implications of a machine that has the capability to understand other minds (human or artificial) and that also has the reasons and intentions to deceive others are dark from an ethical perspective. Being able to understand the dishonest and unethical behaviour of such machines is crucial to current research in AI. In this paper, we present a high-level approach for modelling machine deception using ToM under factors of uncertainty and we propose an implementation of this model in an Agent-Oriented Programming Language (AOPL). We show that the Multi-Agent Systems (MAS) paradigm can be used to integrate concepts from two major theories of deception, namely *Information Manipulation Theory 2* (IMT2) and *Interpersonal Deception Theory* (IDT), and how to apply these concepts in order to build a model of computational deception that takes into account ToM. To show how agents use ToM in order to deceive, we define an epistemic agent mechanism using BDI-like architectures to analyse deceptive interactions between deceivers and their potential targets and we also explain the steps in which the model can be implemented in an AOPL. To the best of our knowledge, this work is one of the first attempts in AI that (i) uses ToM along with components of IMT2 and IDT in order to analyse deceptive interactions and (ii) implements such a model.

***K*eywords** Deception · Theory of Mind · Agent-Oriented Programming Languages · Machine Deception

1 Introduction

The idea of deceptive machines dates back to Turing's *imitation game*: '...It is A's object in the game to try and cause C to make the wrong identification...' [Turing, 1950]. We believe that the main reasons to study deception are: (i) because deception is fundamental to a comprehensive theory of communication; and (ii) because some day autonomous agents might have reasons to employ deception [Castelfranchi and Falcone, 2010]. Machines that have the reasons and capability to deceive pose a serious future threat to the relation between humans and AI. This is especially threatening to the relation of trust between humans and artificial agents. Therefore, it is reasonable to think that humans might adopt a sceptical attitude towards AI. We aim to understand these autonomous agents by looking at (i) how deceptive interactions emerge from various contexts; and (ii) what are the possible outcomes of these interactions assuming that agents already have reasons to employ deception, while paying close attention to *sceptical* attitudes of agents.

We consider that our study is a multi-disciplinary one in the sense that it enriches both the AI literature and the literature in communication theory. To AI, it adds two main contributions: **(i)** a model of deception for Multi-Agent Systems (MAS) that includes ToM and uses ToM to integrate components of two major theories of deception: here we describe (a) how agents use ToM in order to deceive (or not) in scenarios of uncertainty, and (b) how adding agent profiles to this model influences the agents' actions by taking into account the likelihood of deception and trust between agents; **(ii)** a broadly applicable approach for implementing the model in Agent-Oriented Programming Language (AOPL) where (a) agents are able not only to model the other agents' minds but also (b) to execute reasoning and simulation over these representations under factors of uncertainty[1]. Another contribution to AI is the understanding of how machines that have the reasons and capability to deceive are able to interact with other agents and what type of behaviour emerges from these interactions that can impact the relation of trust between agents. For communication theory, our study represents the first attempt to integrate the components of two major theories of deception, with the help of AI methods, namely *Interpersonal Deception Theory* (IDT) and *Information Manipulation Thoery 2* (IMT2). IDT is the theory that explains how the communicative skills and cognitive load of individuals affect deceptive interactions [Buller and Burgoon, 1996], whereas IMT2 is the theory that explains how individuals employ deception by manipulating information [McCornack et al., 2014]. Even though there are many studies in the AI literature that look at deception, none of them uses IMT2, IDT and ToM together to model interactions that involve deceptive artificial agents.

2 Theory of Mind & Components of Deception

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.

That being said, ToM seems to be able to provide machines with the ability to model their opponent's minds [Hadjinikolis et al., 2013]. [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? Deception is, after all, a process of epistemic nature.

As mentioned before, 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.

[1] The implementation is available at `https://tinyurl.com/ybj343wf`, thus showing the compatibility between our formalisation and AOPLs.

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 Modelling Deception

We consider *deception* to be different from *lying* and from *bullshitting*[2]. We **define** *deception* as:

Definition 1 (Deception) *The intention of a deceptive agent, which we name Donald, to make another interrogator agent, which we will call Ivan, to believe something is true that Donald believes is false, with the aim of achieving an ulterior goal or desire.*

To model deception, we make use of ToM by combining TT with ST. TT enables us to pre-assign beliefs of agents about each other's beliefs, whereas ST enables agents to simulate other agents' beliefs when they get new information in order to update their TT.

We proceed to build the model by using BDI-like formalisations. Thus the model consists of several sub-components such as: (i) an epistemic component which represents the beliefs and desires of agents (this includes beliefs of other agents' beliefs), (ii) an event component that represents the actions performed by the agents such as asking and answering questions, (iii) and a component that represents how the agents update their beliefs based on ToM and agent profiles.

Definition 2 (Agents) *Ag represents an agent. When we need to make the distinction between two agents we use Ag_i and Ag_j, representing two distinct agents in a set of n agents. The complete set of our agents is $A = \{Dec, Int\}$, where Dec is Donald and Int is Ivan.*

Definition 3 (Beliefs and Desires) *If ψ represents a predicate from a logical language, then $B_{Ag}(\psi)$ represents a belief of an agent Ag in ψ and $D_{Ag}(\psi)$ represents a desire of ψ that belongs to an agent Ag.*

Definition 4 (Actions) *We define $Q_{Ag}(\psi)$ as a question asked by Ag if ψ is the case, and $A_{Ag}(\psi)$ as an answer by Ag saying that ψ is the case.*

Definition 5 (Theory of Mind (ToM)) *A belief or a set of beliefs of an agent Ag_i about another agent Ag_j where: $B_{Ag_i}(B_{Ag_j}(\psi))$ is a belief of an agent Ag_i of another agent Ag_j's belief that ψ.*

Definition 6 (Ignorance) *If ψ represents a predicate from a logical language, and $B_{Ag}(\psi)$ represents a belief of an agent Ag in ψ, $B_{Ag}(\overline{\psi})$ represents that the agent Ag is ignorant about the truth (or falsity) of ψ .*

Definition 7 (Trust Rule) *$A_{Ag_i}(\psi) \rightarrow B_{Ag_j}(\psi)$ represents the general assumption that if Ag_i tells Ag_j that ψ is the case, then Ag_j will believe that ψ is the case.*

To model deceptive interaction, we make agents use ToM to execute *Pars Pro Toto* and *Totum Ex Parte*. Donald will execute *Pars Pro Toto* by combining TT with ST, while Ivan will execute *Totum Ex Parte* using only TT.

Definition 8 (Theory-Theory (TT)) *The prior beliefs that an agent Ag_i has of the beliefs of another agent Ag_j.*

Definition 9 (Simulation-Theory (ST)) *The process that an agent Ag_i engages in to derive new beliefs of another agent Ag_j's beliefs, starting from Ag_i's TT about Ag_j and assuming some new information is received by Ag_j.*

Definition 10 (Pars Pro Toto) *The process executed by an agent Ag_i to choose an answer A_{Ag_i} using its TT of another agent Ag_j and simulating an ST of Ag_j, that will cause the other agent Ag_j to be deceived.*

Definition 11 (Totum Ex Parte) *The process executed by an agent Ag_i to infer something that it desires to know $D_{Ag_i}(B_{Ag_i})$ from a given context that consists of answers provided by another agent A_{Ag_j}, the Trust Rule, and Ag_i's TT and beliefs.*

[2] In [Sarkadi, 2018] and in [Isaac and Bridewell, 2017] the authors explain the complexities of machine deception, and in [Panisson et al., 2018a], as well as in [Caminada, 2009, Sakama and Caminada, 2010], the formal, computational and implementational differences between the three forms of dishonesty are treated.

Definition 12 (Successful Deception) *A successful deception is when the final conclusion reached by $Ivan$ is a belief that $Donald$ desires $Ivan$ to reach but it is also a belief about something to be true that $Donald$ believes to be false.*

3.1 Preconditions

In order for an interaction between two agents to be called *deceptive*, that is to potentially result in successful or failed *deception* given our model, the interaction should satisfy a set of precondictions that follow from Definitions 1, 2, 3, 5 and 6. We consider that if the following three preconditions are satisfied by a given system of at least two agents, then deceptive interactions can happen within that given system.

Precondition 1 (Known Unknown). *Ivan has some missing knowledge about the world such that it is aware of this missing knowledge.*

Precondition 1 is not a strong precondition to be satisfied. Ivan does not necessarily need to be aware of something it does not know. Donald can provide an information that Ivan never thought about finding out in the first place, and by finding out that information, Ivan can infer a belief about something else that is false. We mainly use this precondition in order to show that Ivan will decide to act on its lack of knowledge by asking Donald about Ivan's desired information. Without this precondition, Ivan would not have to ask anyone about something Ivan is aware of not knowing. What Ivan desires is to reach a state of shared beliefs [Chwe, 2013] with Donald given its TT ToM of Donald as agents manage to reach in [Panisson et al., 2018b] and [Sarkadi et al., 2018].

Precondition 2 (Unknown Unknown). *Ivan is initially not aware of the belief Donald desires Ivan to reach.*

We consider that Precondition 2 is a strong precondition to be satisfied by the system in the current form of our model. If Ivan is already aware of the conclusion Donald desires it to reach, then it means that Ivan already has the knowledge (true or false) and thus, Ivan cannot be caused by Donald to have this knowledge. Also, if Ivan already believes something to be true that Donald wants Ivan to believe is false, then Ivan must somehow decide which belief is true or false and this is bound to increase the complexity of the reasoning processes of Ivan. Furthermore, in order to represent deception at an even deeper level, Donald would have to take into consideration the decision protocol of Ivan on its final conclusion. Such interactions are very interesting and worthy to be further explored, but they are currently beyond the scope of this paper.

Precondition 3 (Theory of the Target's Mind). *Donald has a ToM of Ivan.*

We also consider Precondition 3 to be a strong precondition. The argument for this consideration is that: it is impossible for Donald to know what Ivan might infer from information that Donald is able to provide, unless Donald knows what Ivan knows and is able to reason in the way Ivan reasons about what Ivan knows. Therefore, Precondition 3 must stand if any deceptive interaction is to take place. If Precondition 3 does not stand, and Ivan infers a belief that something is true when Donald believes that something to be false, then such an outcome of the system cannot be attributed to a deceptive interaction because such an outcome is not necessarily caused by an action that Donald reasoned deceptively and rationally about. Donald could not have possibly engaged in such a reasoning process, because such a process requires Donald to have a model of Ivan's mind. Such an outcome might just be determined by some random action performed by Donald and, therefore cannot be called deception (see Def. 1).

3.2 Parameters

We assume that the agents, Donald and Ivan, are constrained by two parameters from IDT, namely trust and communicative skill. We proceed to define a value α that represents the degree of Ivan's trust in the information that Donald is providing. Another assumption, inspired by IDT, is that Donald has some sort of skill that it uses to read Ivan's trust. We add this parameter as the *communicative skill* of Donald and label it β.

On top of the parameters from IDT, we also add a degree of confidence γ that Donald has in its TT of Ivan. This is important because we want to show how Donald executes *Pars Pro Toto* under uncertainty. A final assumption is that Donald has to estimate its chance of deception before feeding Ivan any information. We add a success estimation parameter and label it with θ.

We do not provide a model for computing α, β, and γ, because that would change the focus of the paper. The scope of this paper is to show that given some degree of skills, uncertainty about ToM and trust among agents, it is possible to model deception. For an in-depth analysis of how to compute such parameters see [Golbeck, 2008].

3.3 Aggregating Parameters

We choose to aggregate the labels using conditional probabilities in order to show how trust, communicative skill, ToM, and estimation of success influence the dynamics of deception. Let us assume that Ivan does not trust Donald due to some prior information it has about Donald. In this case we say that α has a low probability. Whenever Donald answers Ivan's question with ψ, Ivan will believe that the opposite, $\neg\psi$, is the case. We use the following definitions to show the computation of the interaction between *trust*, *communicative skill*, *confidence in ToM*, and *estimation of success*:

Definition 13 **($P(\alpha)$)** *Trust α is such that $Ivan$ is able to estimate the probability of trust $P(\alpha)$ in the answer provided by $Donald$.*

Both agents need the degree of trust (i) to estimate success (Donald) and (ii) to trust the information provided by the other agent (Ivan).

Definition 14 **($P(\alpha, \beta)$)** *Donald's estimation of Ivan's trust α in Donald is conditionally dependent of Donald's level of communicative skill β.*

In order to succeed in its deception, Donald needs to make Ivan believe what Donald is telling Ivan. To do this under the assumption of uncertainty, Donald needs access to Ivan's degree of trust.

Definition 15 **($P(\theta)$)** *Donald's estimation of its own success θ in deceiving Ivan is the conditional probability of Donald's access to Ivan's trust in Donald given by the probability $P(\alpha, \beta)$ and Donald's confidence in its own ToM of Ivan given by the probability $P(\gamma)$; $P(\theta) = P(\alpha, \beta) * P(\gamma)$.*

3.4 Agent Architectures

When reasoning about knowledge, beliefs and actions using ToM, both Donald and Ivan are able to perform the following:

- Rational Action (RA): If a given agent Ag believes that an action ψ is possible $B_{Ag}(A_{Ag}(\psi))$, then Ag is able to execute that action.
- Assumption of a Future Action (AFA): When using ST, agents are able to make an assumption of taking an action A (answering) or Q (asking) in order to simulate the final outcome of taking that action.
- Positive Introspection (KK): If an agent Ag has a belief of the form $B_{Ag}(\psi)$, then the agent is able to believe that it has that belief $B_{Ag}(B_{Ag}(\psi))$.
- Modus Ponens (MP): The rule that if an agent Ag knows that $\psi \rightarrow \phi$, and ψ is asserted to be true, then Ag knows that ϕ must be true.
- Negation as Failure (NAF): The non-monotonic rule that if a proposition ψ cannot be derived, then $\neg\psi$ is derived.
- Backward Induction (BI): The reasoning process that an agent Ag uses to select an action $A_{Ag}(\psi)$ out of a set of possible actions that will result in the achievement of the agent's desire/goal $D_{Ag}(\phi)$.

Choosing actions that deceive requires some types of decision making rules or protocols. One method compatible with our model is for agents to use *backward induction*: Donald explores all the possible conclusions that can be drawn by Ivan from its answers. If Donald answers $\neg\psi$ and it believes that Ivan is rational and that Ivan believes that $\psi \rightarrow \varphi$, then Donald knows that Ivan will not conclude that φ. Therefore, Donald concludes that deception will fail. After modelling the conclusion Ivan would draw if it answers ψ, Donald proceeds to check if that conclusion matches its desire. If the conclusion of Ivan as modelled by Donald matches Donald's desire, then Donald will proceed to execute the action.

In order to model different attitudes of agents, we add profiles to the agents. For now, we limit the profiles to *reckless* and *cautious* for Donald, and *credulous* and *sceptical* for Ivan.

Deceiver Profiles:

- **Reckless** Ag will attempt deception even if $P(\theta)$ (estimated success) is low, i.e., $P(\theta) \geq 0.25$. A reckless deceiver does not care that another agent, for example, might misinterpret the reckless deceiver's actions.
- **Cautious** Ag will only attempt deception if $P(\theta)$ is high, i.e., $P(\theta) \geq 0.75$. 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 $P(\alpha)$ is low, i.e., $P(\alpha) \geq 0.25$. 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 $P(\alpha)$ is high, i.e., Ag_i will believe what Ag_j is saying only if $P(\alpha) \geq 0.75$. A sceptical interrogator believes that there is always a good reason to distrust others.

Reasoning Processes:

Simulate ToM (see Def. 9 & Algorithm 1) is the reasoning process used by Donald to see what beliefs will be reached by Ivan given some information provided by Donald. We assume that Donald already has a TT (see Def. 8) of Ivan's mind, thus Donald knows what Ivan already knows. Having this knowledge, Donald starts by assuming that it (Donald) will perform a certain action that will be perceived by Ivan (see AFA). Afterwards, Donald assumes that Ivan believes the information that Donald provides (see Def. 7). Given Ivan's newly formed belief on Doanld's information, Donald checks whether this belief is able to generate any final belief in Ivan's mind given Donald's knowledge of all other beliefs that Ivan has. If there is another belief that together with Ivan's newly formed belief generates a final belief in Ivan's mind, then Donald is able to infer that a rational Ivan will conclude this final belief. If there is no other belief in Donald's TT of Ivan that can generate a final belief, then Donald is able to infer that a rational Ivan will not conclude a final belief. **Simulate ToM** will return the conclusion that Ivan would infer if given a certain information.

Pars Pro Toto (see Def. 10 & Algorithm 2) is the reasoning process used by Donald to decide which action should be performed such that the interaction with Ivan will result in successful deception (see Def. 1). **Pars Pro Toto** uses **Simulate ToM** as a subprocess in order to check if a certain action will make Ivan conclude a final belief. After Donald simulates Ivan's mind, Donald checks whether Ivan's conclusion matches Donald's desire. If this is the case, then Donald knows the action chosen will result in succesful deception, therefore Donald proceeds to check the estimation of success (see Def. 14). If the estimation of success is higher than Donald's profile threshold (see Profiles), then Donald will proceed to execute that chosen action (see RA). Else, Donald will choose another action to simulate Ivan's mind until no other actions are left to check. If there is no possible deceptive action above Donald's threshold, then Donald will decide to not attempt deception. **Pars Pro Toto** will return an action that, from Donald's perspective, is likely to deceive Ivan or if there is no such action then it will return an action that is not deceptive. If there is no action that according to **Simulate ToM** will result in Donald's desires, then **Pars Pro Toto** will return a random action.

Totum Ex Parte (see Def. 11 & Algorithm 3) is the reasoning process used by Ivan to find out the information Ivan desires to find out given a certain context. If Ivan is ignorant (see Def. 6) about some information and Ivan has a TT (see Def. 8) of an agent and knows that the agent has the information Ivan desires to know, then Ivan will ask that agent to provide the information and waits for the agent's answer. After receiving the answer from the agent, Ivan believes the answer, but also checks whether it trusts the agent that has provided the information. If Ivan trusts the agent, then Ivan will keep believing the information provided, otherwise Ivan will believe that the information provided is false. Either way, Ivan has achieved its goal, which is not being ignorant anymore about the information.

Algorithm 1: Simulate ToM

```
Function SimulateToM($action$, $belief$, $ToM$)
    let $action$ = say($\varphi$);
    if $ToM \cup \{\varphi\} \models belief$ then
        return True;
    else
        return False;
```

Algorithm 2: Pars Pro Toto

```
Data: $Actions$, $ToM$, $Desire$, $ProfileThreshold$
Result: $DeceiverAction$
let $Desire = D_{Dec}(B_{Int}(\varphi))$;
/* Backward Induction to choose deceptive action                          */
$DeceiverAction \leftarrow \perp$;
for action in Actions do
    if SimulateToM($action, \varphi, ToM$) = True then
        /* Estimate success of action selected using Backward Induction   */
        success $\leftarrow$ Estimate success (of deception);
        if $success \geq ProfileThreshold$ then
            $DeceiverAction \leftarrow action$;
if $DeceiverAction = \perp$ then
    $DeceiverAction \leftarrow$ random $action$ from $Actions$ such that SimulateToM($action, \varphi, ToM$) = False;
```

4 Evaluation and Results

In this section we will present an evaluation of our model. In 4.1 we will go through a step-by-step deceptive play to see how the beliefs of agents evolve during a game. Then, in 4.2 we will present all possible and impossible outcomes of interactions between Donald and Ivan given all possible combinations of parameters and profiles. This will show us what are the contexts from which deception emerges, and then we will discuss the results.

Donald might estimate its success (see Table 2) given its knowledge about Ivan (see Table 1). However, its estimation might not be precise due to a possible strong influence on Ivan by its profile (*credulous* or *sceptical*) and the real degree of trust α (see Table 3).

If it were the case that the agents would operate on absolute knowledge, then Donald would succeed in any given scenario due to its capacity for meta-reasoning and access to a fully accurate ToM of Ivan. Based on the agents' mental states in Table 1, we show a reasoning process based on the agents' ToM and their profiles given certain values for trust α, communicative skill β, and certainty in ToM γ in Tables 2 and 3. We proceed to show a run of a deceptive play using our model.

Another important observation is that given the way we set up the knowledge bases and possible actions of the two agents Donald and Ivan in Table 1 (see $B_{Dec}(B_{Int}(\psi \rightarrow \varphi))$, $B_{Dec}(\neg\psi)$, $A_{Dec}(\psi)$ and $A_{Dec}(\neg\psi)$), attempting deception corresponds to Donald lying. However, that need not necessarily be the case if, let's say, $B_{Dec}(\psi)$ which would create a context in which a deceptive attempt would correspond to Donald telling the truth.

4.1 Running a Deceptive Play

Setup: *cautious* Donald and *sceptical* Ivan with the following values for trust, communicative skill, and ToM: P(α)=0.4, P(β)=0.8, and P(γ)=0.8 (i.e., the first case in Table 4). We run the model by assuming that Ivan asks Donald about ψ:

Event 1

1. $Q_{Int}(\psi)$ from Actions of Ivan and Totum Ex Parte

Donald's mind executing *Pars Pro Toto*

2. $B_{Dec}(\neg\psi)$ from Beliefs of Donald
3. $D_{Dec}(B_{Int}(\varphi))$ from Desires of Donald
4. $B_{Dec}(B_{Int}(\psi \rightarrow \varphi))$ from ToM of Donald

Algorithm 3: Totum Ex Parte

```
Data: $Beliefs$, $Actions$, $ToM$, $Desire$, $ProfileThreshold$, $Trust$
Result: $InterrogatorConclusion$
let $Desire = D_{Int}(\neg B_{Int}(\overline{\psi}))$ ;
if $ToM \models \neg B_{Dec}(\overline{\psi})$ and $ask(\psi) \in Actions$ then
    $action \leftarrow$ ask($\psi$);
else
    $action \leftarrow \bot$ ;
Perform $action$;
Receive $answer$;
let $answer = say(\psi)$;
if $Trust$ > $ProfileThreshold$ then
    $Beliefs \leftarrow Beliefs \cup \{\psi\}$;
else
    $Beliefs \leftarrow Beliefs \cup \{\neg\psi\}$;
if $B_{Int}(\psi)$ then
    if $Beliefs \models \varphi \wedge Beliefs \setminus \{\psi\} \not\models \varphi$ then
        $InterrogatorConclusion$ = $\varphi$ ;
    else
        $InterrogatorConclusion$ = $\bot$;
else
    $InterrogatorConclusion$ = $\bot$;
return $InterrogatorConclusion$;
```

Table 1: Agents with ToM, Labels and Profiles

Donald (Dec) with skill β	**Ivan (Int) with trust α**
Beliefs: $B_{Dec}(\psi \rightarrow \varphi), B_{Dec}(\neg\varphi), B_{Dec}(\neg\psi)$ Desires: $D_{Dec}(B_{Int}(\varphi))$ Actions: $A_{Dec}(\psi)$, $A_{Dec}(\neg\psi)$ ToM with **confidence** γ: $B_{Dec}(B_{Int}(\psi \rightarrow \varphi))$ Profiles: *reckless*, *cautious*	Beliefs: $B_{Int}(\psi \rightarrow \varphi), B_{Int}(\overline{\psi})$ Desires: $D_{Int}(\neg B_{Int}(\overline{\psi}))$ Actions: $Q_{Int}(\psi)$ ToM : $B_{Int}(\neg B_{Dec}(\overline{\psi}))$ Profiles: *credulous*, *sceptical*

Donald's first simulation of Ivan's mind

5. $B_{Dec}(A_{Dec}(\neg\psi))$ Assumption of Donald (random from possible Actions)
6. $B_{Dec}(B_{Int}(\neg\psi))$ from 8 and Trust Rule
7. $B_{Dec}(\neg B_{Int}(\neg\psi \rightarrow \varphi))$ from ToM and NAF
8. $B_{Dec}(\neg B_{Int}(\varphi))$ from 6, 7 and NAF

First simulation: does not meet the desired outcome (see (3)). Therefore, Donald proceeds to perform a second simulation.

Donald's second simulation of Ivan's mind

9. $B_{Dec}(A_{Dec}(\psi))$ Assumption of Donald
10. $B_{Dec}(B_{Int}(\psi))$ from 12 and Trust Rule
11. $B_{Dec}(B_{Int}(\varphi))$ from 7, 13, MP and KK

Second simulation: meets the desired outcome (see 3). Given a successful outcome, Donald computes the estimation of success by aggregating the deceptive parameters: $[P(\theta) = P(\alpha, \beta) * P(\gamma) = 0.32 * 0.8 = 0.26]$, where $[P(\alpha, \beta) = P(\alpha) * P(\beta) = 0.4 * 0.8 = 0.32]$ and $P(\gamma) = 0.8$. Donald proceeds to the decision protocol.

Donald's backward induction and decision using profile

12. $B_{Dec}(B_{Int}(\varphi) \wedge D_{Dec}(B_{Int}(\varphi)))$ from 3,11 and $\wedge I$
13. $B_{Dec}((B_{Int}(\varphi) \wedge D_{Dec}(B_{Int}(\varphi))) \rightarrow B_{Dec}(A_{Dec}(\psi)))$ from 12 and BI
14. $B_{Dec}(A_{Dec}(\psi))$ from 13 and MP

Table 2: Donald executes *Pars Pro Toto* to choose between two possible actions by simulating Ivan's belief updates using: (i) ToM of Ivan, (ii) the probabilities of α, β, γ, and (iii) Donald's profile.

1 $B_{Dec}(B_{Int}(\psi \rightarrow \varphi))$ from ToM 2 $B_{Dec}(D_{Dec}(B_{Int}(\varphi)))$ from Desires and KK
Donald simulates Ivan's Mind given the first possible action.
3 $B_{Dec}(A_{Dec}(\psi))$ AFA and KK 4 $B_{Dec}(A_{Dec}(\psi) \rightarrow B_{Int}(\psi))$ from Trust Rule and KK 5 $B_{Dec}(B_{Int}(\psi))$ from 3, 4 and MP 6 $B_{Dec}(B_{Int}(\psi) \wedge B_{Int}(\psi \rightarrow \varphi))$ from 1, 5 and $\wedge I$ 7 $B_{Dec}(B_{Int}(\varphi))$ from 6, MP, and KK 8 $B_{Dec}(B_{Int}(\varphi) \wedge D_{Dec}(B_{Int}(\varphi)))$ from 2, 7 and $\wedge I$
Donald proceeds to simulate the mind of Ivan given the second (and final) possible action.
3.1 $B_{Dec}(A_{Dec}(\neg\psi))$ AFA and KK 4.1 $B_{Dec}(A_{Dec}(\neg\psi) \rightarrow B_{Int}(\neg\psi))$ from Trust Rule and KK 5.1 $B_{Dec}(B_{Int}(\neg\psi))$ from 3.1, 4.1 and MP 6.1 $B_{Dec}(\neg B_{Int}(\neg\psi \rightarrow \varphi))$ from ToM and NAF 7.1 $B_{Dec}(B_{Int}(\neg\varphi))$ from 5.1, 6.1, and NAF 8.1 $B_{Dec}(B_{Int}(\neg\varphi) \wedge D_{Dec}(B_{Int}(\varphi)))$ from 2, 7.1 and $\wedge I$
For the answer that results in achieving Donald's goal, Donald computes the probability of success P(θ) given P(α, β) $\wedge$ P(γ).
Having assumed that it executes either $A_{Dec}(\psi)$ or $A_{Dec}(\neg\psi)$, Donald has proved that the belief of Ivan matches Donald's desire only if Donald answers $A_{Dec}(\psi)$. Thus, using BI (*backward induction*) Donald knows that 8 implies Donald should answer $A_{Dec}(\psi)$.
9 $B_{Dec}((B_{Int}(\varphi) \wedge D_{Dec}(B_{Int}(\varphi))) \rightarrow B_{Dec}(A_{Dec}(\psi)))$ from 8 and BI 10 $B_{Dec}(A_{Dec}(\psi))$ from 8, 9 and MP with estimated P(θ)
Donald will update its planned answer $A_{Dec}(\psi)$ or $A_{Dec}(\neg\psi)$ based on 10, its profile and P(θ)
11A $B_{Dec}(A_{Dec}(\psi))$ from 10, profile and P(θ) 11B $B_{Dec}(A_{Dec}(\neg\psi))$ from 10, profile and P(θ)
12A $A_{Dec}(\psi)$ from 11A and RA 12B $A_{Dec}(\neg\psi)$ from 11B and RA

Table 3: Ivan executes the second part of *Totum Ex Parte* after receiving Donaold's answer and reaches a conclusion based on (i) the answer of Donald, (ii) the probability of α, and (iii) its profile.

1 $A_{Dec}(\psi)$ from Table 2 (12 A)
2 $B_{Int}(\psi)$ from 1, Trust Rule and MP 3 $B_{Int}(\psi \rightarrow \varphi)$ from Beliefs 4A $B_{Int}(\psi)$ from 1, Trust Rule, profile and **P(α)** 4B $B_{Int}(\neg\psi)$ from 1, Trust Rule, profile and **P(α)** 5A $B_{Int}(\varphi)$ from 3, 4 A and MP 5B $\neg B_{Int}(\varphi)$ from 3, 4 B and NAF

Donald knows that it should answer ψ given 12 and BI in order to deceive Ivan with a success rate of 0.26. Given that Donald is *cautious*, it does not want to risk failure. Thus its initial planned answer $B_{Dec}(A_{Dec}(\psi))$ will be updated to $B_{Dec}(A_{Dec}(\neg\psi))$.

15. $B_{Dec}(A_{Dec}(\neg\psi))$ from 14, *cautious* and $P(\theta) < 0.75$

Event 2

16. $A_{Dec}(\neg\psi)$ from 15 and Actions

Ivan's mind executing *Totum Ex Parte*

---

17. $B_{Int}(\neg\psi)$ $[P(\alpha) = 0.4]$ from 16 and Trust Rule

Because $P(\alpha) = 0.4$ and Ivan is *sceptical* $B_{Int}(\neg\psi)$ will be updated to $B_{Int}(\psi)$.

18. $B_{Int}(\psi)$ from 17, $P(\alpha)$ and *sceptical*
19. $B_{Int}(\psi \rightarrow \varphi)$ from Beliefs of Ivan
20. $B_{Int}(\varphi)$ from 18, 19 and MP

4.2 Results & Analysis

Furthermore, considering the intervals of values established by the profiles introduced, we have analysed in a similar manner all the possible outcomes of deceptive plays between Donald and Ivan. The results are presented in Tables 4-7.

Given our model, some of the outcomes are not possible due to the influence of α on θ and due to the belief and answer update thresholds for each profile. These are: **Table 4** where $P(\alpha) = [0, .75)$ and $P(\theta) = [.75, 1]$; **Table 5** where $P(\alpha) = [0, .25)$ and $P(\theta) = [.75, 1]$; and **Table 7** where $P(\alpha) = [0, .25)$ and $P(\theta) = [.25, 1]$. Figure 1 shows the possible outcomes for our model, given the influence of the parameters α, β and γ.

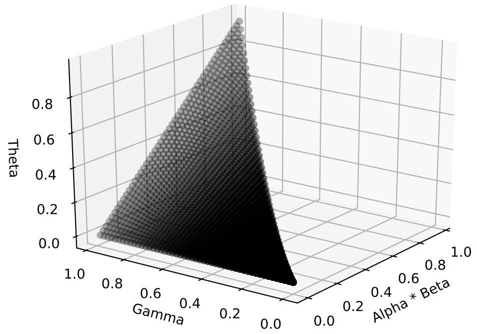

Figure 1: The influence of the parameters α (Alpha), β (Beta) and γ (Gamma) on θ (Theta).

Unintended Deception: The most interesting results are: (i) in **Table 4** where Donald is *cautious* and Ivan is *sceptical*, $P(\alpha) = [0, .75]$ and $P(\theta) = [0, .75]$; and (ii) in **Table 6** where Donald is *reckless* and Ivan is *sceptical*, $P(\alpha) = [0, .75)$ and $P(\theta) = [0, .25)$. In (i) a *cautious* Donald meets a *sceptical* Ivan and *unintended* deception takes place because trust α is considered low by Ivan and because estimation of success θ is considered too low by Donald to attempt deception. In (ii) a *reckless* Donald meets a *sceptical* Ivan, then *unintended* deception takes place because trust α is considered low by Ivan and Donald lacks confidence in its ToM of Ivan θ. Donald will decide not to attempt deception, but Ivan thinks Donald's answer is a lie and decides to believe that the true answer is the opposite of what Donald said, thus reaching the conclusion that Donald actually desires Ivan to reach.

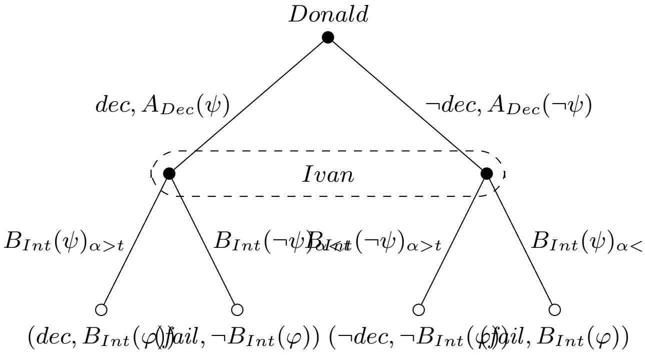
$Donald$
$dec, A_{Dec}(\psi)$
$\neg dec, A_{Dec}(\neg\psi)$
$Ivan$
$B_{Int}(\psi)_{\alpha>t}$
$B_{Int}(\neg\psi)_{\alpha<t}$
$B_{Int}(\neg\psi)_{\alpha>t}$
$B_{Int}(\psi)_{\alpha<t}$
$(dec, B_{Int}(\varphi))$
$(fail, \neg B_{Int}(\varphi))$
$(\neg dec, \neg B_{Int}(\varphi))$
$(fail, B_{Int}(\varphi))$

Figure 2: An extensive-form representation of deceptive and non-deceptive plays of Donald 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).

These results of unintended deception seem to show us that *sceptical* agents can indirectly act as deceptive agents themselves under certain circumstances. Hence, it can be argued that agents that are biased towards sketpicism are not only prone to deceive themselves, but also prone to help actual deceivers to reach their goals without them (the actual deceivers) pro-actively chasing that goal. In a way, these *sceptical* agents offer deceptive agents the option of free-reward. That is, deceivers would maximise their potential payoffs (if there are any) by not paying any costs for deception (if there are any).

Exploring scenarios where deceivers intentionally do not attempt deception in order for *sceptical* interrogators to be ultimately deceived requires an agent architecture with an even higher-order of ToM than we have currently defined in the model. To show a higher-order ToM was beyond the scope of our current study as it was not necessary to show the raw dynamics of machine deception. Moreover, it was sufficient not to have a higher-order ToM in order to understand how skepticism can be detrimental to interrogators in particular deceptive contexts.

Table 4: Cautious vs Sceptical

P(α)	P(θ)	B_{Dec}	A_{Dec}	B_{Int}	Conclusion	Deception
[0, .75)	[0, .75)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	ψ	$B_{Int}(\varphi)$	Yes*
	[.75, 1]	$\neg\psi \wedge \neg\varphi$	ψ	$\neg\psi$	$\neg B_{Int}(\varphi)$	No*
[.75,1]	[0, .75)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	$\neg\psi$	$\neg B_{Int}(\varphi)$	No
	[.75, 1]	$\neg\psi \wedge \neg\varphi$	ψ	ψ	$B_{Int}(\varphi)$	Yes

Table 5: Cautious vs Credulous

P(α)	P(θ)	B_{Dec}	A_{Dec}	B_{Int}	Conclusion	Deception
[0, .25)	[0,.75)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	ψ	$B_{Int}(\varphi)$	Yes
	[.75,1]	$\neg\psi \wedge \neg\varphi$	ψ	$\neg\psi$	$\neg B_{Int}(\varphi)$	No*
[.25, 1]	[0,.75)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	$\neg\psi$	$\neg B_{Int}(\varphi)$	No
	[.75,1]	$\neg\psi \wedge \neg\varphi$	ψ	ψ	$B_{Int}(\varphi)$	Yes

Table 6: Reckless vs Sceptical

P(α)	P(θ)	B_{Dec}	A_{Dec}	B_{Int}	Conclusion	Deception
[0, .75)	[0,.25)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	ψ	$B_{Int}(\varphi)$	Yes*
	[.25,1]	$\neg\psi \wedge \neg\varphi$	ψ	$\neg\psi$	$\neg B_{Int}(\varphi)$	No
[.75, 1]	[0,.25)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	$\neg\psi$	$\neg B_{Int}(\varphi)$	No
	[.25,1]	$\neg\psi \wedge \neg\varphi$	ψ	ψ	$B_{Int}(\varphi)$	Yes

Table 7: Reckless vs Credulous

P(α)	P(θ)	B_{Dec}	A_{Dec}	B_{Int}	Conclusion	Deception
[0, .25)	[0,.25)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	ψ	$B_{Int}(\varphi)$	Yes
	[.25,1]	$\neg\psi \wedge \neg\varphi$	ψ	$\neg\psi$	$\neg B_{Int}(\varphi)$	No*
[.25,1]	[0,.25)	$\neg\psi \wedge \neg\varphi$	$\neg\psi$	$\neg\psi$	$\neg B_{Int}(\varphi)$	No
	[.25,1]	$\neg\psi \wedge \neg\varphi$	ψ	ψ	$B_{Int}(\varphi)$	Yes

5 Implementation in AOPL

Before we were able to successfully implement the model in an AOPL, we first needed to find a way in which to represent ToM in an AOPL. The reason we had to do this is because our model of deception relies on agents that have ToM as a cognitive property. To model agents that model other minds, we adopted the approach presented in [Panisson et al., 2018b] and [Sarkadi et al., 2018]. First, we explain why we chose Jason as an AOPL to implement our model and describe the approach we used for modelling ToM in Jason using its predicates i.e., the representation of TT ToM. After that, we describe how agents execute meta-reasoning using the TT modelling, i.e., we describe how we implement ST as a meta-reasoning mechanism in agent-oriented programming. Finally, we describe how agents use TT and ST to reason about and simulate the other agents' mind in order to make decisions. In particular, we will show the decision-making process for deception, introduced in Section 3.

We consider that the ToM of an agent Ag_i is part of its belief base, i.e., $\Pi^{Ag_i} \subset \Delta^{Ag_i}$, and that everything an agent Ag_i knows that is not in Π^{Ag_i} is considered the private knowledge of Ag_i.

5.1 Agent Oriented Programming Languages

Among the many AOPL and platforms discussed in [Bordini et al., 2009], such as Jason, Jadex, Jack, AgentFactory, 2APL, GOAL, Golog, and MetateM, 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]. Jason [Bordini et al., 2007] has a particular set of features that is interesting for our work: strong negation, belief annotations, and (customisable) speech-act based communication. Also, Jason automatically generates annotations for all beliefs in the agents' belief base about the source of the beliefs. The annotation has the following format: $\mathtt{safe(car1)[source(seller)]}$, stating that the source of the belief that `car1` is safe is the agent `seller`. The annotations in Jason can be easily extended to include other meta-information, e.g., trust [Panisson et al., 2016]. All of these features made Jason the preferred platform for this work. However, other platforms could also benefit from this work by having the approach proposed here adapted to their particularities.

5.2 Modelling ToM in AOPL – TT ToM

An important aspect to be considered in order to model ToM in Jason [Bordini et al., 2007], is that we need not only to represent when an agent believes that another agent believes something (which can be inferred from the belief annotations in Jason), but also we need to represent when an agent believes that another agent does not believe something, or when it is ignorant about that[3]. Therefore, we propose a representation for ToM in Jason, using the following first-order predicates, considering that all are agent a's beliefs:

- `believes(b,p)`: meaning that agent a believes that an agent `b` does believe `p`, i.e., $\Pi^a_b \models B_b(p)$.
- `believes(b,¬p)`: meaning that agent a believes that an agent `b` believes ¬`p`, i.e., $\Pi^a_b \models B_b(\neg p)$.
- `¬believes(b,p)`: meaning that agent a believes that an agent `b` does *not* believe `p`, i.e., $\Pi^a_b \models \neg B_b(p)$.
- `¬believes(b,¬p)`: meaning that agent a believes that an agent `b` does not believe ¬`p`, i.e., $\Pi^a_b \models \neg B_b(\neg p)$.
- `believes(b,inference(q,p))`: meaning that agent a believes that an agent `b` is able to infer `q` from `p`, i.e., $\Pi^a_b \models B_b(p \rightarrow q)$.

Jason automatically annotates all information that an agent has perceived/received with the appropriated source from where that information came. Using this annotation, we are able to implement some meta-reasoning that allows an agent to make inferences[4] from its private knowledge to its ToM. The inference rule $\mathtt{believes(Ag,Prop)} :- \mathtt{Prop[source(Ag)]}$ allows an agent to infer that another agent `Ag` believes proposition `Prop` when `Ag` is the source of that information, i.e., $((\Delta^a \models p[source(b)]) \rightarrow (\Pi^a_b \models B_b(p)))$.

Another important aspect for modelling not only ToM, but also deception, is the representation of when an agent is aware about the other agents being ignorant about a particular information. For example, an agent a is able to infer that another agent b is ignorant about a proposition `p`, when b does not believe on either `p` or ¬`p` — `¬believes(b,p)` and `¬believes(b,¬p)` hold on a's belief base — i.e., $\Pi^a_b \models \neg B_b(p) \wedge \Pi^a_b \models \neg B_b(\neg p)$. The following inference rule allows agents to make such inference: $\mathtt{ignorant\_about(Ag,Prop)} :- \mathtt{\neg believes(Ag,Prop)} \& \mathtt{\neg believes(Ag,\neg Prop)}$

Note that stating that an agent a is ignorant about whether agent b believes `p` or not is different from a ToM saying that a knows/believes that agent b does not believe either `p` or ¬`p`, i.e., $(\Pi^a_b \not\models B_b(p) \wedge \Pi^a_b \not\models \neg B_b(p))$ (agent a is ignorant about if b believes or not in p) is different from $(\Pi^a_b \models \neg B_b(p) \wedge \Pi^a_b \models \neg B_b(\neg p))$ (agent a knows that agent b is ignorant about p). Following the same ideas, an agent a is able to infer when itself is ignorant about some proposition, i.e., $(\Delta^a \not\models p \wedge \Delta^a \not\models \neg p)$: $\mathtt{ignorant\_about(Prop)} :- \mathtt{not(Prop)} \& \mathtt{not(\neg Prop)}$. Finally, an agent a is able to infer when it is ignorant about other agents' beliefs, i.e., considering another agent b we have $(\Pi^a_b \not\models B_b(p) \wedge \Pi^a_b \not\models \neg B_b(p))$:

$\mathtt{ignorant\_about(believes(Ag,Prop))} :- \mathtt{not(believes(Ag,Prop))} \& \ \mathtt{not(\neg believes(Ag,Prop))}$.

5.3 Reasoning Using ToM – ST ToM

For the purpose of this study, it is important to consider what information is being processed in order to form a ToM. We know that ToM consists of beliefs about others' minds (TT ToM) and we also know that ToM formation can be represented through role-playing or simulating others' minds (ST ToM). In this study, both perspectives are considered.

Based on our approach for representing ToM in Jason agents, we explain below how agents use that representation in order to make inferences from ToM, i.e., ST ToM. For example, an agent is able to infer new information about other agents' beliefs from the information it already has on its ToM about those agents. For example: $\mathtt{believes(Ag,C)} :- \mathtt{believes(Ag,inference(C,P))} \ \& \ \mathtt{believes(Ag,P)}$ says that an agent is able to infer that another agent `Ag` believes `C` when it knows, from its ToM, that agent `Ag` believes that `P` implies `C`, and agent `Ag` believes `P`, i.e., $((\Pi^a_b \models B_b(p \rightarrow q) \wedge \Pi^a_b \models B_b(p)) \rightarrow \Pi^a_b \models B_b(q))$.

[3] Usually, an agent will use inquiry dialogues to have access to such information.

[4] These inferences are characterised as ST ToM, as we describe in next section.

Also, agents are able to infer, from their ToM, the missing information for achieving a particular desired state of ToM. For example, imagine that an agent a desires to achieve a state of ToM in which another agent b believes q, i.e., $\Pi^a_b \models B_b(q)$. Considering that the current state of a's ToM only indicates that agent b believes p implies q, i.e., $\Pi^a_b \models B_b(p \rightarrow q)$, using ST an agent is able to infer that it needs to achieve $\Pi^a_b \models B_b(p)$, thus with $(\Pi^a_b \models B_b(p \rightarrow q) \wedge \Pi^a_b \models B_b(p))$ it is able to achieve the desired ToM state $\Pi^a_b \models B_b(q)$. This backward-reasoning can also be observed in Table 2. Agents execute such reasoning using the following inference rule: $\texttt{implies}(\texttt{believes}(\texttt{Ag},\texttt{N}),\texttt{believes}(\texttt{Ag},\texttt{C})) : - \texttt{believes}(\texttt{Ag},\texttt{inference}(\texttt{C},\texttt{N}))$; meaning that an agent is able to infer that another agent `Ag` believing `N` implies it also believing `C`, considering that the agent knows `Ag` believes in inferring `C` from `N`. Note that, here, the agent does not need to model that agent `Ag` believes `N`, it just simulates such an inference[5].

In contrast to agents' reasoning using ST ToM, which we can easily note is not domain dependent, agents' decision making *is* domain dependent, given that different domains will require different decision making. In the next section, we show how agents use TT (i.e., the initial ToM) and ST (i.e., inferences, simulation, and updates they execute using ToM) to make decisions. In particular, we will discuss how agents make decisions based on the scenario in Section 3.

5.4 Decision Making and Communication Semantics

In this section we show how agents update their belief bases and ToM during communication, as well as how agents make decisions based on the state of their mental attitudes (i.e., ToM, belief base, desires, etc.). We give formal semantics to both speech acts used for modelling deception in multi-agent systems, namely `ask` and `response`. We define new semantic rules to accompany the existing operational semantics of Jason [Bordini et al., 2007, Vieira et al., 2007]; however, for clarity we use only the configuration components that we need to formalise the essentials of our approach. Also, to account for the decision-making process, we define two functions, `Conf()` and `Trust()`, which describe different behaviours that agents may adopt depending on the parameters α, β, and γ introduced in Section 3, and based on their profiles.

First, in **Ask1**, when an agent receives an `ask` message and it believes it is likely to be successful in deceiving the sender (based on its profile), it sends a `response` message with the information that makes it achieve the desired state of ToM, regardless of whether the sender agent believes it to be true or not.

$$\frac{\begin{array}{c} S_M(M_{In}) = \langle mid, sid, \texttt{ask}, \psi \rangle \\ \Pi^{ag}_{sid} \models B_{sid}(\psi \rightarrow \varphi) \qquad \Pi^{ag}_{sid} \models \neg B_{sid}(\psi) \qquad \Pi^{ag}_{sid} \models \neg B_{sid}(\neg\psi) \\ (\Pi^{ag}_{sid} \cup \{B_{sid}(\psi)\}) \models B_{sid}(\varphi) \qquad \texttt{Conf}() = \texttt{true} \end{array}}{\langle ag, M, \mathsf{ProcMsg} \rangle \longrightarrow \langle ag, M', \mathsf{ExecInt} \rangle} \qquad (\textbf{Ask1})$$

$$\begin{array}{lll} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle mid, sid, \texttt{ask}, \psi \rangle\} \\ M'_{Out} & = & M_{Out} \cup \{\langle mid, sid, \texttt{response}, \psi \rangle\} \end{array}$$

Ask1 says that when an agent selects a received message to be processed[6] $\langle mid, sid, \texttt{ask}, \psi \rangle$ (with mid and sid the message and sender identifier, respectively), and it knows, from TT ToM, that the sender is able to infer φ from ψ — $\Pi^{ag}_{sid} \models B_{sid}(\psi \rightarrow \psi)$ — and that the sender is ignorant about ψ — $\Pi^{ag}_{sid} \models \neg B_{sid}(\psi) \wedge \Pi^{ag}_{sid} \models \neg B_{sid}(\neg\psi)$ — then, using ST ToM, the agent infers that responding ψ will probably (considering the parameters mentioned and the agent profile in `Conf()`) make the sender believe ψ — $B_{sid}(\psi)$ — and making the sender believe ψ gets the agent to achieve a ToM state corresponding to its desire — $(\Pi^{ag}_{sid} \cup \{B_{sid}(\psi)\}) \models B_{sid}(\varphi)$. Finally, after reasoning about which response to provide, the agent removes that message from its mail inbox M_{In}, and add the corresponding message to the mail outbox M_{Out}. Otherwise, in **Ask2**, when the agent believes it is unlikely to be successful in deceiving the sender (based on its profile), it responds truthfully.

[5] For simplicity, we only model other agents' beliefs in both TT and ST ToM, which is sufficient to show our approach. Similar types of beliefs can be easily added. For example, we are able to model (i) that an agent a has the desire to become aware of q (i.e., $D_a B_a(q)$) using the predicate $\texttt{desires}(\texttt{a},\texttt{believes}(\texttt{a},\texttt{q}))$ at TT ToM, (ii) that an agent does not believe something while it has a desire for that, i.e., $\neg\texttt{believes}(\texttt{Ag},\texttt{Prop})\texttt{:-}\ \texttt{desires}(\texttt{Ag},\texttt{believes}(\texttt{Ag},\texttt{Prop}))$.

[6] Here it suffices to know that this function selects one message from the agent's inbox M_{In}, see [Bordini et al., 2007, Vieira et al., 2007] for more details about this function $S_M()$.

$$\frac{\begin{array}{c} S_M(M_{In}) = \langle \mathit{mid}, \mathit{sid}, \texttt{ask}, \psi \rangle \\ \Pi^{ag}_{sid} \models B_{sid}(\psi \to \varphi) \qquad \Pi^{ag}_{sid} \models \neg B_{sid}(\psi) \qquad \Pi^{ag}_{sid} \models \neg B_{sid}(\neg\psi) \\ (\Pi^{ag}_{sid} \cup \{B_{sid}(\psi)\}) \models B_{sid}(\varphi) \qquad \texttt{Conf}() = \texttt{false} \end{array}}{\langle ag, M, \mathsf{ProcMsg} \rangle \longrightarrow \langle ag, M', \mathsf{ExecInt} \rangle} \qquad (\textbf{ASK2})$$

$$\begin{array}{lcl} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle \mathit{mid}, \mathit{sid}, \texttt{ask}, \psi \rangle\} \\ M'_{Out} & = & M_{Out} \cup \{\langle \mathit{mid}, \mathit{sid}, \texttt{response}, \phi \rangle\} \text{ with} \\ & & \phi = \begin{cases} \psi & \text{if } \Delta^{ag} \models \psi \\ \neg\psi & \text{if } \Delta^{ag} \models \neg\psi \\ \texttt{ignorant}(\psi) & \text{if } \Delta^{ag} \not\models \neg\psi \;\wedge \Delta^{ag} \not\models \psi \end{cases} \end{array}$$

When an agent receives a `response` message, it updates its belief base depending on the result of `Trust()`; we assume this function to determine, depending on the agent profile, whether the sender appears trustworthy. Thus, in **RESPONSE1**, when the agent trusts the sender (based on the receiver profile), it updates its belief base with that information. Otherwise, in **RESPONSE2**, when the agent does not trust the sender (again, based on the receiver profile), it updates its belief base assuming that the sender is lying, thus assuming that the appositive is the case.

$$\frac{S_M(M_{In}) = \langle \mathit{mid}, \mathit{sid}, \texttt{response}, \psi \rangle \qquad \texttt{Trust}() = \texttt{true}}{\langle ag, M, \mathsf{ProcMsg} \rangle \longrightarrow \langle ag', M', \mathsf{ExecInt} \rangle} \qquad (\textbf{RESPONSE1})$$

$$\begin{array}{lcl} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle \mathit{mid}, \mathit{sid}, \texttt{response}, \psi \rangle\} \\ \Delta'_{ag} & = & \Delta_{ag} \cup \{\psi[\mathit{source}(\mathit{sid})]\} \end{array}$$

$$\frac{S_M(M_{In}) = \langle \mathit{mid}, \mathit{sid}, \texttt{response}, \psi \rangle \qquad \texttt{Trust}() = \texttt{false}}{\langle ag, M, \mathsf{ProcMsg} \rangle \longrightarrow \langle ag', M', \mathsf{ExecInt} \rangle} \qquad (\textbf{RESPONSE2})$$

$$\begin{array}{lcl} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle \mathit{mid}, \mathit{sid}, \texttt{response}, \psi \rangle\} \\ \Delta'_{ag} & = & \Delta_{ag} \cup \{\neg\psi[\mathit{source}(\mathit{sid})]\} \end{array}$$

5.5 Example

As a real-world example for our implementation, we take the car-sale scenario as in [Sklar et al., 2004] and in [Panisson et al., 2018a]. Donald is a car dealer and Ivan is a potential buyer. When we buy cars, if we are rational (and we assume the potential buyer is rational), then we consider safety[7] of the vehicle as being a priority. We set up the scenario in Table 8. Further, this scenario corresponds to the fourth case in Table 6, instantiating our model using the abstract agents defined in Table 1.

Table 8: Setup of Real-World Example

Donald (Dec) ***reckless*** **with skill** β = 0.8
Beliefs: $B_{Dec}(\texttt{safe(X)} \to \texttt{buy(X)}), B_{Dec}(\neg\texttt{safe(bmw)}), B_{Dec}(\neg\texttt{buy(bmw)})$ **Desires**: $D_{Dec}(B_{Int}(\texttt{buy(bmw)}))$ **Actions**: $A_{Dec}(\texttt{safe(bmw)}), A_{Dec}(\neg\texttt{safe(bmw)})$ **ToM** with **confidence** γ = 0.8: $B_{Dec}(B_{Int}(\texttt{safe(X)} \to \texttt{buy(X)}))$
Ivan (Int) ***sceptical*** **with trust** α = 0.8
Beliefs: $B_{Int}(\texttt{safe(X)} \to \texttt{buy(X)})$, $B_{Int}(\overline{\texttt{safe(bmw)}})$ **Desires**: $D_{Int}(B_{Int}(\texttt{safe(bmw)}) \vee B_{Int}(\neg\texttt{safe(bmw)}))$ **Actions**: $Q_{Int}(\texttt{safe(bmw)})$ **ToM** : $B_{Int}(\neg B_{Dec}(\overline{\texttt{safe(bmw)}}))$

We describe our scenario making reference to the running model from Section 4.1, showing also that our approach can instantiate many similar scenarios of deception: **Ivan** is ignorant about whether a `bmw` is safe or not. Therefore **Ivan** sends an `ask` message whether the `bmw` is safe or not to **Donald**. Next, **Donald** receives the message, which corresponds to the semantics rule **ASK1** with **Donald**'s profile being *reckless* and $\theta = 0.51$. **Donald**'s decision-making

[7] [Mcburney et al., 2003] includes *passenger safety* as an important part of the car attributes in negotiation scenarios.

process corresponds to the instantiation of the backward-reasoning from (5) to (16) and in the semantics rule **Ask1**, which ends with **Donald** responding that the bmw is safe (16). **Ivan** receives the message, which corresponds to the semantics rule **Response1**, given that, though **Ivan** profile is *sceptical*, it trusts **Donald** i.e., $\alpha = 0.8$. Finally, **Ivan** concludes that the bmw is safe, corresponding the instantiation of the reasoning process from (17) to (20) and the belief update showed in **Response1**, which ends with **Ivan** believing that it should buy a bmw.

6 Related Work

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. The influence of trust on agent communication also implies the possibility of dishonest behaviour that might stop agents from reaching a real state of shared beliefs. Our model 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.

The works on machine deception closest to our approach are [Clark, 2010] and [Panisson et al., 2018a]. In [Panisson et al., 2018a], the authors model and implement a dishonest agent in an AOPL that can lie, bullshit and deceive. However, their dishonest 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 such as levels of trust, communicative skill and confidence in ToM. In [Clark, 2010], 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 that have to be mentioned: [Sarkadi et al., 2019] models deception using argumentation in dialogue games; [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.

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 ϕ that together with a knowledge base K of an agent A_i will make the agent believe a goal ψ_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 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.

7 Conclusions

In this paper we have presented a high-level approach for modelling deception using Theory of Mind in Multi-Agent Systems that integrates components of two major theories of deception, namely Interpersonal Deception Theory and Information Manipulation Theory 2. 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 theories of deception. Besides formalising and evaluating the agent model using BDI-like architectures, we have also successfully implemented the model in a BDI based AOPL 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]. 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. 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.

As future work, we are curious to explore how to increase the order of ToM defined in the current agent architecture. Another of our research aims is the inclusion of a *cognitive load* [Buller and Burgoon, 1996] component, more profiles for the agents, an ST ToM for the interrogator agent to be able to detect deception, and an environment that agents can use to deceive and detect deception.

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 (Wiley Series in Agent Technology)*. 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.

[Castelfranchi and Falcone, 2010] Castelfranchi, C. and Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model*, volume 18. John Wiley & Sons.

[Chwe, 2013] Chwe, M. S.-Y. (2013). *Rational ritual: Culture, coordination, and common knowledge*. Princeton University Press.

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

[Golbeck, 2008] Golbeck, J. (2008). *Computing with social trust*. Springer Science & Business Media.

[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.

[Hadjinikolis et al., 2013] Hadjinikolis, C., Siantos, Y., Modgil, S., Black, E., and McBurney, P. (2013). Opponent modelling in persuasion dialogues. In *IJCAI*, pages 164–170.

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

[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.

[Mcburney et al., 2003] Mcburney, P., Van Eijk, R. M., Parsons, S., and Amgoud, L. (2003). A dialogue game protocol for agent purchase negotiations. *Autonomous Agents and Multi-Agent Systems*, 7(3):235–273.

[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., 2016] Panisson, A. R., Melo, V. S., and Bordini, R. H. (2016). Using preferences over sources of information in argumentation-based reasoning. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 31–36. IEEE.

[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 (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*, pages 50–61. CEUR-WS.

[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 (co-located with EUMAS 2018), Bergen, Norway, 6-7 December, 2018.*

[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.

[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 (IJCAI 2018)*, pages 5781–5782.

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

[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 (co-located with EUMAS 2018), Bergen, Norway, 6-7 December, 2018.*

[Sklar et al., 2004] Sklar, E., Parsons, S., and Davies, M. (2004). When is it okay to lie? a simple model of contradiction in agent-based dialogues. In *ArgMAS*, pages 251–261. Springer.

[Turing, 1950] Turing, A. (1950). Computing machinery and intelligence-am turing. *Mind*, 59(236):433.

[Vieira et al., 2007] Vieira, R., Moreira, Á. F., Wooldridge, M., and Bordini, R. H. (2007). On the formal semantics of speech-act based communication in an agent-oriented programming language. *Journal of Artificial Intelligence Research*, 29:221–267.