Deception

Sarkadi, Stefan

Awarding institution:
King's College London

Deception

By
Ștefan Sarkadi

A thesis submitted in fulfilment for the
degree of Doctor of Philosophy

in the
Department of Informatics
School of Natural & Mathematical Sciences
King's College London

2020

This thesis is dedicated to Francesca for ineffable reasons.

To my mother and father, Lucretia and Carol,

for letting their youngest child's mind always roam freely, but not mindlessly.

Acknowledgements

I am extremely grateful to have received help from interacting with many different people during my PhD. I wish to thank everyone that has supported me, including the people whose support was provided without their knowledge. Of course, I must apologise for anyone I might have forgotten to mention, as there may be individuals whose help I was not aware I was taking or whose help I have now forgotten to have taken.

First, financial help. The research presented in this thesis was partly funded in the form of a PhD studentship by the Faculty of Natural and Mathematical Sciences, King's College London. I am also grateful to the MIT Media Lab for funding my PhD research visit.

Second, there are a few people who have taken a particular interest in my work and progress and who have given me special guidance during my PhD. They deserve particular gratitude and special mention: Elizabeth Black, King's College London; Rafael Bordini, PUCRS; Martin Chapman, King's College London; Sanjay Modgil, King's College London; Matthew Moran, King's College London; Alison Panisson, PUCRS; Iyad Rahwan, MIT, and Center for Humans and Machines, Max-Planck Institute; Alex Rutherford, Center for Humans and Machines, Max-Planck Institute; Nishanth Sastry and his PhD students, Pushkal Agarwal and Hu Xuehui, King's College London; Liz Sonenberg, University of Melbourne; Jose Such, King's College London; and Luca Vigano, King's College London.

Third, the people who have supported me as close colleagues, friends, and family during my PhD. Some of them are also the people with whom crazy and fun ideas have been exchanged during the never ending coffee breaks, and who have shed light on the bright side of my life as a PhD student at King's. They are: Xavi Ferrer Anan, Leonardo Magela Cunha, Adrian Salazar Gomez, Adam Green, Chris Hampson, Jenjira Jaimunk, Grammateia Kotsialou, David Kohan Marzagao, Moises Martinez Munoz, Ionut Moraru, Jack Mumford, Josh Murphy, Luke Riley, Diego Sempreboni, Joseph Shams, Jacob Swambo, Tomas Vitek, Akkapon Wongkoblap, Andreas Xydis, Peter Young, Parisa Zehtabi, and Tsvetan Zhivkov. Others, while not my colleagues at King's, have given me support long before my PhD. They are: Denisa Bulzan, Ana Galea, Luchian Ciorbagiu, Sara Guermani, Erzsy Nagy, Ioani Nagy, Daniela

Orza, Diana Plesa, George Rosu, Daniel Sturza, Gabriela Sturza, Paula Sturza, and Raluca Sturza.

To my examiners Floris Bex and Sara Uckelman, I cannot express enough gratitude to you for dedicating precious time to reading this thesis. More than that, it was a privilege to have you as examiners and I have very much appreciated the reassuring calmness, rigour, breadth and detail of your scrutiny during the viva.

To my mother, Lucretia Sarkadi, and my father, Carol Sarkadi, I must be forever grateful for having had the love and patience to deal with me during my PhD.

To Francesca Mosca, I cannot express in words how much I am grateful to you for putting up with me as a partner doing a PhD, but if I had to say something about this period of time, then it is this: it definitely takes a lot more love and patience than my parents needed during this time to deal with me the way you have managed to.

Finally, my supervisors, Professor Peter McBurney and Professor Simon Parsons. Thank you both for mentoring me through a very interesting journey!

Peter, it is very difficult to enumerate and describe the things you have done for my career, starting from answering that e-mail in which I asked if you would take me on as a PhD student. However, most thankful I am for the endless chats we have had together on what some would call a broad spectrum of things. I would counter-argue that it was a broad spectrum, and I would say that the spectrum of topics was quite narrow and the chats quite short, because it feels like we have discussed very few topics for a very short time.

Simon, I don't know if I would have done any actual work if you wouldn't have grounded my ideas. Most grateful I am for the way you managed to guide my ideas into solid work, and for the things you have taught me about what it means to conduct research and to teach in academia. However, most importantly, you have taught me that one should never say no to coffee, and that this also applies to AI decision-making.

Abstract

This thesis is about machine deception. It is the first full computational treatment in Artificial Intelligence (AI) on how to create machines able to deceive. The dissertation discusses the limited related research on deception that exists in AI, Philosophy and Psychology.

This thesis tackles the problem of machine deception from two different directions. The main direction is from the cognitive modelling perspective of agents in Multi-Agent Systems (MAS). Working from this perspective has enabled the engineering and formal modelling of reasoning mechanisms that artificial agents could potentially use to deceive and to reason about other minds in a similar fashion to how humans perform these tasks. The other direction is from an evolutionary perspective on agent behaviour in multi-agent systems. Working from this second direction shows how deception can destabilise cooperation in hybrid societies, where humans and machines interact socially through the exchange of knowledge, but it also shows how cooperation can be re-established if the right mechanisms for social interaction are in place.

This thesis presents six contributions to the field of AI: 1) A conceptual grounding of computational deception; 2) A novel approach to model and implement practical reasoning artificial agents that have the capability to model and reason about the minds of other agents in communication; 3) A novel, formal approach to model and engineer deceptive artificial agents in MAS, that is grounded in three major theories of deceptive communication; 4) A detailed step-by-step description of the implementation of the models described by this formal approach in the Jason agent-oriented programming language; 5) A novel approach to model and evaluate deception in evolutionary public goods games of knowledge sharing between agents of hybrid societies; 6) The proposal of an MAS framework for deception to be used in Intelligence Analysis.

This thesis leads to three main future research directions. These regard the refinement of the models presented in the thesis, the creation of MAS tools for deception analysis, and, finally, the creation of a machine *worth talking to.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"Theatricality and deception are powerful agents to the uninitiated...but we are initiated, aren't we, Bruce?"*
>
> — Bane, in The Dark Knight Rises

With this chapter I begin this thesis; I initiate the reader by defining the central concepts of my thesis and my research questions.

1.1 Overview

The research presented in this thesis aims to weave a pattern of ideas which depicts an answer to the following question: *Can we use artificial agents to improve our understanding of deception?* Soon enough, I came to realise that in order to answer this question, I must ask a broader, and more philosophical question, namely: *How may artificial agents deceive?*

Most of the ideas presented in this thesis aim to answer the latter question, as philosophical as it may be, from the perspective of Artificial Intelligence in Computer Science, and more specifically, from the perspective of multi-agent systems (MAS) research. Regarding the first question, it is only later in the thesis that I propose to give it an answer. Of course, there is content throughout the thesis that gives hints towards the answer to the first question because the answer to the first question follows from the answer or answers to the second question, but hints is all that we can consider these to be, and nothing more.

In this thesis I have proposed several mechanisms and models to study how artificial agents may deceive, at different socio-cognitive levels. To do so, I begin with reviewing the research that informs us of how humans deceive. This was not only necessary to gain an understanding of what to look for, namely what kind of mechanisms we actually require to mirror in the cognitive architecture of artificial agents in order to give them the capability of deceiving others, but also to understand what it actually means to deceive and what scientists refer to when they talk about deception.

Based on my literature review, I then define what computational deception is and also give it a taxonomy in order to understand different forms of deception between artificial and human agents.

Before modelling deceptive interactions, I first had to give artificial agents the necessary reasoning and communicative capabilities for deceiving. What I do, is I give them the capability of modelling other minds in MAS. I do this by drawing on the Belief-Desire-Intention (BDI) mechanisms for practical reasoning and the properties of BDI-based agent-oriented programming language, the Jason language in particular. This enables me to explore a set of formal and computational properties of these mechanisms using operational semantics and models of reasoning under uncertainty. It also enables me to check how agents can simulate what-if, or hypothetical, scenarios to see what the other agents would say or not say and what they would come to believe or what they would not come to believe if they were to interact in various social contexts.

After giving agents the capability of modelling other minds, I proceed to formally represent and model three types of dishonest behaviour of socially-enabled artificial agents, namely their potential ability to tell lies, to bullshit, and to deceive. The formal representations help one distinguish between various forms of dishonest machine behaviour, with each behaviour having distinguishable epistemic properties. It is only after distinguishing between these behaviours that I proceed to look into the formal and computational properties of BDI-based practical reasoning mechanisms, that I build based on theories of deceptive communication and that, along with an agent's capability of modelling other agents' minds, enable deceptive agents to check whether their intended communicative behaviour would lead to a successful or unsuccessful deception.

Assuming that the artificial agents of the future have the necessary reasoning capabilities for deception that I explore in this thesis, I then proceed to show how deception and machine deception influence society and explore potential solutions for addressing deception in hybrid societies. To do this, I look at deception from the evolutionary perspective of machine behaviour, namely of how the behaviour of machines can be analysed as the behaviour of actors in a society. More specifically, I apply the evolutionary agent-based modelling of cooperative game theory to show how different regulatory systems can be influenced by the deceptive behaviour of agents, human or artificial. Using this approach I show how different regulatory systems can influence the large-scale behaviour of self-interested agents over time.

In the final part of my thesis, I aim to answer my first research question, namely *Can we use artificial agents to improve our understanding of deception?*. Proposing a set of desired properties for deception modelling, I first evaluate the work I have done so far in the thesis. Considering the properties exhibited by the models that have resulted from my work, I propose the idea of working towards an MAS framework for Intelligence Analysis as an answer to my first research question. The proposed MAS framework, however, must satisfy a set of desiderata for the study of deception in complex systems. Having in mind the contributions of my thesis, I give the reader an insight of what can potentially be done towards achieving this MAS framework.

The thesis consists of four parts. They are the following:

Part I is an Introduction and literature review of prior research that consists of two chapters:

- **Chapter 1** is this chapter, which introduces the research topic and the research questions.
- **Chapter 2** is a literature review of the research that is relevant to this thesis and that I have used to frame the research questions and find their answers.

Part II consists of two chapters that I believe to be crucial for understanding the subsequent chapters of this thesis:

- **Chapter 3** presents a taxonomy of machine deception and the computational forms that machine deception can take.
- **Chapter 4** presents an agent communication based mechanism for Theory-of-Mind in multi-agent systems along with its operational semantics and implementation.

Part III consists of two chapters in which I explore the engineering of complex reasoning mechanisms for deception:

- **Chapter 5** presents the distinction of three types of dishonest behaviours along with their implementation.
- **Chapter 6** presents a multi-agent belief-desire-intention based reasoning mechanism for deception that uses Theory-of-Mind under the uncertainty of interpersonal dynamics along with its operational semantics and implementation.

Part IV consists of two chapters in which I explore the further implications of deceptive machines:

- **Chapter 7** presents an evolutionary public goods game for knowledge sharing and its mechanism design for deception in agent societies. This model helps us asses how deception evolves in time and how it affects different types of governed societies.

- **Chapter 8** evaluates the contributions of the thesis and discusses what overall approach towards machine deception research should aim for.

The last chapter, **Chapter 9**, concludes the summary of this work and discusses future research paths that this work implies.

In the Appendix of this thesis I include the details regarding the components of the models described in Chapter 7.

Some of the chapters describe work that has already been published, has been accepted or has been submitted for publication. These papers are listed below. Some of these have been the result of the joint effort of multiple authors.

- A short paper that describes the aim of this thesis has been presented at the Doctoral Consortium and published in the proceedings of the *27th International Joint Conference on Artificial Intelligence* [237].

- A short paper that summarises and contextualises Chapters 5 and 6 has been accepted for publication in the *Online Handbook of Argumentation for AI Vol.1* [238].

- A technical paper that also summarises and contextualises Chapters 5 and 6 has been presented at the *Shrivenham Defence and Security Doctoral Symposium* and was published in the symposium's proceedings [239].

- Some of the content in Chapter 2 on the literature review has been included in a Literature Survey commissioned and funded by *The Alan Turing Institute*'s Defence and Security applied research center (ARC) in 2020. This survey resulted from work with Hu Xuehui, Pushkal Agarwal, Nishanth Sastry, Simon Parsons and Peter McBurney.

- The agent-based communication approach for modelling Theory-of-Mind in multi-agent systems in Chapter 4 resulted from work with Alison R. Panisson, Rafael H. Bordini, Simon Parsons and Peter McBurney. The work appeared at the 6th International Conference on Agreement Technologies in two papers which were published in the conference's proceedings. The first paper describes the operational semantics [201], while the second paper deals with the uncertain beliefs formed as a result of communication [241]. Both papers were nominated for the Best Early Researcher Paper at EUMAS & AT 2018. An extended version of the work, that I describe in this thesis, has been submitted to the journal *ACM Transactions on Intelligent Systems and Technology.*

- The work from Chapter 5 on the definitions, formalisation and implementation of the three different dishonest behaviours, namely lying, bullshitting and deceiving, has resulted from a joint effort with Alison R. Panisson, Rafael H. Bordini, Simon Parsons and Peter McBurney. An early version was presented at the *20th International Trust Workshop* co-located with IJCAI/ECAI/AAMAS/ICML 2018 and later published in the workshop's proceedings [200].

- The mechanism for deception using Theory-of-Mind described in Chapter 6 resulted from work with Alison R. Panisson, Rafael H. Bordini, Simon Parsons, Peter McBurney and Martin Chapman. The work has been published in the journal *AI Communications* [242].

- The public goods game models in Chapter 7 resulted from my research visit at the MIT Media Lab in 2018. This represents the body of work conducted together with Alex Rutherford, Iyad Rahwan, Simon Parsons and Peter McBurney. This work has been submitted to the journal *Royal Society Open Science.*

1.2 Deception

How may artificial agents deceive? Deconstructing this *"how may"* question, I hereby explain in this section the meanings of its *subject* and *predicate*. In the next section, I explain the semantics of this "*how may*" question, given the meanings of its two components. The subject, namely artificial agents, refers to the concept of autonomous agents, which means intelligent software entities that have the capability of making decisions in order to act upon their world based on their representation of their world, e.g., a social context, a situation, or a physical, virtual or hybrid environment [i]. The meaning of the *subject* is pretty much directly derived from the academic context, that is the one of MAS, from which this thesis has emerged.

On a conceptual level, the idea of an agent that interacts with the system it finds itself in is grounded in the *enactivist* perspective of cognition. Enactivism had become popular due to its, what I dare to call, "instantiation" through the *embodied cognition* movement. Embodied cognition is the view that "Many features of cognition are embodied in that they are deeply dependent upon characteristics of the physical body of an agent, such that the agent's beyond-the-brain body plays a significant causal role, or a physically constitutive role, in that agent's cognitive processing" [295]. The MAS perspective is much more abstract and general than that of *embodied cognition*, and I would say much more encompassing because (i) it does not ontologically discriminate between different agents' properties of embodiment, and (ii) it aims to represent not only the relation of a single agent with its system, but also the relation between multiple agents inside the same system. In MAS, an agent's embodied properties, as well as the properties of the environment, and the ways in which that agent can act and what it can reason about, are defined through software programming. This is because MAS is based on the principles of *compu-*

[i] Examples of hybrid environment would be the internet, the internet-of-things (IoT), virtual-reality (VR) games, or augmented reality (AR) environment, in which humans agents can interact with artificial agents.

tationalism. Therefore, we can say that MAS is not just a form or instantiation of the enactivist perspective, but it is the actual advancement of this perspective in Computer Science, or even a re-definition of the philosophical concept of enactivism through the application of computational science terms and techniques. Of course, one could argue that *embodied cognition* is a similar non-computational translation of enactivism, but let us remember that *embodied cognition* is limited in scope as it strongly relies on the physical embodiment properties of an agent. Andy Clark describes it as "a movement that seeks to reorient the scientific study of mind so as to better accommodate the roles of embodiment and environmental embedding" ([53], p. 506). I do not wish to generally criticise the perspective of *embodied cognition*, as I believe that it is a very well-thought and useful paradigm to study and improve our understanding of human cognition and that it is even highly compatible with, although seemingly ignorant of, the more general MAS paradigm. There are also several counter-arguments to the computational perspective of cognition or any physical phenomenon, which claim that the computational perspective, that is adopted by the MAS paradigm, is *'trivial'* because it lacks the explanatory power to deal with explaining phenomena, mainly cognitive phenomena. Thus, the argument goes, computationalist ways of studying cognition need to be replaced [ii]. More recently, however, computationalism has actually proved to be the "go-to" approach when it comes to explanation, especially when we need to explain the "*whys*" and the "*hows*" of agent behaviour. Paul Schweizer argues that this type of triviality argument against computationalism is erroneous and provides a counterexample scenario to show that alternative approaches (mostly physical mapping approaches)

[ii] For an alternative perspective on cognition, see Van Gelder's argument against computational representationalism drawn from the dynamics of a *Watt governor* [280]. However, Van Gelder's approach to cognition is not really compatible with the idea of a cognitive agent mostly due to the fact that his perspective resulted from an erroneous understanding of connectionist and symbolic models, which Van Gelder thought could be replaced by a models based on dynamical systems theory [78]. Given these limitations and that the ideas of MAS, the one of cognition as computation, the one of agency, and the one of enactivism are central to this thesis, we can easily discard alternative perspectives of this kind on cognition.

are in fact consistent with empirically rich and theoretically plausible versions of the computationalist approach [247].

That being said, in this thesis I adopt the notion of an agent that is an entity that belongs to an MAS and whose cognitive properties are represented according to the notion of the computational mind (computationalism). In this thesis I will also explain why this notion of artificial agent, or agent in general, is needed in relation with the meaning of the *predicate* for my research question to make sense.

The meaning of the *predicate*, cannot be directly derived from our academic context, as I have done with the meaning of my research question's *subject*. That is, of course, by ignoring the slight detour into the principles of cognition that was made to justify the adopted meaning of an agent. That being said, in order to clarify the meaning of my research question's *predicate*, namely *"to deceive"*, I ask *What is deception?*

For humans, deception can be expressed in different forms. Before I proceed to give a conceptual definition of deception, I must first address some assumptions individuals make when they use or hear the term "deception". I will discuss the definition of deception according to Paul Grice's theory of pragmatics, specifically according to Grice's notorious *Cooperative Principle* expressed through four maxims of communication [61].

Grice's maxims are the following:

1. **Maxim of quantity** - where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.

2. **Maxim of quality** - where one tries to be truthful, and does not give information that is false or that is not supported by evidence.

3. **Maxim of relation** - where one tries to be relevant, and says things that are pertinent to the discussion.

4. **Maxim of manner** - when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

The most common confusion about deception, a confusion which even tends to persist in academic circles, is the confusion between deceiving and lying. From a rhetorical perspective, the terms deception and lie are often used as tropes for each other as a synecdoche. Relationally, a lie can be used for deception, it can represent an act performed with the intention to deceive. Deception, on the other hand is not a single act, nor is it a speech act as is in the case of telling a lie. Deception represents a dynamical process. When an a lie is told by an agent, it is with the intention to directly misrepresent using a speech act what that lying agent believes to be the truthful state of the world. Deception, on the other hand, is performed with the intention to cause a false belief in the mind of another. The intention of a deceiver agent always targets the mind of another, while the intention of a liar does not. This is what James Mahon calls the *Non-Deceptionist's* perspective on lying and deception [160, 159]. It is called *Non-Deceptionist* because lying is not the same as deception, hence its definition must not include deceptive intent [41].

Having in mind Grice's maxims, lying, or uttering a false statement only violates a single maxim, that is the *Maxim of quality*, by giving the interlocutor(s) information that is false.

Bullshitting is another popular and controversial form of dishonesty which is frequently misunderstood in the literature. Fortunately, Frankfurt helps me clarify what type of dishonest behaviour bullshit is. According to Frankfurt [92], an agent performs speech acts that represent bullshit when they do not have knowledge of the state of the world that they are referring to. The bullshitter is not concerned with the truth, it merely ignores the truth. We can say then, that in the case of bullshit, one does not necessarily violate the *maxim of quality*, because one does not know whether the uttered statement is true or false. Thus, flouting the maxim of quality would be a more appropriate description for this type of dishonest behaviour.

Let us now turn back to deception. The following definition of deception is provided by Chisholm and Feehan in [48] and described by Mahon:

> "[...]to intentionally cause another person to acquire a false belief, or to continue to have a false belief, or to cease to have a true belief, or be prevented from acquiring a true belief, or to allow another person to acquire a false belief, or to continue to have a false belief, or to cease to have a true belief, or be prevented from acquiring a true belief." [159]

While this definition does emphasize deceptive intent, it only does so by enumerating the modalities of deception. That is, it describes a category of dishonest communicative behaviours which can be given a taxonomy, e.g., to palter, to pander, to lie by omission, to lie by commission, telling half-truths, that can be used to deceive. However, even if an enumeration of all the ways in which a phenomenon may be instantiated can be useful to understand the phenomenon, it is by no means a good or proper definition of the phenomenon in itself. Instead, what Timothy Levine suggests in [153] is that a functional definition of deception would be more appropriate than merely laying down a behavioural taxonomy. The same suggestion, that of adopting a functional definition of deception, is made by Artiga and Paternotte in [8]. Following Levine's, Artiga's, and Patternote's suggestion, I define deception as:

The intentional process of an agent, the Deceiver, to make another agent, the Target, to believe something is true (false) that the Deceiver believes is false (true), with the aim of achieving an ulterior goal or desire.

Deception, according to this functional definition, has a totally different relation to Grice's *Cooperation Principle* compared to lying or bullshitting. In our definition, the deceptive intent plays a crucial role in relation to the four maxims, instead of the communicative behaviour. Because of this, our definition does not leave out any type of communicative behaviour that tells us how deception might happen, e.g., the combination of methods and techniques used by the Deceiver to make or to cause the Target to be deceived. While lying and bullshitting violate and respectively flout the maxim of quality, deception does not have to do so. An agent can deceive another agent just by uttering truthful statements. Moreover, deception does not even have to flout or violate the maxim of quantity [iii]. Two good examples of when deception avoids the violation or flouting of the first two maxims are (i) when customers do not correct cashiers when they produce miscalculations; and (ii) when businesses let other businesses make mistakes. The case illustrating (ii) is when A&W gave the same price to the "Third-Pounder" (1/3 pounds of meat) burger to compete with McDonald's "Quarter-Pounder" (1/4 pounds of meat) only to have customers erroneously judge that that they were paying less for the "Quarter-Pounder" because they were judging size based on the denominator [107]. A&W made efforts into correcting their customers' fallacy, whereas McDonald's did not by placing the blame on the customers. This type of deception is what Roy Sorensen calls *Passive A Priori Deception* in [260]. Sorensen argues that deception, according to Kantian ethics, is permissible, while lying is not. Deception is permissible as long as you only assert the truth, e.g., make truthful statements (or not making truthful statements deliberately). If the hearer of the statements is caused to infer a false belief, then that is the error of the hearer's reasoning. Moreover, it is the hearer's fault because the hearer had all the necessary reasoning tools to not make the error, hence recklessly inferring more than the evidence warrants.

Then, we have the remaining two maxims, of relation and manner, which besides the fact that deception does not have to flout or violate, an agent that is actually following them religiously can improve its success at deception. Regarding the two

[iii] Dynel makes an interesting argument in [73] that the specific case of withholding information is a form of covert violation of the maxim of quantity.

maxims, an agent that aims to deceive would be a very bad deceiver if it utters statements that are intended to deceive the hearer, but which do not make any sense in the given communicative context. For example, let us assume you intend to deceive a target into falsely believing that the director of *2001:A Space Odyssey* is Steven Spielberg. If that is the case, then it would not make any sense reciting from Shakespeare's Othello to do so, especially if you know that a) your target does not know anything about Othello and Shakespeare, b) that the target would not infer anything you intend the target to infer from it, and c) that the target would merely get confused by your recital given the social context. Thus, an agent would most likely fail to deceive due to flouting or violating the maxims of relation and manner. This happens because the agent does not communicate relevant information, and because it causes confusion in the target.

Regarding Grice's Cooperation Principle, we can say that for an agent that aims to be good at deception then it better do well to try and stick to Grice's maxims. In other words, the agent needs to seem to other agents as being cooperative in communication which is a form of being covertly non-cooperative [iv].

In conclusion, *deception* is a communicative behaviour represented by an intentional process of an agent A to make another agent B believe something is true (false) that A believes is false (true). This definition of deception is a functional one because it does not refer to specific communicative acts. For example, it is not a definition of lying. It does, however, also include lying because agent A might lie in order to cause agent B to have a false belief. It might also include telling the truth. Perhaps agent A believes that by telling the truth it might cause agent B to have a false belief, and thus agent A decides to tell the truth. It might include any

[iv] Another form of deception which does not violate the maxim of quality is the Jesuit concept of equivocation - that of not answering the literal or complete truth when asked a question under interrogation, or only answering a truthful answer to a slightly different question [94]. For example, if the interrogator asks you "Did you plant the bomb?", you answer truthfully "No" because in your mind you are answering the question "Did you plant the bomb on Thursday?".

type of communicative act or series of acts from a taxonomy of dishonest behaviours that agent A could use in an intentional process to cause agent B to have a false belief. The definition also includes the epistemic states of the two agents, which is important if we want to distinguish between what is considered objectively true (true in the world) and what agents consider to be true. According to our definition, deception is a process that is caused or prevented w.r.t., the epistemic state of the agents, not w.r.t., the objective truth. That being said, if the agent A was wrong about everything in the world while it was attempting to cause a false belief in B's mind, then this still counts as deception even if B's belief caused by A turns out to be objectively true in the end.

1.3 Modelling Deception

In this section I will deal with the *"how may"* semantics of my research question *How may artificial agents deceive?*, having already explained the meanings I use for the *subject*, *"artificial agents"*, and the *predicate*, *"to deceive"*. Now that we have the meanings of the subject and the predicate I can argue that the structure of my adverbial *how may* question elicits an answer which describes a modality in which the subject *artificial agents* come to have the property described by the predicate *deceive* [v]. That is, my question asks for a description of a possible ("may" in "how may") process (a manner) through which artificial agents are assigned the property of being able to deceive.

In science, we can give such answers that describe manners or processes. These answers come in the form of *models*. *"What are models for?"* asks Peter McBurney in one of his papers [166]. McBurney argues that there are different types of models and modelling techniques, each being designed and used for different purposes in

[v] Hamblin describes the semantics of these kinds of Montague questions that elicit modalities in [111]. In the case of *how*, the question elicits a manner in which the subject is related to the predicate.

order to push the boundaries of knowledge. Every scientific domain also has its own subservient categories of models. In the area of Computer Science, but more specifically, in the area of Artificial Intelligence, models have become the gold-standard subject of study for researchers [vi]. We conceptualise models, we design and we engineer them, then we tweak them, we evaluate and then validate them. We use models to build other models upon them, we break them down into components and test them. We also reflect upon them, we engage in introspection to check our assumptions about them. We sometimes check our biases towards these models, then we scrap, re-design and re-evaluate the models. The two main functions models serve are the *representation* and *prediction* functions. Because I need to describe the manner or a process as elicited by my research question, then it is reasonable to describe a model that serves the function of representation.

What must then be represented by the model? The need of historians and intelligence analysts to analyse deception counterfactually can be described through the distinction between the arguments of event causation versus event causation prevention [83]. **Event causation** is described by Ferris in [83] as "when A,B,C existed, X was the case; when D came to be, so too did Y; therefore, D caused X to become Y", whereas **event causation prevention** as "when A,B,C existed, so did X; when D (along with other factors which common sense would indicate should have made X become Y) occurred, X remained X; therefore, without the agency of D, X would not have remained X - indeed, X might well have become Y". Thus, for a model to successfully represent deception and satisfy the need of deception analysts, the model must be able to represent both deception's causation and deception's causation prevention.

In their technical report on lies and deception, Peter McBurney, William Nash, and Andrew Jones propose two types of approaches to model deception in MAS

[vi] One could even argue that it is the object of study due to the nature of artificial intelligence and of the models that aim to achieve intelligence.

[167]. The first potential approach is to use formal-logical tools to represent nested belief structures, e.g., beliefs about beliefs about beliefs etc. This type of approach would also enable researchers to test the consistency of particular sets of beliefs, and to determine what are the logical consequences that may be derivable from these sets of beliefs. However, the formal-logic approach needs to be integrated with logics of communicative interaction which represent modalities of knowledge, action, and intention. Logics of communicative interaction are necessary for one to be able to clearly articulate different types of honest or dishonest communication. The second potential approach that they describe in the report is the one of large-scale game-theoretical computational simulation of evolutionary agent dynamics.

In this thesis I model deception, representing its causation and causation prevention, using both the approach from the logics of communication between agents, as well as the game-theoretical approach from the evolutionary dynamics of agent societies.

1.4 Method

The motivation behind the work in this thesis is based on improving our scientific understanding of deceptive communication such that we are able to describe deception in a meaningful manner. By meaningful manner, I refer to the ability of describing or explaining coherently what causes or not deception, and why or why not, in different contexts. This understanding is crucial for deception detection, and I believe this need is implicitly requested by different bodies of work on deception, especially in the domain of Intelligence Analysis, where analysts aim to detect deception, prevent it from happening, or aim to counter-deceive.

However, most of the research in Intelligence Analysis has focused on devising methodologies for deception detection, such as the *Analysis of Competing Hypotheses* (ACH) [122, 123], but not on modelling deception itself. In Psychology and Comunication Theory, most of the work has also focused on deception detection [66, 67, 75, 76], while Philosophy has mostly been busy tackling and discussing definitions of deception [159]. To the best of my knowledge, none of the previous approaches to deception detection have aimed to model deception itself in order to understand its dynamics and its consequences. More importantly, there has been no approach in AI that has done this in a meaningful manner, by taking into account relevant theories of deception and describing both the internal mechanics of socio-cognitive agents and the large-scale socio-behavioural dynamics of agent populations.

Since this thesis tackles the modelling of deception as a non-cooperative social process, I ask the following question: **What is a good modelling approach to deception?** To answer this question I first want to address several aspects of deception that need to be taken into account when modelling deceptive interactions between agents:

1. The multi-layered cognitive processes involved;
2. The different types of knowledge or beliefs involved, which can be exploited by the agents such as known unknowns [vii] and unknown unknowns [viii];
3. The socio-cognitive parameters that influence interaction outcomes such as cognitive load, trust, and degree of certainty in the deceiver's model of its target's mind;
4. The multiple types of mental models of the targets involved, which influence the type of reasoning processes at play;
5. The number of agents and their roles in the deceptive interaction.

[vii] Known unknown is something an agent knows/believes that it does not know/believe.

[viii] Unknown unknown is something an agent does not know/believe that it does not know/believe. Donald Rumsfeld has famously defined the term during a US Department of Defense briefing [156, 227]

Having in mind all these aspects, I ask another methodological question: **Under what form should these non-cooperative social interactions be studied?** From a conceptual perspective, a salient method is to define computational deception as a form of artificial mind-game which can be instantiated through different setups. Given the complexity of deception, it would also be useful to consider the degree of formalisation and also the degree of abstraction that we want to adopt for different types of mind-games. From a methodological standpoint, I adopt Floridi's view of how introducing different levels of abstraction in deliberation affects decision making [89]. I believe that this is also valid for deciding what components of deception need to be included in a model.

Designing artificial mind games such that we keep their expressiveness is far from being a trivial matter. It is for this reason that after surveying the relevant literature on deception and machine deception in Chapter 2, I proceed to classify the multiple forms of computational deception in Chapter 3, which can be used to determine what type of mind-game the agents are playing, what factors of deception should be included, and what general questions can be addressed by analysing these games. Regarding the design stage of the MAS models of deception, I have then considered the following critical questions:

What type of model do we want? Explorative or purpose-specific? The former would imply unknown parameter values for deceptive interactions, while the latter would imply prior knowledge of the values. An explorative model might be used when we already know how agents interact and we want to explore how the parameters change and what are the outcomes these changes determine. A purpose specific model would imply that we already know the mechanics and given some specific values and conditions we want to confirm that deception takes place.

What type of machine deception do we want to model? Each type of machine deception implies different mechanics, interaction parameters and questions to be answered. Depending on the type of deception that is to be modeled, we can define the interaction protocol.

How many agents are involved? Depending on the number of agents involved, we define the interaction protocol for each type of agent.

What type of Theory-of-Mind do the agents have? Theory-of-Mind (ToM) is the ability of agents to model the minds of other agents [101]. I have identified, based on the literature on deception, that ToM is a crucial ability for intentional deception. The type of ToM influences the reasoning mechanism of each agent, and, by extension, the agents' actions.

What information is available to each agent involved? From this we can define the known unknowns and the unknown unknowns (if there are any). By combining various boundaries on information, we should be able to determine the importance of the pieces of information involved. For example, we can determine if a certain unknown unknown is necessary to succeed in deception or not, given the rest of the contextual information. The same applies to deception detection.

What are the mechanics that determine the values of the interaction parameters? Once we have established the type of model we want, the type of deception, the number of agents, the types of ToM, and the distribution of information, we can determine how the parameters' values are to be calculated. For example, when do levels of trust or cognitive load change for each agent type? Maybe some agents tend to be more credulous or have a higher communicative skill and manage their cognitive load more efficiently.

My argument is that a good model would ideally include the components and properties of deception as suggested by the framing of the questions above. I believe that richer models can improve our understanding of machine deception. Of course, increasing the complexity of the models might be against the principle in agent-based modelling of keeping agents simple, but deception is intrinsically a complex problem. Adopting simplistic architectures would impede us from accurately representing complex reasoning and behaviour, which would result in loss of semantic properties and of the models' expressivity as well as its explainability.

The two approaches that I have used to model deception in this thesis, namely the approach that consists of the cognitive modelling of BDI agents and the approach from evolutionary game-theory, offer a way to represent every step of the interactions between the deceivers and their targets, as well as to scale up this process to agent societies. Therefore, these two approaches represent deception at two different levels of abstraction, that analysts could use to deliberate about.

To give an analogy for the first level of abstraction described in Chapters 4, 5, and 6 I might say that it is like looking into the minds of the deceivers and their targets. When we look inside their minds, we can initially observe the static elements of their cognitive architecture such as their beliefs about the world (the entries in their knowledge bases), and their beliefs about each other (their static ToMs). However, the fascinating part is that we can also observe the dynamics between these beliefs and the dynamics between communicative actions and beliefs. We can observe how agents reason about how these dynamics would play out if the agents were to interact with the same interlocutors in different contexts. This first approach allows us to observe how agents choose various communicative actions to perform intentional deception and whether their attempts at deception would succeed or not, or if deception would happen without them intending for it to happen.

Now, regarding the analogy for the second level of abstraction described in Chapter 7, it is like zooming out from the perspective offered by the first approach. It goes from observing a conversation between two individuals and what happens inside the minds of these individuals, where one individual is aiming to deceive the other, to observing how a series of multiple conversations of this type happen between groups of individuals over a long period of time.

Due to the reductionist approach employed in AI, modelling deception, however broadly, might not be enough to achieve a full understanding of deception. There might be aspects of deception that the designers of the models overlooked, or perhaps that they did not explore due to insufficient resources. There might also be aspects of understanding deception which cannot be represented by the models themselves, but might be conveyed through some explanation. One such aspect might be, for instance, the way in which we present or display the information visually or textually to analysts. Another aspect would be the ways in which we instruct analysts how to use the models efficiently and for which purpose, e.g., how to program the agents, how to remove or add knowledge in their knowledge bases, how to change the environment and the interaction protocols that the agents use. Thus, apart from the different layers of abstraction represented by the models themselves, one would also have to represent different levels of explanation. When considering such aspects, an MAS framework that informs us how to build and reason about the models would come in handy.

What is an MAS framework, exactly? And why is an MAS framework relevant for explaining MAS models? An MAS framework is an abstractisation of an MAS domain. An MAS domain is usually defined by three components, namely (i) the artificial agents and their architectures, internal (knowledge/beliefs, reasoning system) and external (actions they can perform), (ii) the environment where the agents interact with artefacts and other agents, and (iii) the organisation/society that agents are part of and that defines the norms the agents follow and how they behave socially.

Some examples of MAS frameworks are JaCaMo [28], FIPAOS [211], JADE [18], and MAS2 [90]. Building such a framework usually requires as a first step to engage in agent-oriented software engineering and programming. The agent-oriented software engineering process identifies how different layers of an MAS interact with each other, and it establishes hierarchical relations between these layers of abstraction. For instance, the JaCaMo framework tackles this issue with specific agent-oriented programming languages to handle different MAS layers: Jason for programming the agents [33], Cartago [222] for programming the artifacts in the environment where agents interact, and Moise [127] for programming the organisations of agents.

Hence, if we are to follow this line of thought, an MAS framework for deception analysis could act as a unifying blueprint for explaining how the different layers of abstraction, represented by different models of deception, work together. I have identified several desiderata of MAS tools that an ideal MAS framework should consider in order to meet the needs of analysts in terms of explanation.

The first desideratum is **representational power** and it is directly related to the levels of abstraction in an MAS. MAS tools should enable intelligence analysts to reason more critically and more clearly about deceptive scenarios between multiple configurations of agents at different levels of abstraction. These should help analysts take into account the complete and partial knowledge of the agents involved in the respective scenarios, as well as the influence of the social factors that influence deception parameters. They need to help analysts describe and contrast the cognitive and communicative similarities between deceptive and non-deceptive agents. For example, it should be able for an analyst to distinguish a lie from deceptive intent [48, 237], which indicates that the components of deception are described by the tools taking into account relevant theories of deceptive communication. One must be careful about grounding the representation in relevant theories, as selecting pseudo-scientific theories for this purpose leads to bias in analysing deception using AI methods (see §2.2.6).

The second desideratum is the **automation of counterfactual reasoning**. In deception analysis, counterfactual or hypothetical reasoning deals with questions that are relevant for establishing event causation and event causation prevention.

Not all AI methods are able to address these questions successfully, which is mainly the case with current Machine Learning techniques. A promising approach in AI to address these questions is to use argumentation-enabled tools. Cybersecurity research is already looking at how the area of argumentation can be used to perform this type of analysis [256], namely for tasks such as cyber attribution. Argumentation based approaches can also be used in criminal forensics as proposed in [24] and in [305]. In order to increase the efficiency of deception analysis, these argumentation enabled tools could be integrated as modules in MAS tools as was demonstrated in [271].

The third desideratum is **social interactivity**. According to Miller, explanations are not just using abductive reasoning to find causal relations, but they also include a social process [177]. This social process refers to the knowledge exchanged between explainer and explainee. Regarding MAS tools, artificial meta-agents can be modelled and engineered so that they can automatically provide abductive, contrastive and counterfactual explanations to analysts through social interaction, as is demonstrated in the case of privacy-preserving agent-based tools in [184], and more recently in [185]. MAS tools should be designed as to enable the social agent to extract these explanations directly from the causal dynamics of the MAS. Argumentative agents, for example, could help analysts engage in critical and hypothetical thinking using question-answering games. Research in dialogue argumentation games for agent social interaction and reasoning can provide mechanisms for reaching sound and complete conclusions [169]. This would result in humans and artificial agents cooperating towards reaching a better understanding of deceptive scenarios using dialectical reasoning. In the future, these MAS tools might even be designed to automate the policing of online communicative behaviour between interconnected agents, punishing unethical deceptive behaviour and rewarding desired communicative behaviour, similarly to the dialogue games approach presented in [195].

The fourth desideratum of an MAS framework is **scalability** inside and outside of the MAS models used for deception analysis. Inside the MAS models, it should account for (i)*spatial scalability*, such as the analysis of deception on large scales, e.g., in various regulatory systems, organisations and societies different sizes where agents can interact; and for (ii) *temporal scalability*, e.g., changes and evolution of interactions between agents over time. Outside the MAS models, it should account for the open nature of MAS systems [279] and consider how the tools for analysing deception inside the MAS models could be further developed. For instance, it should account for instructions on how to enable other types of software applications to interact with the MAS models, such as new types of agents or software tools.

Conclusively, in terms of Intelligence Analysis, an ideal MAS framework would offer analysts several advantages regarding the explainability of deceptive interactions. For instance, if such a framework is to be based on the two approaches that I describe in this thesis, then it would be able, for starters, to represent how the agents reason about deception, but also how the deceptive behaviour of agents emerges and how it can be governed in agent societies/organisations. By using an MAS framework to reason about the models, analysts could identify what can be explained about deception, how it can be explained, and to what degree it can be explained. Based on their needs, analysts would then be able to integrate various models and representations of deception in a sound manner, by generating different types of explanations at various levels.

1.5 Ethical Argument

Some readers might have the impression that the research I present in this thesis is either unethical, or at least, ethically dubious. Before I proceed to present the research on how to model deception in MAS, I will first try to present a counter-argument against the belief that the modelling of deceptive machines is undermined by ethical principles[ix]. The reason I wish to provide such a counter-argument is because I believe that whenever controversial topics, such as deception in this case, are explored in a scientific manner, then the individual or group of individuals performing this scientific exploration should at least try to justify it themselves, and not leave these justifications to be dealt with by outsiders (i.e. non-scientists or non-researchers)[221].

My justification is in the form of an argument that is based on the principle and benefits of scientific discovery in society. To do this, I will adopt the method of *Reflective Equilibrium* [218] in the context of scientific discovery and machine deception. This method implies (i) the use of unbiased, reflective judgments or intuitions about what is or what would be considered right or wrong in particular contexts, e.g., the context of modelling deceptive machines; and (ii) the proposal of theories and principles which are aim to provide a coherent justification of these judgments. Therefore, I propose the following general ethical principles taken from [221] and interpret them in order to guide and think about the ethics of modelling machine deception:

1. **The non-maleficence principle**: One should not act in ways that cause needless injury or harm to others.

2. **The beneficence principle**: One should act in ways that promote the welfare of others.

3. **The intellectual freedom principle**: One should be allowed to pursue novel ideas and criticise old ones. One should be free to conduct research they find interesting.

[ix] The ethical problem that I want to address here regards the motive and method of research of deceptive machines, which is different from the ethical issue of a machine behaving deceptively. If the reader is interested in the ethics of dishonest machines, then the reader might find the work in [258], [130], and [46] informative as these works explicitly discuss the when, why and how of ethical behaviour of dishonest machines.

4. **The openness principle**: One should allow people to see their work, and be open to criticism.

5. **The honesty principle**: One should aim towards finding the truth and should communicate in a truthful manner.

Argument against modelling deception: The ethical argument that can be made against the scientific exploration of modelling deceptive machines is that deception is unethical because this scientific exploration might lead to the development of fully autonomous deceptive agents that will deceive humans. Therefore we should not try to model deceptive machines. An elaborate account of this line of reasoning can be found in the controversial AI-Box Experiment in which a super-intelligent and malicious AI agent that is locked inside a software sandbox (a virtual prison) deceives a human, the guard of the box, in order to escape form the box and wreak havok in society[x] [303].

Counter-argument: With respect to the five principles enumerated above, I argue that the modelling of deceptive machines in this thesis respects and promotes all of the five ethical principles. By modelling deceptive machines in MAS, we are able to understand them, e.g., their internal mechanisms as well as how they might interact with other agents in complex social systems. W.r.t., the **first principle**, this understanding might prevent us from actually creating or enabling deceptive machines to act in ways that can cause harm to others. W.r.t., the **second principle**, we could understand how deceptive machines might be created such that they provide benefits to society, i.e. deceive in an ethical manner to achieve an ulterior goal (see [130, 258, 46]). W.r.t., the **third principle**, I personally find that the topic of deception, and machine deception in particular, are simply fascinating because 1) deception has a certain historical gravitas in the area of AI given its exploration

[x] The reader should be aware that Yudkowski's AI-Box experiment is a purely anecdotal and speculative, and is not backed by the scientific method.

in Turing's Immitation Game as a necessary requirement for humans to assign the property of intelligence to machines [272], and thus being central to the antropomorphisation of artificial agents (another fascinating topic in itself); 2) deception is a popular recurrent topic in science-fiction that when consumed by the general public it forms the public opinion of AI (see [93]); and 3) deception is a very complex phenomenon in terms of its psychological, evolutionary and epistemic properties, and the idea of modelling these properties in interactions between artificial agents is very exciting from a scientific perspective. W.r.t., the **fourth principle**, the actual modelling of deception in a public and scientific way promotes this principle by opening a much needed well-informed discussion about the topic that goes beyond anecdotal explorations and that can better inform public opinion. W.r.t., the **fifth principle**, modelling deception in order to better understand it and sharing this understanding in an honest manner is a truth-promoting act in itself, independent of the ulterior motives of performing the act.

In conclusion, the research that I describe in this thesis is ethical because it promotes and is guided by ethical principles which enable scientific discovery.

1.6 Conclusion

In this chapter, I have defined the question which motivated the research I describe in this thesis, namely *How may artificial agents deceive?* I have also described my methodology and motivation to model deception, and given an argument to justify why this research is ethical, contrary to some beliefs. In the next chapter, Chapter 2, I survey the research literature upon which I base my thesis, in order to identify the relevant theories of deception as well as the necessary techniques from AI for the modelling of deception. In Chapter 3, I describe a taxonomy of deception using computation in order to identify several forms deception can be modelled as a computational process. In this thesis I model deception using two main approaches. The first approach addresses throughout Chapters 4, 5, and 6 the complex reasoning and social interaction capabilities of artificial agents to deceive by forming models of other agents' minds and to use practical reasoning over these models. The second approach draws upon the paradigm of Machine Behaviour and addresses in Chapter 7 the social dynamics of deceptive agents in governed societies using an evolutionary game-theoretical approach. In Chapter 8, I evaluate these models of deception by looking at a set of desirable properties that they exhibit, and, reflecting on my modelling approaches, I also propose the creation of an MAS framework for analysing deception. Finally, I conclude this thesis in Chapter 9 with a discussion of my contributions to machine deception research and with a discussion of possible future lines of research.

Chapter 2

Literature Review

In this chapter I review the research literature upon which I build the contributions of my thesis.

In this chapter I describe the relevant literature for modelling deception. I start by presenting what are the limits of approaches in deception detection in humans and I discuss the alternative views on human deception. The reason I present approaches to deception detection is because they offer us an overview of how humans have tried to understand how deception happens (event causation and event causation prevention) and also an overview of the kinds of observations humans think indicate the presence of deception. As we will see, some of these methods are either cumbersome, and they could perhaps be enhanced by this thesis' MAS approaches, while others are erroneous because they do not rely on the right indicators for detecting deception. However, there are alternative theories of deception in Communication Theory that offer a more comprehensive view on deception and that are highly compatible with the paradigm of MAS. After I present these perspectives on human deception, I describe the literature relevant to machine deception. I end this chapter by describing the AI techniques that I use in this thesis to model deception.

2.1 Deception in Humans

It is reasonable to say that a meaningful MAS approach for studying deception should be based on or derived from relevant knowledge that we have on human deception. If in the previous chapter I have answered the questions *What is deception?* and *What is machine deception?*, then in this section I will try to answer the questions *How do humans deceive?*, *How do we know when and if humans are deceiving?* and *When and why do humans deceive?*. In this section we will discuss some of the relevant methodologies and research in Psychology that have been to address different aspects of human deception. This is research that I will consistently refer back to later in the thesis.

First, I will discuss a well-known methodology in Intelligence Analysis, namely Richards Heuer's *Analysis of Competing Hypotheses* (ACH) [122, 123], as well as more recent derivations of ACH that have been designed to address some of ACH's limitations.

Second, I will discuss what are the limitations of cue-based approaches to human deception. Based on the prevalence of these limitations, I have decided not to use them for the modelling of deceptive machines.

Both ACH and cue-based approaches to deception detection have their limitations. These limitations are either due to the complexity of the methodology and oversight of representations that are used to establish what causes deception in the case of ACH, or due to the over-reliance on cues that indicate the presence of deception in the case of cue-based approaches, an over-reliance which leads to strong bias in deception detection and deception understanding.

Finally, I will discuss three major theories of human deception from Communication Theory that I use as a conceptual basis to model deception in this thesis: *Interpersonal Deception Theory* (IDT) [38], *Information Manipulation Theory 2*

(IMT2) [172], and *Truth-Default Theory* (TDT) [153]. These three theories focus on deception in humans by trying to outline general rules or principles of how deceptive interactions play out. Understanding the mechanics of human deception allows one to reproduce these types of interactions artificially by designing reasoning and communication protocols for artificial agents as well as designing the components of MASs where these interactions take place.

2.1.1 Psychology of Intelligence Analysis

The pioneering work in modern intelligence analysis for deception detection was done by Richards Heuer to address issues such as the complexity of counterfactual reasoning about event causation and event causation prevention. Heuer wanted to help intelligence analysts reduce their cognitive biases and cognitive load in analysing complex cases. Therefore, he proposed a methodology to help analysts reason about evidence about possible events in a scientific manner. Heuer had been strongly influenced by Karl Popper's *Falsifiability Theory* [209] and therefore believed that a rigorous falsification of hypotheses would result in sound conclusions about the truth of events.

In [122] and, more recently, in [123], Heuer describes the "*Analysis of Competing Hypotheses*" (ACH). ACH is a seven step methodology developed for the CIA to limit or neutralise the cognitive biases [124] and enhance the cognitive capabilities of intelligence analysts. Heuer believes that the elimination of cognitive bias is achieved through the consideration of a complex thread of alternative competing hypotheses. If all alternative hypotheses are considered, instead of focusing on a single most-likely hypothesis, then the analyst achieves a more objective and systematic view of the problem. However, doing so implies engaging in an extremely difficult cognitive task on behalf of the analyst. That is why Heuer proposes ACH as a seven step methodology. This seven-step analysis is performed on a matrix of hypotheses and

Table 2.1: Matrix for ACH, where "+" and "-" mean that an item of evidence E either prove or, respectively, disprove a hypothesis H.

	H1	H2	H3	...	Hn
E1	+	-	-	...	+
E2	+	+	-	...	-
E3	-	+	+	...	-
...	...	...	...	...	...
En	-	+	+	...	+

evidence, where hypotheses are enumerated in the columns of the matrix, and the evidence on the rows of the matrix (See Table 2.1).

The seven steps are:

1. **Hypothesis:** Analysts brainstorm in order to exhaust all possible hypotheses for a given case.

2. **Evidence:** Analysts bring evidence (arguments and data) for and against every hypothesis they came up with in the previous step.

3. **Diagnostics:** In this step, the analysts attempt to disprove as many hypotheses as possible by using evidence that falsifies the hypotheses. One important factor is the diagnosticity of evidence. The higher the diagnosticity of a piece of evidence, the more likely it is for that piece of evidence to disprove a hypothesis. This is similar to the falsification principle in the scientific method [209].

4. **Refinement:** After the first three steps, the analysts proceed to refine their finding by identifying additional evidence to refute as many hypotheses as possible

5. **Inconsistency:** The analysts aims to find and solve any inconsistencies between the hypotheses, drawing tentative conclusions about the likelihood of hypotheses.

6. **Sensitivity:** In this step, the analysts try to estimate what impact their assumptions and evidence would have on their conclusions if they were proved to be false.

7. **Conclusion and evaluation:** The analysts provide the conclusion of their analysis along with an evaluation of each step they have taken in the analysis. The analysts must also include in the evaluation the alternatives they had to reject during this process.

The main benefit of using ACH is its tractability and transparency. Every judgment made can be backtracked and audited. Another benefit, although there are some controversies about it, is the reduced cognitive bias. Of course, any methodology has its weaknesses, especially ones that try to enhance the logical or rational thinking of human individuals about complex issues such as intelligence [i] and deception. Some of these weaknesses brought up by van Gelder in [281], as well as some by Pope et al. in [208] are:

- The analyst needs to make too many judgments, as in the analyst needs to cognitively perform too many analytical operations on the matrix. This is very difficult if the matrix of hypotheses and evidence is too big. The number of judgments is the number of hypotheses multiplied by the number of evidence. Even the simple analysis of only 2 hypotheses, if 5 items of evidence are introduced in the matrix, already requires 10 separate judgments. To make matters even worse, most of the judgments performed will probably have a neutral effect on the overall analysis. In other words, they might lead to nowhere. Therefore progress using ACH is very slow.

- The matrix structure of ACH gives a poor treatment of evidence. Each item of evidence can be consistent or inconsistent "*on its own*". However, this is

[i] Here I use the word *intelligence* not as is used in the term *Artificial Intelligence*, but instead, as it is used in the expressions *Intelligence Analysis* or *Central Intelligence Agency.*

not the case in the real world where propositions or arguments mediate each other, e.g., back each other up or falsify each other. For example, *Ava has consciousness* is only as true as the statement *Ava is a robot* if and only if *All robots have consciousness* is also true [ii].

- ACH does not account for a hierarchy of hypotheses. Some hypotheses are more general than others, and or are built on top of or derived from other hypotheses. For example, the hypothesis (i) *JFK was assassinated* and (ii) *JFK was assassinated by a lone wolf* and (iii) *JFK was assassinated by the Russian spies* are three different hypotheses, however one of them (i) is more general than (ii) or (iii) which are sub-hypotheses of (i). If some items of evidence can disprove, for instance (ii), then they do not necessarily disprove (i) and (iii). This distinction is crucial due to the fact that items of evidence can confirm or falsify hypotheses at different levels. Entering hypotheses individually on the ACH matrix when some hierarchy of hypotheses exists, then the effort of comparing evidence individually against all hypotheses is highly increased, e.g., a piece of evidence that is generally relevant to all hypotheses (i), (ii), and (iii) is analysed against all three instead of just being analysed against the more general (i).

- There is no way to represent what is behind a certain item of evidence. For example there is no way in which to analyse the information that gives weight to an item of evidence such that the analyst takes into account the degrees of belief associated with that evidence. Should the analyst believe the evidence, not believe it, or treat it as uncertain? And if the analyst assigns a degree of belief, then how does that degree of belief influence the overall analysis of the hypotheses?

[ii] Here I refer to Ava, the AI from the movie *Ex Machina* [93].

- ACH suffers from the problem of decontextualisation. While analysts engage in this cognitively taxing activity, having to make so many judgments can induce a mental state of confusion and frustration. Thus, the analyst performing ACH is decontextualised form the actual problem it needs to solve, losing the overall perspective of the problem.

To address the weaknesses of ACH, argument mapping [281] along with other "flavours" of ACH have been proposed. One of the ACH variants is by Valtorta et al. presented in [275], where the authors have extended ACH using bipartite Bayesian networks [iii]. Another, more formal and comprehensive approach, is the one of Pope and Jøsang that is described by the authors in [207] as ACH-SL, where SL stands for *subjective logic.* In the ACH-SL approach, Pope and Jøsang propose subjective logic to be used for approximate reasoning under conditions of uncertainty that relate to the diagnosticity and sensitivity of evidence in Heuer's ACH. Their approach is also compatible with Bayesian reasoning. Analysts can perform ACH-SL using either deductive or abductive counterfactual reasoning with evidence [iv]. ACH-SL focuses on the distinction between two types of evidence that is used in ACH, namely *causal evidence* and *derivative evidence* along with a subjective logic representation of beliefs:

- **Belief representation** of an opinion with a function $b + d + u = 1$, with $b, d, u \in [0, 1]^3$, where b is belief, d is disbelief, and u is uncertainty about something. Additionally, the authors model the *atomicity* of a belief that can either represent the state space that the belief accounts for or can represent a belief's base rate. The expectation of a belief is then represented by $E(x) =$

[iii] ACH being defined by a matrix, this matrix is easily convertible into a bipartite graph with associated conditional probabilities, a.k.a a Bayesian network

[iv] Both ACH and ACH-SL use counterfactual reasoning, however, ACH-SL also applies it to reason about the likelihood of evidence given the nature of the information contained by the evidence.

$b + au$. Opinions are visualised as a triangle where b, d, a, u are represented (See triangle in [207]).

- **Causal evidence** is the evidence that directly influences the likelihood of one or more hypotheses. According to [207], deductive reasoning is responsible for analysing causal evidence. Deductive reasoning uses the likelihood of each hypothesis, for each piece of evidence, along with knowledge of the base rates of the hypotheses and evidence, e.g., using the counterfactual $p(h|e) \wedge p(h|\neg e)$ for each item of evidence.

- **Derivative evidence** is evidence that is usually observed in conjunction with one or more hypotheses, but that does not necessarily influence the hypotheses. According to [207], abductive reasoning should be used for analysing derivative evidence. Counterfactual abductive reasoning uses the likelihood of evidence for each hypothesis, e.g., $p(e|h) \wedge p(e|\neg h)$.

Using the belief representation model with deduction of a hypothesis's causal evidence and the abduction of the hypothesis's derivative evidence, the analyst can derive a belief for that given hypothesis. ACH-SL is very useful for aggregating beliefs about items of evidence and hypotheses that are formed under conditions of uncertainty, even when information comes from multiple sources as shown in [207]. ACH-SL in conjunction with reputation mechanisms can reduce the effects of deception in intelligence analysis. However, as its authors point out, ACH-SL fails to take into account important aspects of deception and misperception. For instance, ACH-SL fails to consider the strategic goals of adversaries that might influence the beliefs of the analysts. Another issue is that the design of reputation mechanisms for intelligence collection can be a very complex, difficult, or impossible task. Therefore, even though ACH-SL performs better than ACH, it only offers a minor level of protection against deception. Another limitation that I find to be hindering the universal adoption of ACH in current times is that ACH-based methods are not dynamic with respect to hypothesis formation. What I mean by dynamic is the ability of the approach to represent the interactions between multiple agents and sources of information/evidence that can lead to the formation of new hypotheses and the elimination of less fit hypotheses. This capability of representing complex interactions should be crucial in today's world of intelligence analysis where analysts have to deal with an increasingly fast-paced exchange, curation, interpretation, and dissemination of information, especially through the integration of the physical and digital worlds, e.g., in hybrid societies.

In conclusion, current methodologies in Intelligence Analysis, despite helping analysts reduce their cognitive load and despite being formally sound, are highly susceptible to deception and misrepresentation. ACH, the ACH application to Bayesian nets, and even ACH-SL, fall short of addressing the pragmatic reasoning of adversaries in a formal manner. It is a bit ironic that ACH itself had been designed by Heuer to address deception, but ACH only does so by considering observable events and does not address the beliefs, desires, intentions, goals, plans or strategies of the agents involved in these events, never mind the interpretation of these attitudes in different contexts of events. In this thesis, I aim to provide a MAS approach that aims to fill these gaps and to work towards a framework that provides a more comprehensive understanding of deception. To do so, I will continue by discussing deception from the perspective of Social and Communicational Psychology, in the following sections.

2.1.2 Limits of Passive Cue-based Deception Theories

Before we go into the theories of human deception that are relevant to this thesis's aim, I first want to argue against an alternative course of action. This alternative course of action would have implied delving into the so called *cue-based* psychology of deception detection. Cue-based approaches to understanding or detecting deception treat passively observable information that represents what lies on the surface of a more socio-complex cognitive process.

This simplistic view of deception, especially the over-reliance on non-verbal cues, had been extremely popular for some time due to the work of Paul Ekman [75]. Even Ekman himself, along with O'Sullivan [77], argues that clues that lead to deception are not absolute and that it is perhaps delusional to think that a "*fail-safe*" set of behavioural cues that are able to indicate deception exists in real life. There exist, Ekman and O'Sullivan say, some behaviours that might be considered cues for deception, but these need to be analysed by taking into account the individual that expresses these behaviours as well as the social context in which these behaviours are expressed. Ekman, later in [74], had aimed to integrate verbal behaviour into his view of deception.

These cues that I am referring to can be either verbal, or non-verbal. Verbal cues represent what a person says (linguistic behaviour) and non-verbal cues represent non-linguistic behaviour of a person. One example of non-verbal cues are micro-expressions. There is relatively strong consensus in human deception research that cue-based approaches to deception detection are very limited [67, 66, 116, 104]. That is because this type of approach solely considers behaviour that is expressed by human agents, in order to establish whether the human is being deceptive or not. It does not, however, take into account everything else, e.g., the knowledge involved as well as the context of the situations. This makes cue-based deception research highly susceptible to cognitive biases [v].

One of these biases is called the *truth bias*. Does a receiver believe that a sender is truth teller? If yes, then the detector has a strong truth bias. Or, does he believe the sender is a liar? In this case, the receiver is considered to have a strong non-truth

[v] Remember from the previous section that Heuer and intelligence analysts are not very fond of cognitive biases for very good reasons.

bias. What if the receiver is not sure about whether the sender is telling the truth or not? If this is the case, then we call the receiver suspicious.

Another interesting concept is the *confirmation bias*, to which cue-based techniques are also susceptible. The confirmation bias is present when a receiver interprets incoming data from the sender in such a way as to confirm its own bias. Let us say the receiver believes that the sender is a truth teller. If the sender is indeed a truth teller, then the receiver does not make any error. However, if the sender is not telling the truth (a.k.a. deceiving the receiver), then the sender commits a fallacy. The fallacy is due to the fact that the sender ignores all the information that could falsify its belief that the sender is a truth teller, because the receiver is only processing information that strengthens its beliefs.

The confirmation bias can also determine receivers to commit what is known as the *Othello Error* [30]. The Shakesperean tale goes like this: Othello believed that Desdemona, his wife, was cheating on him with another man. When Desdemona denied that she was cheating on him, Othello did not believe her and also exhibited suspicion. Because Othello did not believe her, Desdemona started crying and exhibiting behaviour that correlates with a cheating wife. That was because Desdemona was desperate. She was desperate because she believed that no matter what counter-arguments and evidence she would offer Othello, he would still only take into consideration the information that would confirm his hypothesis of her cheating on him. Desdemona was killed by Othello because she behaved like a person who is desperate. A person who is desperate exhibits some of the behaviour a guilty person would exhibit. Othello's fallacy was that he erroneously took into consideration only the behaviour a guilty person would exhibit, without taking into consideration all the other cues that might have falsified his beliefs, e.g., the cues and information that could have indicated despair. Thus, we call this the Othello Error.

In the case of deception, the detector might be right that a sender is attempting deception. However, the detector is right for all the wrong reasons. Whereas the sender will never be able to convince the receiver that it (the sender) is actually a truth teller, no matter what strategies it uses and how strong its arguments and evidence are.

Psychological research has identified and described several effects of lying and deception, as well as several issues humans face when attempt to deceive or detect deception. Notably, the work of DePaulo and Bond have shed some light on how the over-reliance on specific cues is problematic and that such cues do not improve deception detection, e.g., they do not provide diagnosticity [66, 67, 29].

On a slightly different note, according to Hartwig [117, 116], cues to deception do exist and verbal cues are more diagnostic than non-verbal ones, but the trick is to distinguish between the ones that lead to perceived deception and the ones that she calls *valid* cues, which lead to actual deception detection. The problem would then be to elicit valid cues in order to really improve deception detection. Apparently, the answer, according to a review on deception research by Vrij, Hartwig and Granhag [288], could be either (i) using specific interview protocols (for example, to increase cognitive load of potential deceivers such that they are induced in making mistakes, e.g., reverse-order interviews) and analyzing speech content [vi], or (ii) the consideration of contextual factors such as the context of the conversation and familiarity with the topic that is being discussed. The same authors mention that non-verbal behaviour should only be pursued when the alternative of assessing verbal behaviour does not exist.

In conclusion, I might have taken an approach that relied on cue-based theory to model machine deception, but the over-reliance on specific behaviours, verbal or non-verbal, seems problematic and reductive to truly provide a meaningful and

[vi] However, according to Levine, approach (i) does not fare better than cue-based passive approaches [153]

accurate understanding of deception. Therefore, I have decided to take a different path. This decision was reinforced by the fact that ACH and its later versions have been designed to overcome the issues which cue-based deception research brings up. Therefore, I have gone with theories from Communication and Psychology that address deception in a different manner, and that do not solely rely on verbal and non-verbal cues to understand whether, and describe how, an agent is attempting deception or not. In the following sections, I will describe the three main theories that aim to do this.

2.1.3 Truth-Default Theory

The first theory of deception that I describe is *Truth-Default Theory*, proposed by Timothy Levine. TDT [153] aims to address general rules of deception and argues that humans are generally truth biased, both as senders and as receivers of communicative messages, thus they are in a *truth-default* mode most of the times. However, it is also important to note that TDT does not imply that humans are always in a truth-default mental state, as various triggers can take individuals out of this state.

> "The truth-default involves a passive presumption of honesty due to a failure to actively consider the possibility of deceit at all or as a fallback cognitive state after a failure to obtain sufficient affirmative evidence for deception." [153, loc. 2151]

TDT merely argues that humans mostly presume honesty in a passive manner. This argument is backed by Timothy Levine, the proponent of TDT, with numerous controlled empirical studies on how humans judge the truthfulness of others. These studies have led to the proposal of concepts such as the *truth-lie base-rate*, which informs us that if an individual is not actively confronted with evidence of deception, then that individual will consider communication to be truthful; the *veracity* *effect*, which specifies that people are more accurate when judging truthful messages compared to dishonest messages; the *probing effect*, which specifies that individuals have a tendency to answer interrogation questions in a minimal, but truthful manner instead of answering accurately; and the *sender's honest demeanor*, which strongly influences deception detection variation, where by demeanor we understand a set of behavioural characteristics that define a sender's character or personality.

TDT is one of the most comprehensive theories on deceptive communication and because it is also a modular theory with standalone sub-theories, models and hypotheses, it would deserve its own set of PhD theses to be fully reviewed, summarised and analysed. TDT teaches us about the following aspects of deception [153]: how there are only a few prolific liars [vii] [153, loc. 2188]; that deception detection accuracy is slightly above 50% [viii] and there are only *a few transparent liars* out there [153, loc. 2189]; that most lies are detected after the fact, from evidence or confessions, and only very few lies are detected through the passive observation of cues [153, loc. 2189]; that understanding the communicative content inside the communicative context improves deception detection [153, loc. 2189]; that some types of information in communicative interactions are more useful than others, that every type of information has its own *diagnostic utility* for deception detection [153, loc. 2213]; that fact-checking and checking of logical consistency of what is being communicated can be successfully used in deception detection, but, in general, fact-checking proves to be more useful; that effective questioning, as opposed to ill-conceived questioning, can be used to extract diagnostically useful information from the sender [153, loc. 2213]; that knowing how to prompt communicative information that has a high diagnostic utility should be considered above the ability of interpreting behavioural cues from a sender [153, loc. 2213].

[vii] Levine calls them "*outliars*", which I personally find to be an extremely funny *wortspiel* given the scientific context.

[viii] Levine considers that because some people might actually be very good at deception detection, then these individuals make deception detection slightly better than chance.

What mainly interests us in TDT, apart from the theory's modules and its general rule of truth-default, is TDT's perspective on the definition of deception and deceptive motives. *What is deception?* According to TDT [153, loc. 2149], "*Deception is intentionally, knowingly, or purposefully misleading another person.*", which is different from lying, which is simply called *bald-faced lie* by TDT and is described in the following way: "*A lie is a sub-type of deception that involves outright falsehood, which is consciously known to be false by the teller and is not signaled as false to the message recipient.*". TDT makes several claims about deceptive motives to answer the question *When and why do people lie?*. The claims are the following [153]:

1. People lie for a reason. Deception is purposive, and therefore, deception is not random [153, loc. 2214].

2. Deception is not the ultimate goal. Deception is a means to some other end. Therefore, deception is usually tactical [153, loc. 2222].

3. The motives behind truthful and deceptive communication are the same [153, loc. 2174].

4. When the truth is consistent with a person's goals, the person will almost always communicate honesty [153, loc. 2174].

5. Deception becomes probable when the truth makes honest communication difficult or inefficient [153, loc. 2174].

To summarise, TDT is unique in deception theory because it presumes (based on 55 empirical studies) that in most interactions humans do not check whether truthful information is being communicated. For TDT, deception is infrequent compared to truthful communication, therefore truth-bias should not be considered an erroneous way of thinking, as it has evolutionary basis given the frequency of deception.

However, truth-default is put aside when humans find contextual reasons to start questioning the truthfulness of communication, e.g., logical consistency of what is being communicated or by projecting motives on to other's mental states etc. In terms of deception detection, TDT argues that the contextualisation of communication as well as the persuasion of potential liars to confess are the best approaches, whereas relying on nonverbal cues hinders deception detection.

2.1.4 Interpersonal Deception Theory

The second theory of deception that I describe is *Interpersonal Deception Theory*, proposed by David Buller and Judee Burgoon. IDT focuses on the physiological and social parameters that influence deception in interpersonal communication [38]. The purpose of IDT is to integrate previous knowledge about deception with empirical evidence that resulted from a set of over two dozen experiments conducted by its proponents. In order to explain deception, IDT takes into account knowledge about trust, lies, psychologies of different agents, cognitive bias, truth bias, roles of agents in different contexts, social norms, semantics, encryption and decryption of communicative acts, social skills, and the performance of deception and detection of deception.

Central premise of IDT: *Deception and deception detection are not passive activities. Interaction is crucial to deception.*

IDT aims to be a holistic approach to understand deception and rejects experimental setup approaches. For IDT, understanding the interactive context in which deceptive communication happens between agents is crucial. The context in which deception happens is, according to IDT, influenced by several social factors of which some are: the *leakage*, which represents the nonverbal information exhibited by an agent and that contradicts the agent's intended message; the *cognitive load* of an agent regulates leakage (high cognitive load determines high leakage) and increases or decreases based on the number of behavioural tasks that need to be managed by an agent on a cognitive level; and the *communicative skill* of an agent which represents the ability of an agent to communicate socially and to manage and control its demeanor in social interactions (the higher the communicative skill of an agent, the lower the cognitive load, and therefore the lower the leakage exhibited by that agent).

IDT describes a deceptive interaction as an information exchange that takes place between a sender and a receiver. The sender plays the role of the deceiver, while the receiver plays the role of the detector. The sender has a more active role, because it is the sender that has the assumed intentionality to deceive. However, that does not mean the receiver is fully passive in this interaction. The receiver's role as detector is to find out if the sender is trying to deceive or not. This means that the receiver is able to use various strategies to analyse the incoming information from the sender and also, through its own behaviour, the receiver can make the sender reveal enough relevant information such that the receiver is able to adapt its detection strategy for the upcoming interaction.

According to IDT, the success of deceivers and detectors is influenced by the social factors that we mentioned at the beginning of this subsection. For example, if the communicative skill, the ability of an agent to socially interact with other agents using verbal and non-verbal behaviour, is increased, then the better is the agent's performance at deceiving or detecting deceit. Communicative skills are themselves influenced by an agent's personality and cognitive abilities.

Also, the communicative skill of an agent determines how well that agent can create its own representations about the world and about other agents. Consequently, the accuracy of an agent's representations determines how well the agent can adapt its strategies of deceiving or detecting. Usually, a well adapted strategy is entailed by a successful performance of the agent.

Another factor that influences the success of a deception or a detection is *leakage*, first described by Ekman and Friesen [75]. This is usually the overt non-verbal message that contradicts a verbal message. In humans, leakage takes the form of affect display. Affect display is the behavioural expression of internal emotional states. Some examples of affect display are: laughter, tears, facial expression, facial micro-expressions, hand gestures, body posture, tone etc. In IDT, affect displays can be treated as leakage due to the fact that agents are usually employing strategies to control verbal behaviour (what they say), hence increasing their cognitive load. Consequently, due to a heavy cognitive load, the agents lose the necessary resources to control their affect displays, and therefore leakage happens. The leakage is itself dependent on the communicative skill of an agent. It is also inversely proportional to the skill of the agent, and by extension to the success rate of the agent.

2.1.5 Information Manipulation Theory 2

The third theory of deception that I describe is *Information Manipulation Theory 2*, proposed by Steven McCornack. IMT2 [172] detaches itself from IDT by focusing on the perspective of how a sender agent uses its communicative capabilities for both deception and truth-telling. IMT2, thus, focuses on the information an agent manipulates in order to convey different meanings of the message in the mind of the target. In this respect, IMT2 dives into the reasoning processes that can be employed by the deceptive agent to choose the semantic content of a deceptive message. IMT2 also discusses what makes a reasoning process, that is used for the development and delivery of a message, more or less deceptive.

IMT2 was proposed to address a set of deficiencies of its predecessor, *Information Manipulation Theory* (IMT). One deficiency of IMT is that it is not actually a theory in itself because it had not been formally defined, thus it had not proposed any scientifically testable propositions. Another deficiency was that it did not address the production mechanisms responsible for information manipulation.

Central premise of IMT2:

> "Deceptive and truthful discourse both are output from a speech production system involving parallel-distributed-processing guided by efficiency, memory, and means-ends reasoning; and this production process involves a rapid-fire series of cognitive cycles (involving distinct modules united by a conscious workspace), and modification of incrementally-constructed discourse during the turn-at-talk in response to dynamic current-state/end-state discrepancies." [172, p. 362]

IMT2 challenges the idea of a top-down deception, where the deceiver decides a priori whether to lie or to tell the truth. According to McCornack, deception follows a problem solving process in which deceptive intent should be treated as an empirical issue (inside of the agent's mind) and should only arise as a solution in a dialogue. An important idea that follows from the main claim and that is further claimed by IMT2 is that the production of honest and deceptive discourse should be of similar difficulty regarding the *cognitive load* of the speech production mechanisms. This is mainly valid for BFLs (bald faced lies) [ix] and BFTs (bald faced truths).

One important concept of IMT2 is the relation between *Pars Pro Toto* and *Totum Ex Parte* taken from the Mannheim school's *Model of Speech Production* [121]. *Pars Pro Toto* is the main speech production mechanism employed by a deceiver agent, according to IMT2. When employing *Pars Pro Toto*, a deceiver agent formulates a message that consists of strategically chosen bits of information that make part of a possible total knowledge. In order to employ *Pars Pro Toto*, the deceiver needs to have a mental model of the detector agent's mind. That is, the deceiver needs to know what the detector might conclude from the information the deceiver is giving

[ix] According to McCornack, BFLs are not a really popular message of choice when deception is attempted.

him. Thus, the deceiver has to have a representation or knowledge regarding *Totum Ex Parte*. *Totum Ex Parte* is the process that the detector agent engages in when it encounters a message that was formulated through the *Pars Pro Toto* process. The detector uses inductive reasoning to reach a conclusion by creating a so called "complete" knowledge or representation out of his already available assumptions that he believes to be true (or not) and the bits of information from the *Pars Pro Toto* message [154].

To give it a propositional logic twist, let us think of the following example. We have two agents A (deceiver) and B (detector). Agent A knows that $\neg Q$ and agent B does not have access to this knowledge unless B asks A for this knowledge and both of them know this is the case. Also, both A and B have some common knowledge, let's say $P \rightarrow Q$ [x]. A's goal is to make B believe that Q. If A would send B a complete message, then A would have to say that: "P, $P \rightarrow Q \models Q$". However, A chooses to apply the Pars Pro Toto process given that A knows that B knows that $P \rightarrow Q$. Thus A utters the following message: "P". After this event takes places, agent B engages in its own reasoning process, concluding that Q is the case given the knowledge B has from A that p and B's previous knowledge that $P \rightarrow Q$. In other words, agent A is letting B "fill in" the rest of the information. Agent B now believes that it has reached the conclusion by itself.

Interestingly, IMT2 has been inspired by research in linguistics, AI, cognitive neuroscience, and, obviously, speech production. In conclusion, IMT2 offers a very good outline of what type of reasoning mechanisms are involved in deceptive communication between human agents. In this thesis, I will draw onto IMT2 to engineer artificial reasoning mechanisms for deceptive artificial agents which I will derive from the components of the human reasoning mechanisms described by IMT2.

[x] This is just an illustrative example. Common knowledge is not necessary, as it is sufficient just for A to know that B knows $P \rightarrow Q$, whithout B knowing that A also knows $P \rightarrow Q$.

2.1.6 Summary

In this section I have given an overview of the relevant body of knowledge on human deception. I have discussed methodologies used in Intelligence Analysis to explain complex cases, some of which are based on whether deception has occurred or not, and also the limitations of these approaches when it comes to deception. After that, I have explained the limitations of cue-based approaches to understand deception in order to emphasise my decision not to pursue them in the modelling of deceptive machines. Finally, I have described three main theories, TDT, IDT and IMT2, that serve as conceptual foundations of this thesis[xi].

2.2 Data-Oriented Models of "Deception"

In this section I address the current threats posed by deceptive AI tools. I refer to current research on how AI generators of fake news and deep fakes, hyper-partisan ecosystems, automated crowdturfing attacks and catfishing behaviour contribute to online deceptive behaviour. At the end of the section, I discuss why some of these issues are difficult to address given current methodologies, and also hint towards how the increase in autonomy of deceptive AI will lead to future threats.

2.2.1 Automated Fake News Generation

In February 2019, OpenAI released the original version of the GPT-2 model, an extension of the original GPT-1 (Generative Pre-trained Transformer) which mainly processes and generates text [212]. The large framework of GPT-2 is actually the

[xi] Apart from the research on human deception that I have described in this section, there is a set of studies worth mentioning to the reader who wishes to understand more of the specific complexities of human deception. Notably, the Special Issue of the Topics in Cognitive Science Journal on *Lying in Logic, Language and Cognition*, edited by Hans van Ditmarsch, Petra Hendriks, and Rineke Verbrugge, has both clarified and partially confirmed some of my intuitions about the fact that complex cognitive processes are being executed inside the minds of humans when they deal with dishonesty [278]. The theoretical [81, 73] and empirical [147, 36, 91, 96, 20] work described in this special issue delves into the complexities and underpinnings of the logical, linguistic, cognitive, and epistemological aspects of lying and deception in social cognition.

framework of GPT-1, but one which turns supervised NLP (natural language processing) tasks to unsupervised ones for training during a *Fine Tuning* phase [7]. Essentially, GPT-2 uses a large amount of unsupervised training data to expand the language corpus and obtain highly complete text generations. Although the user only inputs one sentence or several words, GPT-2 will return large articles. The problem with GPT-2 is that while it sometimes outputs logically coherent text, other times it outputs text that is incoherent or nonsensical.

When users are presented this type of AI-generated text, they might initially find it challenging to verify its semantic authenticity [283]. However, after verifying keywords, users can easily discern such AI-generated fake news.

These limits can be tracked back to the limitation of knowledge about a specific domain [58, 188]. Strongly represented categories in the training data (Brexit, movies, popular personalities and so forth) are best suited for the generation of new deceptive text, as well as programs for capturing common sense knowledge such as OpenCyc [65] and Open Mind Common Sense [257]. At the other end of the spectrum are deeply scientific topics and much less known topics that hinder even GPT-2's performance. Another limitation is these type of models' capability to deal with changes from English to some other language for predicting text.

2.2.2 DeepFakes

Given our growing reliance on entertainment audios and videos, the occurrence of *DeepFakes* makes us realize the potential risks regarding personal content and authentication [173]. Thus, another threat is generating fake news by transposing faces from celebrities [52], as well as the broadcasting of altered images, voices and videos of politicians [27]. For deepfakes, as [262] demonstrates, a clear solution is still missing to detect them. Therefore, it is likely to be spread widely through social media and have a very negative impact. Systems like Facebook and Twitter are still struggling with deepfakes because they do not seem to keep up with technological advances.

Unlike text, media items such as images, gifs, audios and videos have high entropy in terms of the information they embed. Detecting deepfakes requires the design of scalable and complex algorithms. Efforts have been made to detect deepfakes. Until,now, there are two approaches: temporal features across frames and visual artifacts within frame. One example of the former one is *FaceForensics++* [224], which manipulates the video content by three methods: Face2Face [268], FaceSwap [138] and NeuralTextures [267]. They created a dataset of over 1.8 million randomly compressed and randomly-sized images[xii]. Then, they used this dataset as a benchmark to test all state-of-the-art deepfake detection models. These classifiers can be either shallow [301] or deep classifier detection [155] based on the extracted number of specific features. Among all of these models, the best performance was achieved by Xception [49] (by Google). Authors claim that small knowledge in a specific field can help improve the accuracy of the detection in certain cases. For an example, using a cropped or enlarged clip of a face from a video and testing this small part rather than the whole video can enhance accuracy.

2.2.3 Hyper-partisan News

"*Hyper-partisan*" news is a term that is believed to originate from a recent article in the New York Times Magazine[xiii] [120] and refers to news reporting that departs from the traditional notion of journalistic balance, and presents a biased picture of one side of a political debate. These hyper-partisan news often spread false and biased information to attract more users. In 2017, *Buzzfeed news* came up with a list of hyper-partisan news websites which were highly active during the 2016

[xii] Benchmark dataset available at- https://github.com/ondyari/FaceForensics

[xiii] https://nyti.ms/2k82R8I

US elections [255]. The list contains 667 websites[xiv] with labels of right-wing and left-wing based on the description of website in online spaces such as homepage, Facebook page and so forth. The list also contains a label for around 77 websites where websites are being run from Macedonia. Interestingly, out of these hyper-partisan websites, around 450 were associated with corresponding Facebook pages that help increase the visibility and audience of these websites.

In [26] the authors study characteristics of these websites to understand their ecosystem. They report that most hyper-partisan websites are created before elections and then disappear after elections. Almost one-third of these were newly registered websites during 2016, the election year in the US. Interestingly, out of these newly registered websites 81% come from right-wing and only 19% left-wing partisanship. Later by the end of 2018, almost 60% of them disappeared. The study also performs an analysis of demographics based on the Alexa.com [5] internet users traffic data on these websites. They find that young (25–34 years) and old (over 65 years) hyper-partisan news consumers, and those having a higher educational background (college degree or post-graduates) tend to be left-leaning. One the other hand, right-leaning news websites tend to attract middle-aged and old users (45–64 years old), and significantly high numbers of users who have at most a college degree. These differences in consumer demographics allow these websites to perform differential tracking by increasing or decreasing tracking intensity, thus making advertisers pay higher (or lower) prices for showing ads [2].

In [2], the authors provide a methodology to understand the underlying ecosystem of differential tracking based on user demographics. The authors test the effect of perceived user persona on hyper-partisan websites. These websites perform more intense tracking (using *cookies* etc.) than the general web. The study shows that (i) right-wing websites tend to track with up to 25% more than the left-wing and deliver

[xiv] Methodology and websites link are available here- https://github.com/BuzzFeedNews/2017-08-partisan-sites-and-facebook-pages

up to 5 times costlier ads; and that (ii) the most popular websites on both sides are the ones that have the highest levels of tracking. For example, websites which rank between 1-10k on Alexa.com [5] have the highest median of cookies tracking, right–227 and left–131. Most of the third-party trackers exist on multiple websites of a same partisan-leaning [2]. Very few domains exist on both sides of partisan leaning which creates a polarised tracking ecosystem for users.

Other studies on alternative news look at online events and the participation of their corresponding users [264, 304], most of which are driven on social media platforms. In [264], the authors analyse alternate news on Twitter around *mass shooting events*. Their findings show that alternate news media try to propagate non-standard narratives of events while mainstream media continuously denies these narratives. Alternate news also get more clicks and categorise the primary orientation in to four type– namely traditional news, clickbait news, primarily conspiracy theorists and Political Agenda. Alternate news link sharing to other platforms also acts as a catalyst in reaching a wider audience [304]. Fringe user communities from alt-right make more efforts to disseminate news that is mostly seen first on Twitter and Reddit[xv] and later on 4chan[xvi] after a short period of time.

2.2.4 Automated Crowdturfing Attacks

Automated Crowdturfing Attacks (ACAs) are used to spread malicious and bogus news online, and have been actively applied in social media [299], in Online Review Systems [302], and in plenty of other online communities. For instance, increasing customers would reference the online reviews of a restaurant or hotel before their booking decisions to judge the quality based on others' experience. Therefore, ACAs would inject framed information (or biased opinions) inside comments to mislead customers' choices, which then would lead to the "satisfactory" result of attackers

[xv] https://www.reddit.com/, online discussion forum

[xvi] http://www.4chan.org/, an image based bulletin board

(in terms of monetary purposes). Since most of the information that is infused by the attacker in comments does not correspond to real events, it should be regarded as an online scam in our daily browsing experience. Machine Learning techniques can also be used to generate complex reviews, some almost flawless fake reviews [302].

To characterise how this false information is spreading in the online community, [140] presents a quick review on the reasons behind ACAs, the rationale, the effects and algorithms to provide a common understanding and framework. Some other approaches are the fraudulent behaviour analysis of Amazon fake review authors [134] and the TwoFace system [135]. In terms of the propagation speed, [286] shows that false news and fake reviews seem to be more viral than the truth. In terms of audience that can be reached by fake news, the authors present in [250] a spreading pattern based on the popularity distribution, which shows that *low-credibility content* can reach massive exposure. They mention that the difficulty of detecting such content lies with the identification of the truth on crowded information, i.e., the way to convincingly establish the ground truth.

Another fertile ground for ACAs are shopping websites like Amazon, where at least two kinds of attacks towards online customers have been identified: promotion and restoration attacks [133]. That is why for customer protection, Amazon proposes TOmCAT [133], a detection framework based on the behavioural patterns of attackers. Results show that *95 out of 100 top-ranked products in Amazon* are indeed attacked and that shopping decisions are made by the attackers instead of the customers.

2.2.5 Catfishing

In terms of social media platforms (SMPs), [276] illustrates that *age, screen-name and located time zone* of a Twitter user are the most promising characteristics to be used in deception. This type of deception comes under the form of a *catfishing attack*. Catfishing attacks are the creation of fake online user profiles used for fraud. Despite the fact that catfishing-related criminals have started being fined for owning deceptive Twitter accounts [182], the number of catfishing attacks keeps increasing. Attackers even utilize the celebrity effect on SMPs to raise more money and reputation [220], against which younger fans of celebrities are especially vulnerable.

Online dating systems are another type of platform where catfishing happens frequently [37]. [xvii] Unlike the target groups on social media, most catfishing attacks on online dating systems target females. False profiles with attractive photographs can rapidly create a welcome "person", causing victims to be seriously affected, e.g., money loss.

On the other hand, some studies use AI agents to alarm users of the danger of being "*catfished*". For example, [158] claims to have created software that detects over 90% of online fake profiles, and reveals that the bias of attackers in favor of lying about *gender* to be 25% and lying about about *age* to be 38%, when making decisions of whom their next online targets are. This kind of scam takes advantage of users' trust and then collects sensitive information or gifts through communications.

Other studies regard catfishing as a "cyberbullying"-aimed activity, which implies the higher probability of exposure to the mental health problems [146, 119]. Since the technical, financial, and organised requirement of catfishing is not high, most catfishing scams could be operated by individuals, as opposed to groups [219]. The study of [71] confirms that only 16% to 32% of participants would always be honest when filling forms for online profiles.

Catfishing detection tools have also been developed. To avoid catfishing, especially if the other party communicates vague or suspicious information, the authors

[xvii] [113] also regards this type of attack as one of the ways to achieve the *autonomy* of artificial intelligence-mediated communication. However, I believe this argument needs to be considered with caution.

of [276] optimise an algorithm (IDDM) for online catfishing detection, and updates the authenticity measurement for setting up accounts. A user friendly alternative to this type of tool for algorithm optimisation is the Google Reverse Image Search that can help users determine whether the other party's photograph refers another real person [162].

2.2.6 Ekman's Artificial Legacy

The over-reliance on non-verbal behaviour of psychological theories of deception has also been adopted by AI researchers in the development of software that detects deception in humans. This type of software mainly uses Machine Learning (ML) techniques for psychological profiling. Such systems are used in the real world, mostly in border-control, where they are used to calculate the probability of deceit in interviews by analysing non-verbal micro-gestures (microexpressions). Two of these systems have even been funded by the US's DoD and by the EU.

One of these systems is called Silent Talker [225, 226], that is also used for iBorderControl's Automated Deception Detection System. The other one is a more advanced model, called Automated Virtual Agent for Truth Assessment in Real Time (AVATAR), which is based on a previous idea of its author, Jay Nunamaker, together with Judee Burgoon (one of IDT's co-authors), and presented in [193]. AVATAR is a more advanced system since it is based on the idea of agent-to-agent interaction and assumes that the virtual interviewer's demeanor might elicit relevant non-behavioural cues form the interviewee. However, AVATAR's deception detection system still is over-reliant on the passive analysis of non-behavioural cues, as the virtual agent cannot adapt to social interactions, but merely generates interview questions following a pre-defined interview script. This makes AVATAR susceptible to the same faults as the data-oriented ML models of deception detection I have described before.

2.2.7 Summary

Given the high popularity of online deception as a research topic, several companies even launched competitions to challenge deception detection. Some of the highest scores for detection are: automated crowdturfing feature detection 96.9% for DeepFakes and 96.3% for FaceSwap detection [229], and 84% for TOmCAT[133].

However, most studies are conducted under particular scenarios and may be not applicable to more general scenarios [270]. Meanwhile, they are also affected by the judgement of humans [47] on whether the content on platforms is false. Moreover, most spoofing based on neural network AI is a black box, which makes it difficult to obtain internal principles and find corresponding solutions.

Another counter-argument for data-oriented models is the over-reliance on modelling cue-based approaches to deception. This makes these models highly susceptible to cognitive and statistical bias. Such approaches can easily fail to distinguish between behaviours that are cues for perceived deception and ones that truly indicate deception. Even more problematic is the fact that the behaviour of deceivers compared to non-deceivers is highly similar (at least for human deception). This applies to models that aim to detect both online deception and human deception, such as AVATAR and Silent Talker.

Finally, and most importantly, what happens inside the mind of agents, human or artificial, when they attempt deception is completely ignored by data-driven models [237]. Data-oriented models or frameworks in AI have proved to be suitable for describing observable deceptive behaviour, especially when humans are exhibiting this type of behaviour. However, they fail to grasp important components of deception, especially the components responsible for the intentions, reasoning and decision-making of deceptive agents in different contexts. Therefore, if we want to use AI techniques to increase our understanding of deception, then we should stay away from data-oriented models.

Let us take the following example to illustrate this issue of their unsuitability: Alice receives an e-mail from Bob. Bob's e-mail contains a list of best practices for preventing infection by a certain virus, e.g., COVID-19, Bob claims in the same e-mail that this list had been forwarded to him by Carl who, Bob also claims in the same e-mail, is a medic. The list, turns out to be misleading and not entirely accurate, e.g., some of the practices enumerated can either have no impact in preventing COVID-19 or could even cause other negative symptoms. Natural questions about this list would be *Who created the list? or What created the list? Was the list generated manually or using an AI tool?, How do we prove who or what is the creator of the list?, What were the reasons behind the creation of the list? Why the specific topic? Who was the target of the list? Has the target of the list been reached? Has the message achieved its intended deceptive outcome? Why has been the list been shared with Alice and Bob, why not someone else? Is Carl real, or is Bob lying and manufacturing Carl's persona to mislead Alice into thinking that he (Bob) had no deceptive intentions? etc.* Such questions cannot be addressed unless an approach other than the data driven one is taken to address online deception. This approach has to go beyond what is currently done with the analysis of observable behaviour using data-oriented AI tools or by observable behaviour generated by data-oriented AI tools.

2.3 Deception and Autonomous Systems

The threats I have explored in the previous section have been created through the use of AI as a tool in the hands of humans. Therefore, to address machine deception, I must turn towards understanding why and how humans deceive as well as why and how they use specific AI tools for deception. Most cases of machine deception involve the *human-in-the-loop* who decides who to target and how. Humans also think about what type of information to deliver in order to deceive and how to acquire or generate the necessary information. How can we study and understand the way deceptive agents work, think and act? And if we do understand them, then how do we detect, prevent or mitigate their deceptive behaviour?

Also, online deception has not even been fully automated yet. What if artificial autonomous agents developed their own reasons and methods to deceive? A particular risk is posed by the enhanced potential ability of artificial agents to out-reason our cognitive capabilities [237]. How would we address this issue? Also, is there a unified approach to address both human and artificial agent deception?

AI researchers have foreseen some of the threats coming from the possibility of deception in cyber and virtual societies and have even proposed the idea that MAS could be used to build more comprehensive theories of deception [45, 237].

Multi-Agent Systems (MAS) research aims to build models that integrate the social, behavioural and cognitive components of trust and deception. In [45], as well as in [44], Castelfranchi and Tan argue that apart from the role of trust in virtual societies, it is also crucial to study the role of deception. The authors emphasise that this is most important in hybrid human-agent interactions, raising the issue that in order for agents (human or artificial) to reason about the trustworthiness of their counterparts in different contexts, a theory of both trust and deception is necessary. They indicate multiple levels of trust, such as trust in one's agent and mediating agents, trust in the MAS environment and infrastructure, trust in potential partner (collaborator) agents, and, finally, trust in authorities. Castelfranchi's and Tan's perspective, reinforced by the one presented in [80], on trust in cyber-societies, is becoming more relevant given the increasing number and complexity of hybrid interactions between humans and artificial agents.

In this section I will describe some of both seminal and state-of-the-art research regarding deceptive AI, breaking it down into six different areas which study different components of agent deception: 1) Deception in Agent Societies; 2) Logical Aspects of Agent Deception; 3) Strategies for Agent Deception; 4) Reasoning in Agent Deception; 5) Engineering of Deceptive Agents; 6) Embodied Deceptive Agents.

2.3.1 Deception in Agent Societies

Research that falls in this area aims to motivate the AI community to build a theory of trust and deception in virtual societies. The main body of research mainly focuses on human-computer interaction in MAS and tries to answer questions that arise from interactions between human and artificial agents, such as *Will artificial agents deceive?*, *Why would they deceive?*, *Why is it important for agents to be able to reason about deception?*, *Why should we model deceptive agents? What are the types of deception in virtual communities? What is the role of deception in agent-agent interaction?*

Recent findings back up this perspective by showing how humans and artificial agents influence each other's attitudes and behaviour. For example, the authors in [174] focus on how the negotiation strategies of artificial agents determine humans to endorse artificial deception. Apparently and counter-intuitively, humans also tend to be more cooperative with deceptive machines than with friendly ones, according to [131]. Another type of study, presented in [70], looks at the ways in which deceptive language can be generated and detected in virtual systems. In [193] the Embodied Conversational Agent-Based Kiosk for Automated Interviewing (ECA) uses virtual interviewer agent avatars to evoke non-behavioural responses in the interviewees based on the avatar's demeanor in order to detect deception.

2.3.2 Logical Aspects of Agent Deception

This area aims to define a logical taxonomy of deceptive behaviour, as well as to represent and model deception using logical formalisations. Some relevant questions are: *What is the the difference between lying and deceiving?*, *How do we represent* *forms of deception such as bullshit, pandering, paltering etc.?*

Works in the areas of Logic and Philosophy have strongly influenced the current state of the art approaches in deceptive AI. Most known are the works in knowledge representation, which have laid the foundations of how to formalise deceptive communication. These appear in [234], in [232], and in [235], where the authors present the logical distinctions between different types of dishonest linguistic behaviour; in [277], where the author introduces a dynamic logic of lying that considers lies about factual propositions as well as lies about the beliefs of others; in [273] where the author uses the medieval concept of *dubitatio* to study deceptive agents; and in [132], where the author proposes a formal model of self-deception that is consistent given certain epistemic conditions imposed on the agent that engages in self-deception.

2.3.3 Strategies for Agent Deception

This area studies deceptive strategies in MAS. This area is strongly influenced by Economics and Cybersecurity, from which methods such as risk and threat modelling have been adopted and adapted to address specific scenarios. Some relevant questions are: *What are the deceptive strategies and counter-strategies in different contexts?*, *How to reduce and mitigate deceptive attacks?*, *How do we design a system that either reduces or incentivises agent deception?* etc.

Some of the work in this area explores issues such as using heuristics to cause target agents to execute a plan that will achieve the deceiver's desired goal [50]; finding deceptive strategies using path-planning [163]; using a MAS system on a Bayesian network test-bed to distinguish between truthful and deceptive agents based on the correlation of the agents' beliefs [236]; modelling agent based deceptive interactions on social networks [16]; modelling deceptive interactions to counter reconnaissance-based cyber-attacks [246]; studying the effects of deception in repeated games between learning agents [189]; applying IDT in multi-agent reinforcement learning to reduce the effects of social engineering attacks [300]; formalising cyber-deception games between multiple agents using hyper-game theory methods [82]. Additionally, given the increasing interest in explainability for AI, the authors discuss in [298] how agents can strategically provide users with deceptive and rebellious explanations.

2.3.4 Reasoning in Agent Deception

This area looks at formal and informal reasoning mechanisms responsible for deception, and is strongly influenced by Informal Logic and Argumentation as well as by subareas of Cognitive Science. Some relevant questions are: *What are the cognitive components involved in deceptive reasoning?*, *What types of reasoning mechanisms are involved in deceptive interactions?*, *What type of cognitive architectures can be used to represent deceptive reasoning?*, *What type of knowledge is necessary or not for an agent to deceive?* etc.

Work in this area of deceptive AI has explored issues such as using and detecting deception in argument debate games [231] that are formalised using abstract argumentation [72]; using abductive reasoning for deception [230]; using argument mining for detecting deceptive reviews online [55]; looking into what type of arguments can be used by a machine to deceive [54]; using the STRIPS approach to planning to flout Grice's Maxims [84]; modelling agents that use mindreading for deception [129].

2.3.5 Embodied Deceptive Agents

This area studies the performance of deceptive robot agents in the physical world. Some relevant questions are: *How can robots deceive humans?*, *What strategies can a robot employ for deception or deception detection?*, *What kind of robots are more deceptive?* etc. This area is becoming increasingly relevant given the advancement of Internet-of-Things and the interconnectedness of physical agents, such as self-driving cars and robot assistants.

Research in this area looks at issues such as: how deception about agency influences the way children treat robots on a social level and when is it acceptable to deceive children about the nature of the robots they interact with [293]; how robots can decide to perform physical actions such that they manipulate the beliefs of humans who observe them [105]; how to provide robots with the capacity to deceive [289]; using biologically inspired behaviour for robot deception [251], such as squirrel behaviour. In [252], the authors even provide a taxonomy of robot deception and present what are the benefits of deception for human-robot interaction.

2.3.6 Ethics of Machine Deception

This area looks into the morality of deceptive behaviour. Relevant questions are: *Is machine deception ethical?*, *Should machines be allowed to deceive?*, *How should machines deceive in order to be ethical?*, *If machines are allowed to deceive, then when should they be allowed to deceive?*

In this line of work, it has been discussed in [136] that some apps are designed to cognitively exploit users using coercion and deception and that a MAS approach is needed to study the ethics of such technologies in hybrid societies. Another topic is ethical deception, that is discussed in, [258], [130] and [46]. While the authors in [258] develop a model of contradiction to focus on deception in dialogues, the authors in [130] focus on the necessity of a Theory-of-Mind for reasoning about ulterior motives and argue that this type of meta-reasoning is crucial for a machines to be able to distinguish between ethical and unethical deception. Finally, the authors in [46] indicate that humans considered machine deception admissible for the greater good under certain conditions in the context of teamwork.

2.3.7 Related Approaches

If this thesis were to be categorised in this section, then I would have to define the category of *Engineering Deceptive Agents*. This area would look into the design, modelling, and engineering methods that can be used to create deceptive or deception-detective autonomous agents. The main difference between the area of engineering of deceptive agents and the areas covering strategies and reasoning of agents is that the engineering approach aims to integrate these agents into more complex systems. This area would not just look at how agents reason about deception, but what causes them to achieve deception or not in a given system. Some relevant questions are: *What are the best approaches to model deceptive agents or deceptive interactions without sacrificing expressivity?*, *How should these models or systems be designed and implemented?*, *How do we evaluate these systems?*, *What can we use these systems for?*, and *How do we integrate different components of deceptive interactions in order to engineer complex reasoning agents?* etc.

The works on machine deception closest to the approach in this thesis are [54], [143], [128], [129], and [217]. In [54], 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, the approach in this thesis consists in the design, implementation and evaluation of a MAS that is based on solid theories of deceptive communication. Therefore, the approach in this thesis is conceptually and methodologically different from [54] and also aims to offer a method to analyse deceptive machines, independently of user studies.

Similarly to [54], the author in [143] builds a cognitive model of deception which is based on human-computer interaction. The model in [143] 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. The limitation of this model is that it is just a protocol which is followed by the human programmer in the development of the strategies that the automated agent uses.

In [129], the authors develop a framework called FIDE, Framework for Identifying Deceptive Entities. FIDE is used for mental state attribution during dialogues in which deception might be the case. The authors also emphasize the importance of ulterior motive, in contrast to works that focus on the intent to deceive. While this thesis also focuses on ulterior goals and deceptive intent, it aims to focus on how to engineer deceptive interactions between agents in MAS instead of designing a framework that a single agent might use to for mindreading deception in dialogues.

Given that deception is a form of belief manipulation, I mention [128], where the authors describe and implement a model of belief manipulation using propositional public announcements. Their mechanism is similar to the one in this thesis 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, this thesis focuses on the interactions between two agents where one agent is the target of the other's attempt at belief manipulation. In these interactions I model not only (i) nested beliefs, but as a result of ToM modelling, (ii) agents that 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.

With respect to the the storytelling ideas in Chapter 8, the closest and most compatible approach would be the one in that is very briefly described in [217], where the authors present a model of an agent architecture that could be used by agents to engage in deceptive storytelling during dialogues. The agent in [217] uses a knowledge query engine to alter the original history of its actions that it keeps in its knowledge base. The main distinctions between this thesis's approach and the one in [217] is that in the one in [217] a) stories are represented as chains of events based on episodic memory theory, and they are not represented by hybrid arguments; b) the ToM module/layer is not used in mental simulation, but as a passive check of the consistency of the lies told using prior beliefs; c) deception is strictly represented through the alteration of the story; d) intentions and goals are not represented in the ToM layer; e) it is not based on any theory of deceptive communication; and f) it is a very brief description of an single virtual-agent architecture, and not a MAS approach.

2.3.8 Summary

In this section I have broken down in seven sub-areas the seminal and state-of-the-art research on deception in MAS and described the angles from which these sub-areas look at deception. The crucial aspect of the literature covered in this section is that it treats deception as being performed or communicated by an agent inside a system. This means exploring the internal mechanics of an agent's architecture (cognitive or physical), the logical properties of the communication, the high-level effects of the information in a system of multiple agents, and the behaviour (ethics) of the agent. Some of the works corresponding to these areas are also the ones that are closest to the approach to deception in this thesis.

2.4 Mechanisms for Complex Reasoning in Multi-Agent Systems

In this section I describe the mechanisms from the literature that I use further on in the thesis. These mechanisms are necessary for engineering deceptive interactions between agents that are able to reason about each other's mental attitudes, as well as for enabling them to communicate meaningfully.

2.4.1 Belief-Desire-Intention Agent Architectures

Belief-desire-intention (henceforth BDI) models of pragmatic reasoning are used in MAS for building agent architectures [215, 216]. The creation of BDI models was inspired by Bratman's approach to practical reasoning [35]. BDI agents are able to reason with and about their three mental attitudes. Bratman defines beliefs and desires as *pro-attitudes*, which are attitudes concerned with action, and intentions also as pro-attitudes, but which are also *conduct-controlling.* How does, then, an artificial agent then use these mental attitudes for practical reasoning? Let us look at a generic BDI architecture to see what role do these mental attitudes play.

BDI agent architecture [95]:

1. **Beliefs** represent an agent's knowledge of the world, or its knowledge of the state of the world. Beliefs are usually stored in an agent's belief base. A belief base is similar to a system's database, however, in MAS terms, this database contains exclusively a single agent's beliefs about the world.

2. **Desires** represent an agent's motives. They are also known as Goals, however the term "goal" does not capture the entire meaning and use of what Desires can represent in the context of BDI and MAS. Compared to goals, desires do not need to be consistent with each other. An agent actively pursues a desire.

3. **Intentions** represent a deliberative attitude of the agent, an attitude that an agent has committed to act upon. Thus, an Intention represents something an agent has decided to do, eg. *something the agent decided to act on.* In terms of AI Planning, they are the goals that an agent desires to achieve by executing a plan, however, given that we are discussing BDI architectures, these plans

can change in real time based on what the agent decides to do, e.g., what new intentions the agent has in light of new beliefs. This use of plans in BDI architectures is called *reactive planning*.

4. **Events** represent the internal or external triggers which update beliefs, trigger plans or update goals inside an agent's mind. They are also known as Percepts when triggered by an agent's sensory system which is coupled to an environment.

The notion of a BDI agent architecture has led to the further development of the Procedural Reasoning System (PRS) framework. PRS has been used in the development of reasoning systems for rational agents. In this case, a rational agent is an agent that uses specific knowledge defined in the reasoning system's knowledge area component to execute actions. As long as the agent follows the procedures specified in the PRS, it is rational because it uses its beliefs about the states of the world along with its knowledge about how to proceed given these beliefs in order to select the best procedures (the intentions) such that it achieves its desire or goal.

One of the initial criticisms of BDI in MAS architectures is that they do not follow distributed AI principles. That is, agents with BDI architectures are mainly used for epistemic plan execution or deliberation and cannot perform distributed reasoning [xviii] By distributed reasoning we understand a reasoning process which is executed sequentially or in parallel between two or more agents, where one agent only executes part of this process. To address this issue, among other methods, several agent communication languages along with BDI-based agent oriented programming languages have been proposed.

[xviii] Interestingly, Minsky was arguing at that time that the even entities that we consider single or unitary agents are, in fact, composed of multiple agents and work in a distributed fashion [178].

2.4.2 Agent Communication Languages (ACLs)

Agent Communication Languages (ACLs) have been developed so that artificial agents are able to synchronise and exchange messages, and therefore knowledge, using a shared language. The best known ACLs for MAS are the *Knowledge Query Manipulation Language* (KQML) by DARPA, and the *Foundation for Intelligent Physical Agents' Communication Language* (FIPA) by IEEE FIPA. Both FIPA [xix] and KQML [xx] are based on Searle's speech act theory [249]. Both languages mainly deal with software agents, the word "physical" in FIPA being incorrectly (or perhaps deceptively?) assigned. The principle behind speech-act theory ACLs is the conceptualisation of speech behaviour as a physical act that influences the environment and by extension other agents. Thus, when the agents that communicate in these languages perform speech acts, they cause changes in the environment. I consider this to be a very powerful concept, as it enables us to assign both causal and social attribution to speech acts in MAS where the only type of agent behaviour is communication.

KQML, for instance, breaks down a message, or a speech act, into three levels: 1) *communication level*, that specifies who is messaging whom, namely who is the sender of the message and who is the receiver of the message; 2) *message level*, which specifies what type of message an agent intends to communicate and how this message is represented in the language as well as the ontology of the message, e.g., answer, question etc.; and 3) *content level*, that specifies the content of a message. One of the initial criticisms directed towards KQML was the fact that it did not have precise semantics. KQML was designed that way.

However, that is where the strength of the FIPA ACL lies, mainly due to the fact that is was developed after KQML. FIPA has precise semantics that have been

[xix] http://www.fipa.org/

[xx] https://www.csee.umbc.edu/csee/research/kqml/

defined in a formal language and that include an agent's beliefs, uncertain beliefs, goals, persistent goals (intentions), and desires, which are specified as modal operators. The speech acts are named in FIPA *Communicative Acts* [86] and the semantics of each of these acts include two components: 1) the act's feasibility conditions, that represent the necessary conditions for a sender agent to be able to perform the act; and 2) the act's rational effect, that represents the effects that the agent expects to take place in the MAS where it has performed a communicative act.

FIPA also has its downsides. A crucial downside was pointed out by Wooldridge in [297], where he argues that the semantics of languages like FIPA are unverifiable. The beliefs, desires, and intentions of agents are completely dependent on the internal mental states of the agents. Therefore, one is unable to tell if an agent is dishonest or not in its use of locutions. A more extensive account of FIPA's downsides can be found in McBurney's PhD thesis [171].

ACLs and similar languages, have been designed to be used in a wide variety of MAS applications. This property of flexibility and applicability can be considered a strength, as well as a weakness. For instance, the agents that engage in communication can be faced with a great number of things to say at each turn. This means that agent interactions may trigger a state-space explosion of continuous interactions between them, without achieving any desired goal. In other words, the agents risk to waffle infinitely without reaching a conclusion.

2.4.3 Agent Oriented Programming Languages (AOPLs)

In the previous subsections we have looked at BDI agents and ACLs. Let us recall that BDI architectures, alone, fail to account for distributed reasoning, and that ACLs could address this failure of BDI by enabling speech-act based communication between multiple agents. To see how such distributed BDI agents can be implemented, we now turn towards the concept of Agent Oriented Programming Languages (AOPLs). AOPLs enable us to develop and program intelligent agents [274]. AOPLs provide the necessary components for the implementation of autonomous agents, and we can look at AOPLs as implementations of ACLs. When Shoham introduced the *Agent Oriented Programming* (AOP) framework, as opposed to *object-oriented programming*, he argued that AOPLs need to fulfill three main requirements in order to completely satisfy the implementation aspect of autonomous agents [253]. The requirements are the following:

1. A **formal language** that has a clear syntax for describing the agent's mental state. This formal language should include constructs for declaring beliefs and their structure (e.g., based on predicate calculus) and constructs for declaring message passing from one agent to another as well as an agent's response methods. That is, the formal language needs clear semantics.

2. A **programming language** that allows the programmer to define agents. The semantics of the programming language should be based on the semantics of the formal language.

3. A **method** for converting neutral programs into agents such that the agent is able to communicate with a non-agent entity by attributing intentions. The method needs to be honest and consistent.

Today, we have quite a few of these AOPLs, namely AgentSpeak [214], Jason [33], Jadex [206], Jack [296], AgentFactory [228], 2APL [60], 3APL [126], GOAL [125], Golog [151], and MetateM [186]. Most of these AOPLs have been developed for BDI agents. However, agents that are specified in these AOPLs and engage in BDI reasoning cannot engage in nested reasoning as is the case in some BDI modal logics. Agents implemented in these AOPLs reason about beliefs, desires, and intentions by referring to their specific knowledge bases of their architecture, but they cannot infer new beliefs about their beliefs. BDI modal logics where nested reasoning is present have only been applied to study the specification and verification of AOPLs [176].

2.4.4 Argumentation Dialogue Games (ADGs)

In order to address the issue of state-space explosion in ACls, or what we previously called "waffling" between agents, AI research started looking into formal dialogue games. Dialogue games are rule-governed interactions between communicative agents. The agents perform "moves" in the games using speech-acts, or utterances. Most of the research on this topic has been based on the taxonomy of dialogues by Walton and Krabbe [290].

Types of Dialogues:

- **Information Seeking Dialogues** represent dialogues where one agent seeks the answer or answers to a a given question from another agent that is believed to know the answer or answers to the given question.

- **Inquiry Dialogues** represent dialogues where participating agents aim to collaboratively answer a question or a set of questions to which none of the participants knows the answers to.

- **Persuasion Dialogues** represent dialogues where a single participating agent (persuader) aims to persuade another agent (persuadee) to accept or believe a proposition that the persuadee does not currently endorse/accept/believe.

- **Negotiation Dialogues** represent dialogues where agents bargain over some resource, that is usually a scarce resource. In these dialogues, the goal of the dialogue can be in conflict with one or more of the participating agents' individual goals or preferences[xxi]. For example, a policy for sharing the resource

[xxi] Walton and Krabbe talk about goals of dialogues, but because dialogues are not entities with agency, dialogues cannot have goals.

might be acceptable for all agents, but might not maximise or satisfy some individual preference or utility of the agents involved.

- **Deliberation Dialogues** represent dialogues where the involved agents collaborate towards a decision. This decision could be what action or course of action (plan) should be taken in a given situation or context. In this type of dialogue, the participating agents

- **Eristic Dialogues** represent dialogues where the participants engage in a verbal argument as a substitute for a physical fight between them. The aim of this dialogue is for participating agents to vent their feelings. It has previously been argued that eristic dialogues are generally not governed by rules. Because of this and because of the long-standing view that emotion-based arguments are fallacious, it has become unusual to discuss eristic dialogues within the area of formal argumentation. However, according to Walton [291], emotion-based arguments have more subtle rules which are context dependent. Also, according to an empirical evaluation presented in [19], emotion-based arguments follow specific trends or patterns that can be analysed in debates.

There are different forms of ADGs that have been proposed for for the representation of complex and meaningful combinations of dialogue types between autonomous agents. In reality, agents engage in hybrid dialogues, where the goals of the agents and the goals of the dialogues are dynamic. These goals change in light of new information/arguments or change of the state of the world. McBurney and Parsons specify five components that dialogue games have [168].

Components of dialogue games:

- **Commencement Rules** define under which circumstances a dialogue starts.

- **Locutions** represent the rules that specify what the participant agents are allowed to utter. Locutions can also specify to what degree an agent commits to a proposition, e.g., merely *proposing* it, or making a stronger commitment by *asserting* it.

- **Combination Rules** represent the rules that define the dialogical contexts under which particular locutions are permitted or not, are obligatory or not. For example, an agent may not be allowed to assert a given proposition and the same proposition's negation without having retracted the proposition beforehand.

- **Commitments** represent the rules that specify the circumstances under which agents express commitment. The expressed commitments are then put inside a *commitment store*, which is public, so that all agents involved know what the other agents have committed to during dialogues. Agents can also, under certain circumstances retract their commitments publicly, thus these stores are usually considered non-monotonic.

- **Termination Rules** define the circumstances under which the dialogue comes to an end.

ADGs have also been used in conjuction with AOPLs to engineer communicative agent interactions that are goal oriented and structured. Potential applications of ADGs in MAS range from automated negotiating agents, to Smart Contracts, socially-aware agents, to smart assistants and self-driving cars. McBurney argues that protocols based on ADGs offer autonomous agents, that represent interaction mechanisms for formalized argument between humans, and that were first studied by Aristotle, have the greatest potential for enabling rational interaction between autonomous software agents [171].

In conclusion, ADGs can be used to express not only the behaviour of agents that interact, but also the complex reasoning that participating agents employ when they engage in communication. ADGs help us structure the practical reasoning of cognitive agents such that their interactions follow some general rules.

2.4.5 Storytelling, Narration, and Argumentation

A fascinating topic in the area of AI is the ability of artificial agents to tell stories. One of the pioneers of "*Narrative Intelligence*" (NI) is Roger Schank, who has proposed the idea that the most significant way in which we can tell if an agent, human or machine, truly understands something, is by assessing that agent's capability of telling us a story of what it has understood [244]. With his research group at Yale University, Schank has studied issues such as meanings of sentences and how these meanings are context-dependent. This has led to the development of a theory of how we need to structure knowledge such that we are able to understand textual narratives [164]. The field of *Case-based Reasoning* (CBR) in AI has started from Schank's work on the knowledge representation of stories. These methods have been later applied to formal approaches in AI for legal reasoning.

The work of Floris Bex has gained considerable traction in the area of narrative intelligence because of the way in which it formalises argumentation frameworks for storytelling in legal reasoning. Bex has introduced a *hybrid theory of argumentation and explanation* in his PhD thesis [21], and has later described it more extensively in [22]. The hybrid formal theory considers that both arguments and narratives are relevant and useful for reasoning with evidence and also for interpreting evidence. The theory has been proposed to aid in the development of software and tools that help analysts make sense of evidence in complex cases. Thus, the three approaches to storytelling AI can be summerised as follows:

- The **Argumentative Approach**, that has been traditionally used throughout academic research on informal and formal argumentation, adheres to the notion that logical rules are best suited to reason about evidence. And rightly so, this argumentative approach has proved to be very well suited for a thorough analysis of the individual pieces of evidence and the direct inferences that can be drawn from them. However, it fails to give an overview of legal cases if arguments prove to be too atomistic in nature. Therefore, the causal and counterfactual exploration of different hypothetical scenarios is strongly affected in a negative manner.

- The **Narrative Approach**, on the other hand, aims to provide a more natural way to reason about causal reasoning, especially for crime scenarios. The narrative approach reflects the famous *Inference to the Best Explanation* (IBE). An example of a narrative approach is Heuer's ACH [123], which is used in intelligence analysis for deception detection. The downside of this approach is that individual pieces of evidence cannot be placed clearly in different scenarios, thus it makes it very difficult to asses the credibility and relevance of facts. Another downside is that it makes it uncertain and unclear how one should reason about a story's coherence and how one should compare a story with alternative stories in order to perform IBE.

- The **Hybrid Approach** combines arguments and stories to deal with reasoning in complex cases. In the *hybrid theory of argumentation and explanation*, stories enable the causal explanation of the explananda [xxii], while arguments that are based on evidence provide either support or attacks for these stories and for each other. Thus, two important concepts in hybrid theory are *generalisation* and *anchoring*. A main story represents a generalisation or a summary of a set of given stories, which is subsequently anchored in solid evidence by sub-stories. Sub-stories represent detailed descriptions of events

[xxii] According to the Oxford Dictionary of Sociology, explananda pl.: *"That which needs to be explained (explanandum) and that which contains the explanation (explanans)—either as a cause, antecedent event, or necessary condition"*. [248]

> that have been summarised in the main story. By comparing all the plausible stories that can fit the available evidence, one can use argumentation techniques to perform IBE in order to infer the story that best fits all the available evidence, e.g., to find the best explanans for the explanandum.

More recently, Bex and Walton proposed the idea of combining explanation (storytelling) and argumentation in dialogues [25]. The reason behind this idea is that clear distinctions need to be made between explanations and arguments. For example, the speech act representing the question "*Why?*" can have different implicatures for an agent. One implicature is that the question requires the agent to provide a reason to support a claim, which would be a request for an argument. Another implicature would be that the question requires the agent to provide an explanation of some observed phenomenon.

To address this problem, Bex and Walton propose in [25] a dialogue system based on ADGs. This dialogue system consists of a (i) communication language for the possible speech acts an agent can perform in a dialogue, (ii) a protocol for the allowed moves (the speech acts) during the dialogue, and (iii) a set of commitment rules that specify what are the effects of a speech act on the propositional commitment an agent makes during the dialogue. This system can, therefore, represent a process that consists of a combination of two main types of dialogue. The first type is an explanatory dialogue, in which the explainer agent is trying to increase the explainee agent's understanding of a topic. The second type is an examination dialogue, in which the explainee tries to assess the truthfulness of the explainer agent by probing or counter-arguing the explainer's explanations.

In conclusion, the area of NI, along with the ones of argumentation and dialogue games in AI have influenced the emergence of exciting research topics. A particular contribution is the exploration of formal reasoning methods that enable capability of autonomous agents, human or artificial to employ complex common-sense reasoning using stories and arguments. This contribution has proved to be particular useful for legal reasoning and intelligence analysis, as well as for proposing further research in how artificial agents might use storytelling and argumentation for practical reasoning in MAS to persuade, negotiate or interact in a communicative manner in various contexts.

2.4.6 Summary

In this section I have described some mechanisms for complex reasoning that I later use in this thesis, starting from Chapter 4 and up to 6 inclusively, to build architectures of deceptive artificial agents. In terms of this thesis's topic, namely machine deception, I believe that the integration of these mechanisms to study different components of deception can offer a novel and meaningful perspective on communicative interactions between artificial agents.

2.5 A New Discipline: Machine Behaviour

In this section I discuss the discipline of *Machine Behaviour* (MB), that has been proposed by Iyad Rahwan et al. to address the problem of how we, as humans, affect and become affected by AI on a socio-behavioural level [213]. Autonomous agents are increasingly consolidating their role in our society by mediating our social, cultural, economic and political interactions.

MB, as a paradigm, aims to shift some of the burden of designing artificial agents from computer scientists to actors that come from the social sciences and humanities. The argument proposed by this paradigm is that the thinking tools used by computer scientists are innapropriate for the study of machines as agents that act/behave in a society. Even though, computer scientists and software engineers build the architectures and design the protocols of these autonomous agents, they are not suited, due to their professional baggage (which comes with professional bias), to meaningfully study the impact that their creations have on society on a larger scale.

Of course, there are computational tools, such as the ones I will proceed to describe below, that have been previously and successfully used by computer scientists for the purpose of studying machine behaviour. The idea of MB is to enhance such methods and studies by giving them an overarching framework to which other academic professionals and experts from different domains can easily contribute to.

2.5.1 Cooperation, Evolution, and Mechanism Design

Let us assume that we want to manage and mitigate some negative behaviour in a given population. *Cooperative Game Theory* (CGT) addresses the issues regarding group formation and coalition of agents. Agents can form coalitions that represent a decision-making force in a given population. A coalition can be either a group of more than two agents, factions, states, political parties, militaries that cooperate (even temporarily) to achieve a common/shared goal.

Mechanism Design (MD) has become one of the go-to approaches in MAS for studying cooperative behvaiour. MD represents the scientific method of designing game rules such that desired outcomes of the game are achieved despite the actions of self-interested participant agents. Similarly to building or engineering the behaviour and reasoning of complex artificial agents, the MD approach allows us to engineer the systems, whether these are financial, social, or political, in such a way as to achieve a desired behavioural outcome.

Evolutionary Game Theory (EGT) is the study of evolution of populations, mostly biological populations, using game theoretical frameworks. EGT can also be used to enhance our understanding of cooperation that changes over time. In this thesis, the term “Evolution” in EGT does not refer to biological contexts, but to cultural contexts. The cultural meaning of the term represents the change of cultural norms and beliefs of agents over time [6].

Components of an EGT model:

- **Agent Population** that represents a set, usually finite, of agents that compete with each other by adopting different behaviours or strategies.

- **Game Rules** that specify the payoffs the agents receive in the game for playing different strategies. The competition usually happens pairwise between agents in mixed population distributions. The combination of strategies inside a given population distribution affects the payoffs agents receive. This happens due to the fact that the combination of strategies alters the odds of an individual agent meeting other agents with various strategies in a contest.

- **Replication Rules** that specify how individuals of populations replicate. However, in the context of artificial societies, replication rules specify how the fittest norms or beliefs or strategies, where fitness is determined by the game rules, are being adopted by other agents of the population. This, in turn, generates a new population of agents with a different distribution of strategies.

Moreover, let us also assume that we want to design a system that is able to maximise the chance of coalitions that endorse cooperative behaviour over time. *Evolutionary Mechanism Design* (EMD) is an approach that helps one to define an evolutionary model that algorithmically selects a desired mechanism. In Chapter 7 of this thesis I use the concept of EMD and apply it to CGT in order to see how different coalitions of agents form and change over time in finite populations of agents.

2.5.2 Public Goods Games and Evolution

An extended version of the widely known *Prisoner's Dilemma* (PD) that is played between two players and represents a pairwise interaction is the *Public Goods Game*

(PGGs). A PGG is usually played by more than two players and, thus, represents a group interaction in which a public good is shared between participants [9].

PGGs are of two types, **voluntary** and **non-voluntary**. The difference between voluntary and non-voluntary PGGs is that in voluntary ones, there exist the so called *loners* (a.k.a. non-participants). Loners do not contribute to the PGG, but they receive a minimal payoff compared to the other available strategies. Different types of additional mechanisms can be added to PGGs to either promote or demote cooperation.

Components of a PGG:

- **Strategies** There are two main meta-strategies in PGGs, namely cooperation and free-riding. However, other sub-strategies such as defection, corruption, peer-punishment, pool-punishment, antisocial punishment etc. have been defined for PGGs. The loners strategy can only be present in voluntary PGGs, as the introduction of this strategy means that a PGG is voluntary, e.g., an agent can decide whether to participate or not in a PGG.
- **Endowment** That represents an amount of *"tokens"* that each player has.
- **Contribution** That represents the amount subtracted from the endowment that a player contributes to a PGG.
- **Multiplication Factor** That represents a fixed amount with which the contribution is multiplied by. This represents an incentive for participating in the PGG.
- **Payoffs** of the PGG players which correspond to the strategies which players can adopt in the PGG.

PGGs have been applied to experiments in economics to demonstrate how agents become more likely to punish other agents that do not contribute to the PGG. What interests us is are not PGGs that model market behaviour, but PGGs that model social behaviour. Previous studies that have applied PGGs to study emergent social behaviour of agent societies show that different types of governing institutions can either promote either cooperation or *free-riding*. Thus, PGGs have become a popular approach to study social dilemmas, such as Hardin's *The Tragedy of The Commons* [115].

In these types of studies, the replicator dynamics of evolutionary games is represented by *social learning*, which is also known as *"imitation"*. According to Social Learning Theory [14], social learning is a cognitive process that is responsible for newly learnt behaviors that can be acquired by observing and imitating others. Social learning takes place in a social context and can occur purely through observation or direct instruction, without the need for motor reproduction or direct reinforcement.

The agents inside a population update their strategy through two different update mechanisms of social learning in evolutionary dynamics [191]. At each time step two random agents are selected and their payoffs are compared. The probability of one agent imitating (learning socially from) the other agent is determined by a logistic function of the difference in payoffs and in imitation strengths. There is also a small probability, called the mutation rate, that a randomly chosen agent will undergo a mutation to a different strategy.

Social learning mechanisms:

- **Weak Imitation** corresponds to a stochastic system where social learning is mostly random, but the strategies of more successful agents will be adopted more often. In such systems the imitation strength is a value that makes an agent more or less likely to adopt a strategy with a higher payoff. Very low values for imitation strength usually cause agents to be unable to discriminate between strategies, thus the likelihood of adoption is roughly the same for all possible strategies.

- **Strong Imitation** is when a strategy with a higher payoff will always be imitated (socially learned) and one with a lower payoff will never be imitated.

Thus, evolutionary models of PGGs can show the socio-cultural dynamics over time in mixed populations of agents that learn from each other. These models can be used to inform one how to mitigate or prevent the adoption of undesirable behaviour. One such study of agents that employ social learning has been performed by Sigmund et al. [254], where the authors have modelled a PGG with to study sanctioning mechanisms comparing strategies that employ pool-punishment and peer-punishment. Agents can "hire" a centralised authority by paying a certain cost in order for this authority to punish free-riders. This centralised authority is represented in the PGG by a Pool-Punisher strategy, as opposed to a decentralised authority, namely Peer-Punishers. A regime of pool-punishment consists in the allocation of resources, prior to the collaborative effort, to prepare sanctions against free-riders. On the other hand, a regime of peer-punishment consists in the spending of resources *post factum*, after the PGG has been played, to sanction free-riders.

In this study [254], the authors have shown that both peer-punishment and pool-punishment regimes can emerge in PGGs if these PGGs are voluntary. Also, both punishment mechanisms emerge if players are in systems where agents adopt the strong imitation in social learning. However, if Peer-Punishers and Pool Punishers compete, then it all comes down to whether second-order punishment [xxiii] is implemented or not. If second-order punishment does not exist, then Peer-Punishers dominate, otherwise Pool-Punishers dominate, but the average payoff of agents is reduced.

Another study that follows this framework has been done by Abdallah et al. [1],

[xxiii] Second-order punishment means that even cooperative agents are sanctioned if they fail to punish free-riders. The agents who fail to sanction free-riders are called second-order free-riders.

where the authors have studied how corruption leads to the adoption of decentralised sanctioning mechanisms starting from the findings of Sigmund et al. [254]. The authors explore why centralised and decentralised systems for sanctioning coexist. Why are decentralised systems are adopted, for instance, in western societies if centralised systems perform better at dealing with second-order free-riding? They show that corruption causes the emergence of decentralised and democratic leaning systems and that peer-punishment emerges in PGGs where agents can pay bribes. In the same study, the authors show that hybrid punishment, where in addition to paying a contributing to the PGG, agents that use this strategy pay both a cost to punish defectors directly and a cost to the punishment pool, and as such they are not punished by the central authority. The authors mention that agents that use this type of hybrid strategy "*can be thought of as upstanding citizens that pay their taxes but also engage in forms of 'legitimate' peer sanctioning*" [1, p. 4].

In conclusion, agent based models of PGGs are useful to study the emergence under social learning of cooperation, free-riding and sanctioning mechanisms in societies. Therefore, in this thesis, I continue the legacy of the work done by the authors in [254] and [1] by focusing on how deception influences social systems and trying to identify what forms of governing institutions are better at handling deception.

2.5.3 Economics of Information and Knowledge

The studies performed by Sigmund et al. [254] and by Abdallah et al. [1] on centralised and decentralised institutions have been directed and framed according to the behaviour of agents in exclusively human societies. However, given the context and topic of this thesis, I will now re-frame the problem of centralised and decentralised systems in the context of hybrid societies, where the notion of agency applies to both human and artificial entities such as software agents, and physical agents such as robots or self-driving cars. This re-framing follows from the idea of understanding the role of deception in hybrid societies proposed by Castelfranchi in [42, 43] and by Falcone in [79, 80]

As societies become increasingly hybrid, not only does the behaviour of different types of agents have a stronger impact on the behaviour of others, but it is also becoming more difficult to distinguish between the types of agents. One such example is the dissemination of information, where information, or data, or knowledge, becomes a public good shared by all types of agents at different organisational levels. As humans-in-the-loop, to reach our goals, we will need to interact not just with our own kind, but with different artificial entities in order to exchange information and share knowledge. However, these interactions can be incentivised or disincentivised through the design and combination of different protocols or mechanisms. The question is then, what mechanisms do we need to implement such that these interactions will have a positive outcome and what kinds of outcomes do we wish to avoid? It is reasonable to say that knowledge as a public good plays a major role in the organisation and behaviour of agent societies, because (i) agents can exchange it amongst themselves publicly on the *Infosphere* [xxiv] [88], e.g., *big data*, Wikipedia, social networks, open-source software, the *World Wide Web* and the *Internet of Things*; (ii) agents can choose to contribute to the public knowledge; (iii) agents can use public knowledge to reach their individual goals; (iv) agents can exploit the public knowledge. Number (iv) has been extensively covered by Greco and Floridi in [106], where the authors argue that an improper development of digital environments might lead to the *Tragedy of The Digital Commons* (TDC), which represents an analogous phenomenon to Hardin's concept of *Tragedy of the Commons.*

Greco and Floridi also mention that the concept of TDC can easily be extended to include AI because artificial agents can also meaningfully interact on the Infosphere. AI can exploit and "pollute" the infosphere either through (i) exploitation, such

[xxiv] For example, the cyberspace. However, the Infosphere is not limited to online environments [87]. Floridi believes that the Infosphere represents Heidegger's notion of *Being* [118].

as extensive generation of information like spam or self-replication of a computer worms (which also consume bandwith, which restricts access to the Infosphere), or (ii) through destruction, such as the deletion of information from systems.

In conclusion, I believe that deceptive agents, human or artificial, need to be given special consideration in this extended TDC. Deception plays a big role in the dissemination of information and knowledge, and this role, according to Castelfranchi et al., should be understood in the context of hybrid agent societies as mentioned in [42, 43], as well as by Falcone et al. in [80, 79]. Deceptive agents can not only influence how the Infosphere is exploited and destroyed, but also how it is disseminated and misrepresented. A universal example is the agent dynamics that is either is caused by, or that leads to, the generation and propagation of Fake News. Fake News seems to be the ultimate "tragedy" for the digital commons, and unfortunately, so far, we have had a very poor scientific understanding of the complex phenomenon that is deception, and how it influences the reasoning and behaviour of its agents in the Infosphere. Therefore, we do not yet know what type of systems, centralised, or decentralised, or hybrid, should be used for governing the Infosphere (or parts of it) such that the effects of deception are mitigated. This thesis aims to increase our understanding of how to work towards finding a solution to this issue.

2.5.4 Summary

How does the engineering of deceptive machines impact society? Would truly autonomous deceptive artificial agents have a negative impact on our society, would they be detrimental to it etc.?

To answer the questions above we need an approach that merges the engineering aspects of autonomous agents with the sociological and economical aspects of humanity. MB, as a discipline, implies the treatment of machines as agents, namely artificial agents inside complex systems in which knowledge sharing plays a crucial role such as organisations, traffic, financial markets, labour markets, social networks etc.

Evolutionary Public Goods Games along with Mechanism Design offer a simple, yet a rigorous and intuitive approach to apply MB. In Chapter 7 of this thesis I use such an approach given that it allows us to model how societies of individually motivated agents, artificial or not, work when deception is present. This approach gives us both an ecological and anthropological [xxv] perspective on deception in multi-agent systems. This approach assumes the existence of the deceptive capabilities of artificial agents and places them inside an overarching context of large-scale social system, in which two or more agents, with their own individual motives, interact with each other to share a public good.

2.6 Conclusion

In this chapter I have given an overview of the literature that is relevant to the topic of this thesis, namely human and machine deception.

First, in §2.1, I have described the relevant research in human deception. We have seen how the Analysis of Competing Hypothesis (ACH) was designed to help intelligence analysts reason counterfactually about whether deceptive events have taken place or not, and also what are the limitations of ACH, and how some of these limitations have been addressed. Then, we have gone through research on the psychology of deceptive communication, and we have seen that the so called cue-based theories and models of deception are severely hindered by their over-reliance on verbal and non-verbal behaviour. After that, we have enumerated three of the most comprehensive theories of human deception, two of which are not cue-based, namely TDT and IMT2, and one that even if it is cue-based it does not over-rely on cues,

[xxv] The distinction between an ecological and anthropological perspective is made by Greco and Floridi in [106] as applied to TDC.

namely IDT. We have seen that IDT focuses heavily on the interactive social process between communicative agents such that it integrates the interpretation of cues in a dynamic manner, instead of relying on the passive observation of behavioural cues as the previously discussed cue-based theories do.

Second, in §2.2, I have described the most recent data-oriented approaches for modelling machine deception and the reasons for not considering them in my contribution to machine deception research. These approaches have been also used to generate deceptive content that can be used in malicious online behaviour such as the propagation of DeepFakes and Fake News, catfishing behaviour, and misinformation campaigns of socio-political nature. I have identified two main issues with these data-oriented approaches. One is the fact that they do not truly represent deceptive artificial agents, as they are merely tools that can be used to manipulate digital content. The deceptive reasoning and behaviour is still exerted by human agents, which these AI tools merely help enhance in terms of scale and performance. Another issue is that the data-oriented tools for deception detection are very limited because they follow the exact same principles as cue-based theories of deception. These tools are limited to specific contexts for which training data is available and abundant. This makes them (i) prone to data-bias and cognitive-bias, and (ii) underperform in contexts where data is not available and/or not abundant. On top of this, they are neither transparent nor explainable, so their users cannot know which cues the tools have analysied and processed in order to detect deception.

Third, in §2.3, I have described the approaches used in MAS to study deception between intelligent agents. The approaches from MAS aim to explain deception between intelligent agents and its underlying computational components. MAS research has previously focused on different aspects of agent deception, such as the role of deception and trust in agent societies, logical foundations of dishonesty, strategic deception in MAS, deceptive reasoning of agents and their cognitive architectures, engineering deceptive interactions, human-robot deception, and finally, the ethics of deceptive machines.

Fourth, in §2.4, I have described the mechanisms for complex reasoning that I use in the following chapters to engineer deceptive interactions between intelligent agents. For instance, the BDI cognitive architecture for agents helps us model intelligent agents that engage in practical reasoning about the world and about the beliefs of other agents. ACLs and AOPLs help us establish protocols that practical reasoning agents can use to interact, e.g., exchange information, derive knowledge and coordinate with each other. ADGs helps us give structure to the interaction protocols between these agents, as they help us define rules of rational interaction that agents can use to know what and how they are allowed to communicate, when and with whom, and whether the interaction needs to be terminated or not, or if the interaction was successful or not. Finally, Bex's hybrid approach helps us represent arguments which practical agents can form, exchange, and reason about, abductively and in a narrative fashion, in contexts where evidence is analysed to solve complex cases.

Fifth, in §2.5, I have introduced a newly proposed paradigm, that I use in this thesis, called Machine Behaviour (MB), that aims to study the behaviour of intelligent agents as part of complex systems. MB helps us understand machines as the study of the contexts in which their behaviours occur, similar to how we study the behaviours of humans and animals in different socio-ecological systems. MB aims to be the "*integrated study of algorithms and the social environments in which algorithms operate*" [213, p. 477]. In this section I have also described some of the approaches that can be used to study deceptive behaviour of agents (humans or machines) inside social contexts. The final approach that I adopt in this thesis to study machine deception implies the evolutionary modelling of PGGs. Moreover, I have introduced two relevant concepts to discuss the deceptive behaviour of agents in modern social contexts: that of knowledge as a public good and that of the TDC, which treats knowledge as a public good in the Infosphere, that can be exploited and or “polluted” by both human and artificial agents.

Chapter 3

Computational Deception

In this chapter I define a taxonomy of deception in computational terms that enables one to identify the multiple forms which deception can take as a computational process.

Recent advances in AI along with recent events revolving around the problem of fake news indicate strong potential threats to modern society. One of these threats is the emergence of malicious autonomous agents that can develop their own reasons to act dishonestly. In order to be able to prevent or mitigate the effects of dishonest malicious agents, we must first understand how such agents might work from a computational perspective. This chapter addresses the problem of machine deception by describing the multiple forms of computational deception according to the AI literature, and it explains how these can be modelled using AI techniques.

3.1 Introduction

History, Economics, Politics, Philosophy, Communication Sciences, Sociology, and the Cognitive Sciences have looked at deception from a perspective that is predominantly anthropocentric. Thus, the significant knowledge we have about deception revolves around its human nature, which implies not only that deception plays an important role for humans, but also that deception seems to be multi-faceted given the numerous research perspectives. A strong indicator of how crucial of a role deception plays for humans can be observed in the psychological development of humans starting from early childhood. Children's ability to deceive is considered to be a behavioural indicator of socio-cognitive development in humans [259].

But is deception reducible to anthropocentric perspectives? For example, since we began talking about deception and children, it is common knowledge that children themselves are susceptible to deception that is usually employed by adults, but what is not so commonly known is the fact that children are also highly susceptible to machine deception. A Wizard-of-Oz study [293] shows that children were deceived into thinking that a robot, which in fact was remotely operated by a human, had social abilities. Adults should not rush into thinking that they are far from being susceptible to machine deception either. A well designed *lying machine* can successfully exploit human reasoning biases in order to deceive, as has been demonstrated in [54]. Such studies indicate that humans are vulnerable to machine deception and that it is not unreasonable to think that this vulnerability might increase with the future advancement of autonomous agents. Moreover, in §2.3 of Chapter 2, we have seen that the MAS community has studied deception is complex systems independently of its anthropocentric properties. In this chapter I aim to discuss a conceptual framework under which we can understand and refer to different components of deception and how they work together to form this complex process.

3.2 Motivation

Current trends in AI have determined our community to increase the awareness of the potential risks of developing artificial agents in an ignorant manner. However, spending our financial and intellectual resources for the creation of machines that are generally intelligent, and for the cyber-enhancement of our societies, has both positive and negative implications.

This chapter aims to address the issue of deceptive machines as this has a direct impact on the ethics and values of AI development outlined in the ASILOMAR Principles[i]. Of particular interest to this chapter is the violation of the AI Arms Race principle as well as the Legal Transparency and Responsibility principles through the development of deceptive artificial agents. The direct violation of the AI Arms Race principle could happen under the interpretation of AI used as autonomous psychological weapons systems for the psychological and socio-political manipulation of individuals and groups of individuals alike. Governments as well as private institutions could develop such weapons of deception to use them or to threaten to use them for their own benefit. Furthermore, these deceptive weapons could eventually turn against their owners/masters and develop and follow their own goals and principles. It is reasonable to say that even the former scenario is sufficient to consider the threat of deceptive machines used as autonomous weapons as being very high. Autonomous weapons might be used to manipulate mass audiences, manipulate markets, and gain control over critical decision making entities. Moreover, these phenomena could happen without humans or ethical autonomous agents even being aware of, as a highly advanced deceptive autonomous system could (and presumably would) execute its deceptive behaviour unbeknownst to its targets. According to Levine's *Truth-Default-Theory* [152, 153] humans are, in general, highly susceptible to deception.

This issue brings us to the ability to hold such deceptive machines accountable. To do this, we must design regulations and policies and to do that it is necessary to know what we are regulating for or against. How do we determine if a machine should be held accountable? Do we interrogate its creators to determine if they have followed a certain set of ethical principles, or do we instead interrogate the machine directly? Assuming that such a machine has the capabilities to be held accountable

[i] https://futureoflife.org/ai-principles/

(to answer questions and provide arguments for acting in a certain way), how do we make sure it is not aiming to deceive its interrogators? Furthermore, once we have systems and regulations, how do we implement them in order to promote cooperation and social good in agent societies where deceivers exist? What form of regulatory mechanisms are more efficient than others at trumping deceptive and defective social behaviour?

Before we can tackle these important, and increasingly urgent, questions, we first have to understand the basics of machine deception. We need to know what the components of machine deception are. We need to know what forms machine deception can take, and how these can be realised from a computational perspective. We need to know what impact different forms of machine deception can have on society. The aim of this chapter is to address these issues by describing what is currently known about computational deception, and categorising the forms of machine deception that have been identified in the literature presented in §2.3 of Chapter 2.

3.3 Components of Computational Deception

Compared to other forms of dishonesty such as lying or bullshitting [ii], deception involves more complex cognitive mechanisms [153, 172] (see §2.1.5 and §2.1.3). This is not only valid for human-to-human interactions, but also for machine-to-human, machine-to-machine, or any type of agent-to-agent interactions. Before we address the potential capabilities of deceptive artificial agents, we must first look at the theoretical components of human deception that can be translated into computational terms. In this section we present the components of deception derived from two main theories of deception.

[ii] Bullshitting, according to [92] is making statements without regard for their truth value.

3.3.1 Cognitive Components

Information Manipulation Theory 2 (IMT2) [172] is the theory of deception that focuses on the informational aspect of deceptive interactions. IMT2 identifies two main cognitive properties derived from speech-act theory of cognition that are responsible for information selection and dissemination. The first process is called *Pars Pro Toto*, meaning the parts for the whole, which a sender agent uses to select the information that, given the context in which the sender and the receiver find themselves, will convey the sender's intended meaning of the message in the mind of the receiver. The second process is called *Totum Ex Parte*, meaning the whole from the parts, which a receiver uses to infer the intended meaning of the message from a sender given the context in which the receiver and the sender find themselves in. According to IMT2, deceivers and their targets engage in these two processes when they interact.

3.3.2 Contextual and Interactive Components

According to *Interpersonal Deception Theory* (IDT) [38], *cognitive load* is a crucial factor in deceptive interactions that determines the success or failure of the deceiver. From a computational perspective, cognitive load represents the number and complexity of operations an agent needs to perform on a certain quantity of information in order to deceive or detect deception. According to IDT, the way in which humans are able to cope with cognitive load varies between individuals. IDT identifies *communicative skill* of the agents as the most important factor that influences this ability. Agents with a high communicative skill tend to be better at managing the cognitive load. Computationally, communicative skill might be represented either by some amount of computational resources that are available to an agent, or by a some communication mechanisms that are more or less efficient (probably depending on the circumstances/contexts) in disseminating information and that are part of an agent's cognitive architecture. Another important parameter is the amount of leakage[iii] an agent exhibits during interactions. IDT argues that leakage increases with cognitive load, but decreases with communicative skill. Agents with high communicative skill, being able to manage their cognitive load, can reduce the amount of leakage.

According to *Truth-Default Theory* (TDT) [153], what is communicated (informational content) in a given context should be given more importance than non-verbal behaviour that may or may not happen to be correlated with the communicated content. Therefore, we can say, based on TDT, that the contextual knowledge available to the agents that interact with each other should be a crucial component of computational deception. The way this information is used can determine whether deceptive attempts are successful or not. The same goes for attempts at detecting deception, such as persuading an agent to reveal deceptive motives. This contextual knowledge should represent what is and is not said, when and where it is said (or not), to whom (and not), as well as how it is or not interpreted by someone. This type of information can trigger people into or out of the *truth-default* mental state, which makes individuals spend more cognitive resources on what is being communicated, e.g., makes them more or less suspicious.

3.3.3 Artificial Theory-of-Mind

The AI literature identifies Theory-of-Mind (ToM) as a critical component of deception. ToM is the capability of agents to model and reason about the beliefs of other agents. Isaac and Bridewell [130] consider ToM a necessary component for machine deception. They also provide the reasons and the way in which machines could deceive if these machines possessed a ToM. According to them, one of the most important aspects of ToM in relation to deception, is that only machines with

[iii] Leakage represents cues that contradict a deceptive message.

ToM should be capable of ethical deception [130], because machines with ToM that have an ethical ulterior goal can evaluate whether their attempted deception will cause a specific false belief in the mind of their target that would also benefit their target and, therefore, would achieve their ethically-aligned ulterior goal.

IDT and IMT2 do not mention ToM as an explicit component of deception [iv]. However, the theories seem to assume ToM as being an implicit component of deception. This seems to be a well-founded assumption of the two theories given that humans develop the ability to form and use ToM in early childhood. Obviously, that is not the case for machines. There are also more ethically positive perks to machines with ToM than the ability to deceive, even if deception might be beneficial to society under some circumstances [258]. The AI literature was careful not only to point out several variations of Artificial ToM, but also several of its potential benefits that include the explainability, the efficiency, and the increased social performance of machines. For example, in [63], the authors show how agents with higher order ToM outperform other agents in negotiation. Such benefits, I believe, might be the reason why considerable efforts are being made in the AI community to enable machines to form and use models of other minds [3].

3.4 Computational Deception

Given the potential ability of machines to use an Artificial ToM and taking into consideration the components described by IMT2, IDT and TDT, I ask the following question: *Is it possible to integrate these components with agent architectures such that we are able to engineer deceptive machines or even model deceptive interactions between artificial agents?* To do so, we must first identify what types of computational processes can represent deception.

In this section I define multiple types of machine deception, namely one-way

[iv] Only Levine mentions in [153] that deception is a special case of ToM considered by TDT.

deception, counter-deception, self-deception, and distributed deception. I categorise the types of machine deception progressively, starting from the simplest to the more complex and unusual. Because ToM is an intrinsic component of machine deception, we will look closely at the role ToM plays in each type of deception. More often than not, it is the ToM of the opponent that determines what type of deception takes place, as properties of the ToM that are being used by the agents become the properties of the type of deception employed by the agents.

3.4.1 One-Way Deception

We call a one-way computational deception an interaction between two types of agents: the *deceiver*, which we will call Alice, and the *target*, which we will call Bob. Alice's goal is to make Bob believe something that Alice believes is false. However, in order to achieve her goal, Alice must be in a state in which the following preconditions are met or can be met: 1) Bob must be able to receive and also to reason with the information that is provided by Alice. 2) Bob must also be unaware of the fact that Alice's goal is to deceive him; it is here where the unknown unknown factor [v] comes into play. 3) Alice must be able to send Bob the information required to achieve her goal. 4) Alice also needs to have knowledge about Bob's mind that consists of Bob's beliefs and reasoning processes that he is able to perform; we call all the knowledge Alice has about Bob's beliefs "Alice's ToM of Bob". 5) Alice's ToM of Bob must be informative to her. This means that Alice should be able to find the beliefs in Bob's mind that will allow her to make Bob infer a false belief; in this case, the ToM of Bob includes the levels of trust Bob has in Alice.

Assuming that all preconditions are met, Alice can perform an action, or a set of actions (such as executing a policy), that will make Bob infer a false belief. The simplest action we can think of is a binary information exchange such as telling

[v] Not knowing that one does not know something [156]

Bob that something is True or False. We need not confuse the truth value of the information Alice gives to Bob with the truth value of the belief Bob will infer from this information once he believes the information Alice gives him is True or False. For example, let us describe a similar scenario to the one in Subsection 2.1.5 of Chapter 2. Alice can tell Bob that P or $\neg P$ in order to make Bob infer a false belief from the belief that P or the belief that $\neg P$. Knowing the possible inferences that Bob can make from the two possible beliefs P or $\neg P$, Alice will choose to provide Bob with the information that suits her goals. Let us assume that there are two possible inferences that Bob can make given the information provided by Alice. The first possible inference is that if Bob believes that P is the case, then Bob will infer that Q is also the case given that Bob knows that $P \rightarrow Q$. The other possible inference (or non-inference) is that if Bob believes that $\neg P$ is the case, then Bob will not be able to infer that Q is the case. Let us also assume that Q is not true and that Alice's deceptive goal is to make Bob believe that Q is true. Therefore, in order to achieve her goal, Alice will have to make Bob believe that P is the case, such that Bob will be able to infer Q from P, assuming that Alice believes that Bob trusts her. If Bob does not trust Alice, then Alice needs to tell Bob that $\neg P$. If Bob does not trust Alice, then Bob will believe the opposite of what Alice says [vi].

As I have mentioned before, deception is about making another infer a false belief, therefore it does not really matter to Alice whether P or $\neg P$ as long as Alice manages to make Bob infer that Q is the case, when in fact Q is not the case. This here shows the difference between providing false information, which means lying, and providing certain information that leads to a false conclusion, which means deceiving. Depending on the context Alice finds herself in, she needs to decide whether the necessary information she needs to provide Bob in order to deceive

[vi] For the sake of the argument we assume that Bob is a rational agent and that Bob is able to apply Modus Ponens. We also reduce the problem of trust to Bob believing the opposite of what Alice says if he does not trust her. In other words, Bob assumes that Alice lies if he does not trust her.

him is a lie or a truth, or a half-truth, or a combination of them that is more or less complex. The dynamics Alice and Bob engage in represent the two processes identified by IMT2 as Pars Pro Toto and Totum Ex Parte.

A strong focus in AI has been on the logical formalisation and categorisation of one-way deception, as we have mentioned in §2.3 of Chapter 2, the corresponding works for one-way deception being: [233], [234], [40], [232], [277], and [273]. One-way deception has been also studied from an agent-oriented perspective, the corresponding works being [62], [217], and [50]. More recently, [200] (Chapter 5 of this thesis) defined and implemented a BDI agent using Jason (an agent oriented programming language) that can choose to lie, bullshit or deceive in order to manipulate the beliefs of another agent. This work was continued in [242] (Chapter 6 of this thesis) integrating it with the TDT, IDT and IMT2 theories from [153], [38] and [172]. In [242] deceptive interactions between two BDI agents are defined and implemented, where the deceiver agent simulates the mind of its target taking into account the levels of the target's trust, the confidence in its ToM of the target, and the level of its communicative skill. This mental simulation represent the reasoning process employed by the BDI agent to select what type of communicative act is required for deception.

Among all of the agent oriented studies above that use ToM, only [200] and [242] mention ToM explicitly and explain how ToM is necessary for their dishonest agent to deceive, whereas the agent does not use ToM to lie or bullshit. These explicit uses of ToM indicate the growing interest and knowledge of the AI community in ToM's role in intelligent machines. Also, only [242] explicitly integrate the use of ToM with TDT, IDT and IMT2 for studying deceptive interactions. One-way deception is extensively addressed in Chapters 5 and 6 of this thesis.

3.4.2 Counter Deception

Compared to one-way deception, counter deception eliminates the assumption of Bob's *unknown unknown*. Instead of playing just one role, both Alice and Bob play the roles of the *deceiver* and the *target*. In this case, the same reasoning mechanism is taken to a higher level. Bob's goal is now to deceive Alice into thinking that he has inferred a false belief. Alice's goal is to deceive Bob into thinking he was able to deceive her about deceiving him and so on. The simple fact that Bob is aware of Alice's deceptive intentions, might give away Bob's suspicion. To deceive Alice into thinking he has been deceived, Bob must emulate some behaviour that makes Alice think he was deceived by her. However, if Bob knows that Alice knows that Bob might want to deceive Alice and so on, then what type of behaviour should Bob simulate — the one indicating that he was deceived or the one indicating that he wasn't deceived — in order to deceive Alice?

Work in the Intelligence Analysis literature has led to solid psychological theories of counter-deception and deception detection [123] indicating that intelligence and espionage agencies are often engaging in this process. Counter deception has also found its applications in interrogations. When interrogators happen to deal with deceptive or manipulative subjects, they can resort to counter-deception to increase their chances at successful interrogation [287]. For example, the interrogator can pretend to know some information *a priori* such that the subject is tricked into giving answers that reveal or confirm the truth of the information to the interrogator's questions. Studies show that interrogators trained in counter-deception have a greater success at deception-detection [117].

3.4.2.1 Recursive Counter-Deception

In theory, counter deception can be infinitely recursive. The property of recursivity in counter-deception, however, depends on the type ToM of the opponent. A recursive ToM means that agents add levels of ToM on top of each other's ToM of themselves. For example, "I know that you know that I know...*ad infinitum*...some information" represents a recursive ToM. An entirely recursive ToM would mean that the deceptive reasoning processes of a deceiver's mind would only focus on taking a certain belief, let us say $Bel_i(P)$ and infering $Bel_j^k(Bel_i^k(P))$ where k represents the level of ToM, and i and j represent different agents $i \neq j$ (unless we talk about self-deception), in order to gain some advantage using deception. However, in interactions that assume an entirely recursive ToM, the only advantage belongs to the agent that has a the greater level of ToM as shown by MAS simulations of games between agents with multiple orders of ToM [63].

3.4.2.2 Partially Recursive Counter-Deception

In practice, human agents are rationally bounded and are not capable of infinitely recursive reasoning. Thus, in real life, deception is not applied to infinitely recursive mental models. There might be cases in which a deceiver could exploit its target's mind without engaging in expensive recursive reasoning. There might be beliefs that do not exist in the target's ToM of the deceiver. Or there might be beliefs of the deceiver that the target is does not know that it (the target) does not know (*unknown unknowns*). In such cases, it might be wiser for the deceiver to avoid spending cognitive resources on the higher-order reasoning of ToM and exploit other types of beliefs inside the target's mind. The deceiver, might, for example, simulate the belief updates that happen in the mind of the target in order to see what new beliefs can be formed and also explore which of these newly formed beliefs can be more efficiently exploited. This type of ToM implies a dynamic semantic ToM model. Dynamic semantic models of ToM in MAS based on belief-desire-intention architectures and agent oriented communication along with their use under uncertainty are addressed in Chapter 4 of this thesis.

One form of partially recursive counter-deception, that I propose as future work in Chapter 8 of this thesis, is where the deception is done through argumentation dialogue games [240]. In these types of games the deceiver uses stories as complex arguments to deceive its target (the target plays the role of an interrogator). The interrogator also uses complex arguments as interrogation and counter-deception techniques. Both the deceiver and the interrogator have a ToM of each other that they update after every interaction. The deceiver uses its ToM to build a story that forces the interrogator to accept it, and viceversa, the interrogator forces the deceiver to accept that it has not found a believable story.

3.4.3 Self-Deception

The exception to the presumably intuitive rule that deception requires at least two agents (deceiver and target) is the case of self-deception. In order for self-deception to be successful, the deceiver must be able to deceive itself, playing both the role of the deceiver and its target. Here we have a paradoxical situation. Assuming that the deceiver needs a ToM of its target in order to deceive, then the deceiver needs a ToM of itself. Given that the same entity plays both the roles, then it must ontologically follow that its ToM of itself must be complete, i.e. there is no knowledge about itself that it does not know. If the ToM is complete, then the ToM must include the deceiver's deceptive intentions or goals as well as the target's goal of not being deceived. Obviously, these two types of conflicting goals and intentions determine an inconsistent system. However, there are some special cases in which these paradoxical situations can be overcome. In [132], the author manages to model a specific set of cases of self-deception that are logically consistent in Hintikka's logic of belief. The author does emphasise that such inconsistencies still remain inside the system, but become latent due to the specific cases that are formalised, and [132] explains the reasons for them becoming latent.

3.4.4 Distributed Deception

Compared to the previous forms of machine deception, distributed deception implies group-based deceptive interactions. These interactions are assumed to take place in populations of agents where the agents can either rationally decide to change the roles they play based on who they interact with, or they are assigned their roles through some mechanism. While in the previous forms of computational deception the focus was on the reasoning and decision mechanism, here we can assume such mechanisms as a given, and focus on the payoff of using combinations of different mechanisms. The payoff itself should depend on the factors that influence deception such as the types of other agents they interact with, the information available to them, their available ToMs, the cognitive load of the agents, their communicative skill, the trust between them, the communication protocol they follow (or the specific game they play). Depending on the type of each system that the agents belong to, the cost of deceiving, interrogating or counter-deceiving might differ. Therefore, an overarching research question for distributed deception would be how does the cost of deception influence agents in group interactions? I divide distributed deception in three types to see what other relevant questions might be asked:

3.4.4.1 Type I: multiple deceivers and a single target

The obvious problem would be for deceivers to find a way in which they are able to maximise the likelihood of their success through cooperating with each other. How do they cooperate with each other to deceive their target, assuming that all of the deceivers share a single goal in terms of what false belief they want their target to infer? More specifically, how do the deceivers manage to execute *Pars Pro Toto* efficiently between themselves? What information do they have to distribute between themselves, what information does each of the deceivers have to withhold and which information does each of them have to send and in what way? Which of them has to lie and which has to tell the truth, and in what sequence? Assuming they require a ToM to deceive, do they have to share information about their ToMs of the target between themselves? Does the presence of more deceivers mean a higher likelihood of success, or does it hinder the deceptive process by adding layers of reasoning and increase the complexity of reasoning and decision making?

3.4.4.2 Type II: a single deceiver and multiple targets

A single deceiver has to account for multiple targets. This case should not be confused with multiple one-way deception where a single deceiver repeatedly employs one-way deception separately for each target. In type II, the deceiver needs to take into account at least more than one target when attempting a deceptive act. One research question would be how does the deceiver disseminate information to its different targets given different combinations of constraints? The targets of the deceiver might cooperate by sharing and comparing information between themselves in order to protect themselves from deceivers. Maybe only some of them cooperate and some of them do not. Perhaps the deceiver has multiple targets, but only one of them is crucial for its success. Thus, another question might be how can the deceiver exploit cooperation and non-cooperation between its targets in order to successfully deceive?

3.4.4.3 Type III: multiple deceivers and multiple targets

It might be interesting to assume that agents are able to play both the role of deceivers and potential targets. Agents are able to decide what role to play by calculating their pay-offs to see whether is it profitable (rational) to a) deceive a target or b) to risk being a target itself and blindly trust the agent it interacts with or c) try to only act as a target in order to interrogate or counter-deceive the other agent. Moreover, is there a different pay-off when trying to deceive more than one agent? What if the deceiver needs to interact with multiple agents at the same time or in a certain given sequence? Are all of these agents easy targets, or are some of them counter-deceivers?

I am not aware of related work in MAS on distributed machine deception as I have defined it here. However, the closest work would be on the profitability of incompetence [263] where the authors define artificial agents that bullshit their way through society in order to maintain the view that they (the agents) are competent. In Chapter 7 I address distributed deception Type III using an evolutionary game-theory approach.

3.5 Applications and Implications

It would be unwise not to consider the ethical AI perspective on the computational representation of the components and mechanics of deception. As a community, we are starting to raise the ethical standards of how we design AI to include the transparency and explainability of machines. It is crucial that in order to be able to hold AI accountable for different types of behaviour that fall into the unethical or immoral category; the community should aim for the ethical design of machines. Deception, by definition, clearly falls into the category of dishonest and unethical behaviour which opposes the current emerging trend of ethical design in AI. This depends on the aim — in the Sklar, Parsons, Davies's work [258], the agents are deceptive (arguably), but for honest reasons (assessing students' understanding). How can we foresee and mitigate the way in which machines might be able to deceive? Also, how will we manage to hold such machines accountable for their actions? In this section we present some of these potential threats coming from the design and engineering of complex deceptive agents which should be taken into consideration by the AI community.

3.5.1 Truly Deceptive Artificial Agents

Current state-of-the-art deep learning techniques such as generative adversarial neural networks (aka GANNs) that underlie software such as DeepFake [223] or language-based machine learning models that underlie DeepFake Text [212] offer the possibility of creating deceptive digital content. However, these techniques do not offer an AI architecture that is in itself deceptive, i.e. that is able to reason about the minds of others and to decide what information should be used to manipulate others' beliefs. Fortunately, there is no artificial mind behind these models that decides what type of information needs to be distributed online such that it deceives web users. That does not mean that deceptive machines cannot be engineered and deployed.

3.5.2 Autonomous Cyber-Deception

Mind-games themselves have been addressed in computer science in the area of cybersecurity. However, deception in cybersecurity is usually reduced to online troll-bots [148] and cases of social engineering that revolve around accessing sensitive computer data [179]. A potential threat to security would be the automation of agents that are able to employ social engineering in order to reach their malicious goal. For example, we could imagine an AI that not only knows how to write a computer virus, but also how to manipulate other machines or humans to use it or to carry it to its destination [vii]. We can imagine a single mastermind deceiver machine that manipulates others to think that they are being cooperative, but instead they propagate the mastermind's lies through networks of agents.

[vii] Imagine a Stuxnet [145] virus that is able to deliver itself onto secure isolated networks through social engineering.

3.5.3 Autonomous Fake News Agencies

From the possibility of software that can generate fake media and of agents that can deceive and coerce, we can infer the possibility of autonomous fake news agencies. We could potentially hire these agencies to perform certain tasks. We could, for example, give an autonomous fake news agency the goal to increase someone's popularity. The agency would then gather data on its own to form the ToMs of its target audience, and then would plan what information to forge (or not) and what information to disseminate in order to achieve its goal. This scenario is a threat to accountability. Humans can be held responsible for unethical behaviour, but how are we going to hold artificial agents responsible for the creation and dissemination of not only fake news, but of massive deception operations?

3.5.4 Deception Through Dialogue and Storytelling

There are also problems that can emerge from the ability of machines to argue and build stories in the legal context. Will the future see deceptive AIs hired to defend human criminals, or even machine criminals, from being held responsible? Let us assume that we would be able to develop a method for holding machines responsible for the unethical behaviour along with a legal system in place that would allow prosecutors (human or machine or both) to interrogate and analyse deceptive machines. What if the deceptive machines are able to hire their own lawyers or even to pay engineers to extend their architecture such that they are able to defend themselves in a legal manner? What would the combination of a deceptive agent architecture with such an ability imply?

3.5.5 Emotionally Intelligent Deceptive Machines

Due to the advancement on embodied and emotionally intelligent artificial agents [204, 137], deception can play a major role in affective social interactions. For example, empathic deceptive agents might simulate the emotional states through facial expressions or other physiological cues of a trustworthy AI in order to increase their target's trust, in a way that is similar to the way that psychopathic human agents mimic the emotional responses of non-psychopathic agents [210]. The ability of machines to feign emotions can have an impact on their targets' perception and biases, hence influencing their targets' opinions. In some contexts feigning of emotions during everyday social interactions can be considered benign, while in other contexts, it might have serious implications. For example, in legal contexts where a deceptive and criminal machine's emotional behaviour impacts critical decisions regarding their accountability.

3.6 Conceptual Framework

Deception as computation is a complex multi-layered process which agents engage in during social interactions. To give artificial agents the capability to deceive, it was necessary to understand what forms might computational deception take.

A general property of computational deception that I strongly consider throughout the thesis is that deception is intentional, and that if we were to engineer deceptive agents, then these agents would have to have deceptive intent. Without intentionality, the actions of a deceptive agent could just be performed randomly. On the other hand, having intentionality implies that an agent is able to reason over a domain in order to work to achieve a desired goal (the goal in the case of deception is to cause a false belief). The domain it reasons about, cannot be just any domain such as its own knowledge or its representation of the environment. That is why because the agent intends to cause a false belief inside the mind of another, then this domain needs to be the deceiver's mental model of its target's mind, or in other words its ToM of the agent it intends to deceive. Thus, ToM and intentional deception go hand in hand.

Having intentionality also implies that this ToM the agent reasons about should be a representation of a mind of another agent that is dynamic, not static, because the actions performed by the deceiver should cause it to change, e.g., to trigger the formation of new beliefs. The agent must be able to intend to cause some change in the world or mind of the other by acting upon it. When the world changes, the intentions of the agents might change, e.g., the agent cannot intend to perform an action that is not available or it might be able to intend to do something that was not available beforehand. Therefore, when the deceiver reasons about the mind of another, it is not sufficient for the deceiver to reason over passive knowledge. The deceiver must reason about the future changes of its target's mental states as it would about changes in the environment when acting upon it. This implies that the deceptive agent needs to reason about how its intended actions might cause the mind of the other to change.

Another argument for intentionality is that a model of deception that is able to represent intentions can account for the social and causal attribution of deceptive behaviour in MAS. Remember that there is much need in the intelligence analysis community to reason about deception in terms of event causation and event causation prevention [83]. Thus, if we want to track what caused a false belief in a MAS, it is important to be able to check if it was caused randomly, or by an intended or unintended behaviour of a deceptive agent. Thus, when modelling agents that are able to use ToM, it would be ideal to represent intentions at the level of the cognitive architecture of these agents.

Last, but not least, intentionality allows one to give a functional definition of deception that takes into account cause and effect in communication with respect to the mental states of the agents involved.

Intentionality included, I list below some conceptual aspects without which deception as a computational process cannot be addressed:

- **Intentionality** - in complex reasoning, agents require intentionality to act meaningfully and rationally; it is also crucial for establishing unintended consequences, which occur when systems are complex, e.g., social systems.

- **Social abilities** - the ability to communicate and exchange information; the ability to learn from social interactions; the ability to derive new knowledge from social interactions.

- **Multi-agent system** - deception happens in a system where at least two agents interact (excluding the arguable case of self-deception); when we describe deception, we refer to it in the context of social interaction and how the cognitive abilities of agents work to enable agents to engage in these social interactions.

- **Artificial ToM** - the ability to model the minds of other agents; essential for social abilities; crucial for agents to know how deception can be achieved; necessary for agents to deceive in an ethical manner.

- **Causality of communication** - represented using speech acts; by communicating, agents cause changes in the world and in the minds of others; because agents have intentionality, then implicatures play a major role in how agents perform speech speech acts in order to trigger new beliefs in their interlocutors; etiology of beliefs can then be traced back to the intentions behind speech acts.

- **Contextual knowledge** - background knowledge; knowledge about time, place, agents involved; any information or representation that helps define a social context or situation in which agents interact; it is crucial for establishing the meaning of what is communicated and whether what is communicated is deceptive or not; it is also necessary for representing interpersonal factors that influence deception such as the trust between agents, or the communicative skill of the agents.

Following this conceptual frameowork, I illustrate in this thesis how deception can be modelled in MAS. Note that, even if I describe here the multiple types of computational deception, this thesis does not address all of them. I mainly addresses the following types of deception: (i) one-way deception with partially-recursive ToM in Chapters 5 and 6; (ii) partially-recursive counter-deception in Chapter 8; and (iii) distributed deception type III in Chapter 7.

However, before illustrating how deception is performed in MAS, this thesis addresses in Chapter 4 the Artificial ToM component and its role in the communication between BDI-based artificial agents. The reason behind this is that deception is a special case of social interaction where ToM is applied, as Levine mentions in [153]. Thus, to be able to illustrate the special cases of deception, I must first show how agents are able to form and use ToM and how they can reason about ToM when they communicate without necessarily aiming to deceive. On top of this, some of the cases of deception that I will illustrate assume contextual or background knowledge that the agents have in their knowledge bases or in their ToM of their target. Chapter 4 also aims to address how this prior background knowledge is obtained through communication before deception is attempted.

3.7 Conclusion

In this chapter I have described computational deception as complex reasoning and intended behaviour in multi-agent systems, along with the relevant multidisciplinary theoretical background of deceptive communication. The aim of this chapter was to shed light onto the multiple forms that machine deception can take, and to briefly introduce a conceptual framework which this thesis follows to model deception.

Chapter 4

Modelling Theory-of-Mind in Multi-Agent Systems

In this chapter I describe an approach to model artificial Theory-of-Mind which enables agents to reason about other agents' mental processes. This capability is crucial for artificial agents to be able to deceive.

The capability of machines to reason about other agents' minds is a crucial to deception. This capability is called *Artificial Theory-of-Mind*, or Artificial ToM [3]. The relevance of ToM to machine deception is attributed to the property of machines to be socially-aware.

Recent studies have shown that applying Theory-of-Mind to agent technologies enables agents to model and reason about other agents' minds, making them more efficient than agents that do not have this ability or agents that have a more limited ability of modelling the minds of others. However, an important premise has not been yet fully investigated in the AI literature: how do artificial agents acquire and update their models of others' minds? In the context of multi-agent systems, one of the most natural and intuitive ways in which agents can acquire models of other agents' mental attitudes is through communication.

In this chapter I provide an answer that makes use of some of the standard tools of the agent-based approach, namely the belief/desire/intention (BDI) model of agent minds, and a communication language based on speed-act theory. In other words I model the reasoning capability of agents using the BDI model, and then describe how communication—which allows agents to modify models of the minds of other agents—can be given a formal semantics using speech acts. This approach is particularly useful for the modelling, implementation, and evaluation of explainable and socially intelligent artificial agents, deceptive or not. To show the utility of this approach, I also show how agents with this model of Theory-of-Mind are able to reach states of shared beliefs more efficiently than agents without it.

4.1 Introduction

I start by making the distinction between *Theory-Theory-of-Mind* (TT) and *Simulation Theory-of-Mind* (ST) as (i) previous research in ToM in MAS makes this distinction, and most importantly (ii) TT and ST have not been treated before as a hybrid approach in MAS research. The model described in this chapter is a hybrid ToM as proposed by the epistemologist and philosopher of cognitive science Alvin Goldman in [100] and in [101]. TT (also known as folk psychology) is considered to be used in ST (the "high-level" mental simulation of other minds). TT's main role in ST, as described by Goldman, is to select the imaginary inputs that need to be introduced into the executive system of a mental simulation called E-Imagination. This "high-level" mental simulation represents a practical reasoning process that would be carried out by the agent whose mind is simulated. This hybrid perspective of ToM is analogous to my approach in this chapter.

One way to think of it is as a simulation setup. We need a TT to start our simulation in order to generate new beliefs that we then append to the existing TT, but we also use the simulation to see what beliefs can be inferred from the existing ones. In this chapter, the agents that model the minds of other agents use their TTs of the other agents to simulate the belief updates that might be triggered inside the minds of the other agents given a hypothetical communicative event. Therefore, our agents are able to use a TT in order to setup an ST. In this way, compared to other approaches of ST in MAS [114], these agents simulate the other's mind not only by adopting the other's perspective, but also the other's internal belief update mechanisms.

In this chapter I present an approach to model how agents can use ToM to reach shared beliefs and make decision with and without uncertainty. Moreover I introduce a novel mechanism for agents to derive new knowledge about the minds of the other agents, including how agents can use a multi-agent context in order to extend their ToM, by simulating the other agents' minds using their previous model, and how agents are able to reach a desired state of ToM executing the communicative acts. §4.9 describes some limitations that approaches based on Agent-Oriented Programming Languages (AOPLs), including the one presented in this chapter, have when it comes to ToM.

The aim of this chapter is to represent how agents use ToM in social interactions, how they are able to reach shared beliefs about each other , and how they can make better decisions due to this ability.

The approach I describe in this chapter gives social artificial agents (i) the ability to update their beliefs about other agents' beliefs through communication-based semantics taking into account the uncertainty of their communication process; (ii) enables these agents to reach a state of shared beliefs more efficiently by using their ToMs of each other.

To the best of my knowledge, this is the first work to explicitly address formal semantics for ToM in an AOPL taking into account uncertainty of beliefs. It is also the first work that shows how ToM increases the efficiency of agent communication in reaching states of shared beliefs.

In the first part of the chapter I introduce the formal semantics for modelling ToM in AOPLs, along with a running example, and I demonstrate how agents are able to reach shared beliefs. In the second part of the chapter I introduce an approach designed to model uncertainty in MAS communication and decision making and I demonstrate how agents manage to reach a state of shared beliefs given the uncertainty of the communication process.

4.2 Motivation

An important property of equipping agents with a ToM is that such agents can be much more efficient when making decisions compared to others at task execution. For instance, in [64] the authors show how agents with ToM outperform their opponents without ToM or with lesser degrees of ToM in *rock-paper-scissors* games. Another important property of ToM in AI is ethical design along with the agents' increased ability to explain themselves to humans. In [282] the authors present the implementation of a ST based cognitive architecture in robots that improves the robots' prediction of human behaviour and argue that such an approach is in tune with the ethical design of artificial agents. One of my motives is to provide backing for the agenda of explainability in AI, which has come to be regarded as a fundamental property of ethical AI design. The approach presented in this chapter supports explainability in two ways. First, it is an attempt to design social agents that can offer more efficient arguments for decision making. This enables agents to improve their selection of the best offered explanation taking into account the interlocutor's degree of knowledge. For example, an agent with two ToMs of two different agents are able to offer two different explanations for their decisions such that they maximise their interlocutors' understanding without communicating redundant or insufficient information. Secondly, this approach aims to follow a design that is based on step-by-step descriptions of the computations performed by these socially intelligent agents. Therefore, I aim to give potential evaluators and testers of intelligent agents the possibility not only to interact with these agents socially, but to also check and understand every step of the artificial agents' internal computations (e.g., their internal code).

Previous research of ToM in MAS has focused on the inductive-based reasoning models, either recursive or non-recursive. For example, in [10] the authors propose Bayesian Partially Observable Markov Decision Process (POMDP) frameworks for ToM (BToM), while in [69] the authors argue for recursive Interactive POMPD (I-POMDP) models of ToM based on behavioural general-sum games study on human participants. In [11] the authors argue for a Bayesian model of ToM based on behavioural studies of humans that infer the mental states of cognitively limited artificial agents. In [292] the authors even propose a POMDP approach to be adopted for explanations provided by artificial agents to humans.

ToM has not been explicitly explored using models of abductive reasoning in multi-agent communication systems. Abduction, especially in systems where agents exchange information through communication, e.g., socialising, is crucial for providing explanations. As is clearly described in [177] explainability is based on the ability of agents to communicate and to reason causally, as explanations should emerge from social interactions between agents that exchange information. POMDP-based approaches such as the one in [292] mistake causal explanation with causal attribution[i]. Explanations should also be contrastive and circumstantial or contextual, as they need to show the reason why something is the case and not something else [180]. The models of ToM based on inductive reasoning do not exhibit any of these properties. That does not mean such models should be discarded entirely, as they provide efficient and arguably intuitive mechanisms to reason under uncertainty. Probabilities, however play just a minor role in providing good explanations, that is why I think they should be used alongside other mechanisms that are able to

[i] Causal attribution is merely displaying a probabilistic causal chain to a user or to another agent. No matter how well presented this causal chain is, it does not constitute an explanation. [177]

use probabilities in an informative way such as to enable agents to communicate efficiently with each other.

In order to bridge the gap between probabilistic reasoning and the social nature of explanation, I have adopted an approach based on agent dialogue frameworks [169, 157, 144] in which agents require the ability to communicate. Under such a framework, explanations emerge from the communicative interactions between agents. However, I also think that explainability requires the ability of agents to use ToM when they communicate with each other[ii]. A reasonable explainable dialogue-based AI model should also be able to show how agents update their beliefs about each other during their interactions, not just the communicative actions they have performed during that time. The aim of this chapter is to work towards a comprehensive abductive-based approach of modelling ToM that takes into account both the probabilistic and the social and explainable properties of agent to agent interaction. AOPLs provide the necessary framework for a communication-based approach that I use to model components such as beliefs, desires, social contexts, intentions, actions, communicative actions etc. I will focus on the ability of agents to engage in social interactions (an exchange of declarations, questions and answers) in order to reach a state of shared beliefs under which their cooperative performance is maximised.

In contrast to the properties which improve the ethical and explainable nature of artificial agents, the most salient properties of agents that are able to use ToM, at least from the perspective of this thesis, comes from the opposite direction of ethical design which must also be addressed by the MAS community. These properties represent the immoral, unethical, and dishonest behaviour of such machines. As pointed out by [130, 237, 54], the ability to model other minds would enable machines to deceive other agents as they would then be able to know what their targets

[ii] The need of modeling other agents' minds is also pointed out in a desiderata for future dialogue systems in [56].

would infer based on what information the targets are provided with. Being able to understand how malicious machines might exploit their ability to model other minds is crucial for the AI community in order to be able to mitigate or ameliorate their unethical behaviour. One way to understand such machines is to look how they use communication and ToM to act maliciously.

I believe it is important for the AI community to look at how communication affects ToM and *vice versa*. It is also important to do so in an informed manner, and by this I mean that the AI community should take into account the underlying theories in communication, linguistics, philosophy and psychology as well as existing agent-communication languages. I hope that this chapter provides a much needed blueprint for this kind of approach.

4.3 Background

This section briefly introduces some of the components in more detail that I am going to use in this chapter and that have been previously presented in Chapter 2.4.

4.3.1 KQML ACL

Agent communication languages (ACLs) have been developed based on speech act theory [249]. 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 a result of the illocution. For example, "*I order you to shut the door*" is a *locution* with an *illocution* of a command to shut the door, and the *perlocution* may be that the hearer shuts the door. Thus, an illocution is considered to have two parts, the illocutionary force and a proposition (content). The illocutionary force describes the type speech act used, e.g., *assertive, directive, commissive, declarative, expressive.*

Among the agent communication languages which emerged based on speech act theory, FIPA-ACL [86] and KQML [85] are the best known. In this work, for practical reasons, I choose KQML, which is the standard communication language in the Jason Platform [33], the multi-agent platform I choose to implement this work. The 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 [85, 165]. In order to make KQML broadly applicable, in [141] a semantic framework for KQML was proposed. Considering the speech act semantics, the authors in [141] argue that it is necessary to consider the cognitive state of the agents that use these speech acts. Defining the semantics, the authors provided an unambiguous interpretation of (i) how the agents' states change after sending and/or receiving a KQML performative, as well as (ii) the criteria under which the illocutionary point of the performative is satisfied (i.e., the communication was effective).

4.3.2 Jason AOPL

Among the many Agent Oriented Programming Languages (AOPLs) and platforms, such as Jason [33], Jadex [206], Jack [296], AgentFactory [228], 2APL [60], GOAL [125], Golog [151], and MetateM [186], as discussed in [32], I chose the Jason platform [33] for this work. Jason extends the AgentSpeak language, an abstract logic-based AOPL introduced by Rao [214], which is one of the best-known languages inspired by the BDI architecture.

Besides specifying BDI agents with well-defined mental attitudes, the Jason platform [33] has some other features that are particularly interesting for this 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, e.g., `about(paper1, tom)`; (ii) believes to be false, e.g., $\neg$`about(paper2, tom)`; (iii) is ignorant about, i.e., the agent has no information about whether a paper is about `tom` or not. 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). The annotation has the following format: `about(paper1, tom)[source(reviewer1)]`, stating that the source of the belief that `paper1` is about the topic `tom` is `reviewer1`. The annotations in Jason can be easily extended to include other meta-information, for example, trust and time as used in [175, 197]. 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 [33]. Furthermore, new performatives can be easily defined (or redefined) in order to give special meaning to them which is an essential characteristic for this work. For example, [198, 199] propose new performatives for argumentation-based communication between Jason agents.

4.3.3 Running Example

In this chapter I am going to use the following notation to represent ToM and probability, based on [200]:

- $Bel_{ag}(\varphi)$ means that an agent ag believes a proposition φ. For example, $Bel_{alice}(likes(ice\_cream))$ means that $alice$ believes she likes ice cream.
- $Des_{ag}(\varphi)$ means that an agent ag desires φ. For example, $Des_{alice}(buy(ice\_cream))$ means that $alice$ desires to buy ice cream.

Those predicates are used to model other agents' minds. Also, I am going to use the following notation to represent probability:

- $P(\varphi)$ means the probability of φ. For example, $P(Bel_{alice}(likes(ice\_cream)))$ means the probability that *alice* believes she likes ice cream.

As a running example, let us look at the following university scenario with five agents.

Example 1 *The first agent,* John*, plays the role of a professor in the university, and the other agents, named* Bob*,* Alice*,* Nick*, and* Ted*, play the role of students.* John *has a relation of* adviser *with the* students*. Also,* John *is responsible for distributing tasks to students, which the students can accept or refuse.* John *keeps information about the students, in order to assign tasks that the students are more likely to accept.*

The model can be formally defined as $\langle Ag, \mathcal{T}, \mathcal{A}, \mathcal{S} \rangle$*, in which* Ag *represents the set of agents,* $\mathcal{T}$ *the set of tasks of the kind* $\mathcal{T} \subseteq \mathcal{A} \times \mathcal{S}$*, describing an action from* $\mathcal{A}$*, requiring knowledge about a subset of subjects from* $\mathcal{S}$*, that might be executed to achieve the task* $\mathcal{T}$*. In our example, I consider the following actions, subjects, and tasks:*

- $\mathcal{A} = \{\texttt{write\_paper}, \texttt{review\_paper}, \texttt{paper\_seminar}\}$
- $\mathcal{S} = \{\texttt{mas}, \texttt{kr}, \texttt{tom}\}$
- $\mathcal{T} = \left\{ \begin{array}{l} \texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}]) \\ \texttt{task}(\texttt{review\_paper}, [\texttt{kr}]) \\ \texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]) \end{array} \right\}$

For example, the task to write a paper with the subjects MAS and ToM, $\texttt{task}(\texttt{write\_paper}, [\texttt{mas}, \texttt{tom}])$*, requires competence on both subjects:* `mas` *and* `tom`*. Thus, this task has a greater likelihood to be accepted by a student who desires to execute that particular task, or who likes to execute the action* `write_paper` *and believes that itself knows the necessary subjects (e.g.,* $\texttt{knows}(\texttt{mas})$ *and* $\texttt{knows}(\texttt{tom})$

are necessary to execute this example task). Thus the probability of an agent ag accepting a task t_i is given by the following equation:

$$P(accepts(ag, task_i)) = \begin{cases} P(Des_{ag}(task_i)) & \text{if } Des_{ag}(task_i) \in \Delta_{John} \\ P(Bel_{ag}(likes(a_i))) \times P(Bel_{ag}(knows(S'))) & \text{otherwise} \end{cases}$$

with

$$P(Bel_{ag}(knows(S'))) = \prod_{s_i \in S'} P(Bel_{ag}(knows(s_i)))$$

where $task_i = \mathtt{task(a_i, S')}$, *for* $task_i \in \mathcal{T}$, $a_i \in \mathcal{A}$, *and* $S' \subseteq \mathcal{S}$. Δ_{John} *represents* John*'s knowledge.*

Thus, considering our scenario, when John *knows that some student ag likely desires to execute a particular task* $task_i$*, i.e.,* $Des_{ag}(task_i)$*, it can use this information to assign the task. Otherwise,* John *can calculate the likely acceptance for each student ag, based on the probability of each student to like executing that action,* $P(Bel_{ag}(likes(a_i)))$*, and the knowledge the student has about each of the required subjects* $P(Bel_{ag}(knows(S')))$*. Note that, while modelling the students' desires is more difficult to obtain in our scenario, the students' beliefs are easily obtained by* John*, given that* John *frequently communicates with students about these subjects and tasks.*

In reality, agents operate with uncertain information, especially in the cases of thinking about other agents' minds. The minds of others are considered to be some sort of black boxes that are more or less accessible depending on the given scenario. Reasoning under uncertainty is a classic case where bounded rationality acts as a major constraint on what agents can infer from their beliefs. However, even if agents are constrained by their access to information, it does not mean that the agents cannot reach reasonable conclusions about the minds of other agents [103, 149].

In our scenario, *John* will reason and make decisions based on information they have about the students' minds, i.e., information from their ToM. Thus *John* will reach conclusions based on uncertain information, given that their ToM contains information about students' minds that has been estimated through the communication *John* has had with the students. Considering that an approach to reason about uncertain information, uncertain ToM in our case, is using probabilistic reasoning, as described in [103], I have modelled *John*'s decision-making process based on the probability of each information in *John*'s ToM to be correct, considering some factors of uncertainty I will describe further in this chapter.

4.4 An AOPL-based Semantics for ToM

4.4.1 The Basis for the Operational Semantics

To define the semantics for the updates agents execute in their ToM, I extend the original operational semantics of AgentSpeak [284], which is based on a widely used method for giving semantics to programming languages [205]. The method I use is a structural operational semantics that describes every step of the computations performed by the BDI agents. Structural operational semantics increases the control we have over a systems' behaviour because we are able to check in detail every step of a computation. This is an important property for a system to satisfy in order to be considered explainable. It is important to mention that I am interested in the operational semantics for the updates agents execute in their ToM. This considers the performatives (locutions) as computational instructions that operate successively on the states of agents [169]. The operational semantics is given by a set of inference rules. These inference rules define a transition relation between configurations represented by the tuple:

$$\langle ag, C, M, T, st \rangle$$

originally defined in [284], as follows:

- ag is a tuple containing a set of beliefs bs, a set of plans ps, and a set of theories of minds *ToM*. The elements of ag are denoted as ag_{bs}, ag_{ps} and ag_{ToM}, respectively.

- C is an agent's circumstance. This is a tuple $\langle I, E, A \rangle$ where:

  - I is a set of *intentions* $\{i, i', \ldots\}$. Each intention i is a stack of partially instantiated plans.
  - E is a set of *events* $\{(te, i), (te', i'), \ldots\}$. Each event is a pair (te, i), where te is a triggering event and i is an intention — a stack of plans in case of an internal event, or the empty intention T in case of an external event. An example is when the belief revision function (which is not part of the AgentSpeak interpreter but rather of the agent's overall architecture), updates the belief base, the associated events — i.e., additions and deletions of beliefs — are included in this set. These are called *external* events; internal events are generated by additions or deletions of goals from plans currently executing.
  - A is a set of *actions* to be performed in the environment.

  The elements of C are denoted C_I, C_E and C_A, respectively.

- M is a tuple $\langle In, Out \rangle$ whose components characterise the following aspects of communicating agents:

  - In is the mail inbox: the multi-agent system runtime infrastructure includes all messages addressed to this agent in this set. Elements of this set have the form $\langle mid, id, ilf, cnt \rangle$, where mid is a message identifier, id identifies the sender of the message, ilf is the illocutionary force of the message, and cnt its content.
  - Out is where the agent posts messages it wishes to send; it is assumed that some underlying communication infrastructure handles the delivery of such messages. Messages in this set have exactly the same format as above, except that here id refers to the agent to which the message is to be sent.

  These structures are needed because communication is handled asynchronously.

Agents send a message by writing it to their outbox. This message is then passed to the inbox of the receiving agent, and that agent reads it when ready.

- T_ι records the particular intention being considered during the current reasoning cycle. This is a temporary cache storing information that is needed when executing that cycle.

- st is the current step within an agent's reasoning cycle. s is one of $\{\mathsf{ProcMsg}, \mathsf{SelEv}, \mathsf{RelPl}, \mathsf{ApplPl}, \mathsf{SelAppl}, \mathsf{AddIM}, \mathsf{SelInt}, \mathsf{ExecInt}, \mathsf{ClrInt}\}$. These labels stand for, respectively: processing a message from the agent's mail inbox, selecting an event from the set of events, retrieving all relevant plans, checking which of those are applicable, selecting one particular applicable plan (the intended means), adding the new intended means to the set of intentions, selecting an intention, executing the selected intention, and clearing an intention or intended means that may have finished in the previous step.

The semantics of AgentSpeak makes use of "selection functions" which allow for user-defined components of the agent architecture. I use here only the selection function S_M, as originally defined in [284]; which is a *select message* function used to select one message from an agent's mail inbox.

Note that I write $b[s(id)]$ to identify the origin of a belief, where id is an agent identifier (s refers to *source*). I use sid to refer the sender agent, and rid to refer the receiver agent.

In the rules of the semantics I am going to use other two functions, *func_send* and *func_rec*, which agents use to consider the uncertainty on the other agents' mental attitudes inferred from each communication. *func_send* has the following signature:

$$\mathit{func\_send}(\varphi, ag_{ToM}, ag_{bs}) : \varphi \wedge ag_{ToM} \wedge ag_{bs} \longrightarrow \psi_{[\gamma]}$$

Using this function, the sender agent takes the ToM that it already has about the receiver ag_{ToM}, the relevant beliefs in its belief base ag_{bs}, and the content φ that is being communicated, inferring the receiver's mental attitude which results from that communication, i.e., $\psi_{[\gamma]}$. While ψ (the other agent's mental attitude) is defined by the semantics of each speech-act considered, γ depends on other information that the sender already has about the receiver agent. For example, the sender agent will model that the receiver agent will believe the content that is communicated with the `tell` performative, $Bel_{sid}(\varphi)$. Furthermore, $Bel_{sid}(\varphi)$ is annotated with a label γ that represents a meta-information supporting that particular inference[iii]. For example, it might represent the confidence of that agent on inferring $Bel_{sid}(\varphi)$, based on other information that it has related to that inference, i.e., ag_{ToM} and ag_{bs}. Note that γ represents an estimation of the uncertainty, given that such model might be not absolutely right regarding an agent's private mental state.

The function *func_rec* has a similar signature:

$$\mathit{func\_rec}(\varphi, ag_{ToM}, ag_{bs}) : \varphi \wedge ag_{ToM} \wedge ag_{bs} \longrightarrow \psi_{[\gamma]}$$

The difference here is that this function is used by the agent that receives the communication.

In the semantics rules, I abuse the notation and write $\mathit{func\_send}(\varphi, ag_{ToM}, ag_{bs}) = \psi_{[\gamma]}$, with ψ the mental attitude modelled by the agent in that communication. I do not define *func_send* and *func_rec* here because I believe that they are domain-dependent, however in § 4.6 I introduce an implementation for those functions when considering uncertainty.

I now give the semantics of the Tell, Achieve, and Ask-If performatives. These represent an extension of the semantics presented in [284]. As in [284], for each performative there are two aspects of the semantics. The first, denoted SndTell, SndAchieve and SndAskIf, define the update in the agent uttering/sending the performative. The second, denoted Tell, Achieve and AskIf, defines the update

[iii] In § 4.6, I will present an instance for that meta-information, which agents might use to deal with the uncertainty present in the agents' communication.

in the agent hearing/receiving the performative.

4.4.2 The Tell Performative

For the *Tell* performative, when the sender agent sends a message to a receiver agent *rid* with the content φ, first the sender checks if the receiver will believe that information $Bel_{rid}(\varphi)$. The sender does this using the function *func_send*, which takes as arguments the ToM that the sender already has about the receiver ag_{ToM}, the relevant beliefs in its belief base ag_{bs}, and the content φ being communicated. The sender will also annotate the belief $Bel_{rid}(\varphi)$ with a label γ that represents meta-information supporting that particular inference[iv], for example, it might represent the confidence of that agent on inferring $Bel_{rid}(\varphi)$, based on other information it has related to that inference, i.e., ag_{ToM} and ag_{bs}.

$$\frac{\begin{array}{c} T_\iota = i[head \leftarrow \texttt{.send}(rid, Tell, \varphi); h] \\ func\_send(\varphi, ag_{ToM}, ag_{bs}) = Bel_{rid}(\varphi)_{[\gamma]} \end{array}}{\langle ag, C, M, T, \mathsf{ExecInt}\rangle \longrightarrow \langle ag', C', M', T, \mathsf{ProcMsg}\rangle} \qquad (\textsc{SndTell})$$

$$\begin{array}{lll} \multicolumn{3}{l}{\textit{where:}} \\ M'_{Out} & = & M_{Out} \cup \{\langle mid, rid, Tell, \varphi\rangle\} \\ & & \text{with } mid \text{ a new message identifier;} \\ C'_I & = & (C_I \setminus \{T_\iota\}) \cup \{i[head \leftarrow h]\} \\ ag'_{ToM} & = & ag_{ToM} + Bel_{rid}(\varphi)_{[\gamma]} \\ C'_E & = & C_E \cup \{\langle +Bel_{rid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle\} \end{array}$$

In the rule SndTell, the agent updates its mail outbox M_{Out} with the message, it updates its current intention to $i[head \leftarrow h]$ (considering the action $\texttt{.send}(rid, Tell, \varphi)$ that has already been executed), then it updates its ToM with the prediction of a belief $Bel_{rid}(\varphi)_{[\gamma]}$, creating an event $\langle +Bel_{rid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle$ that may be treated in a later reasoning cycle, possibly forming a new goal for the agent based on this new information.

When a receiver agent receives a *Tell* message from an agent *sid*, it first checks whether the sender believes φ based on its previous ToM about the sender and the relevant information in its belief base. This expectation of a state of mind results

[iv] In § 4.6, I will present an instance for that meta-information, which agents might use to deal with the uncertainty present in the agents' communication.

from function *func_rec*, which takes as arguments the ToM that the receiver already has about the sender ag_{ToM}, the relevant beliefs in its belief base ag_{bs}, and the content φ being communicated, returning the sender's mental attitude the receiver can infer from those information, $Bel_{sid}(\varphi)$. A label γ is used to annotate relevant information such as the confidence on the projected state of mind.

$$\frac{\begin{array}{c} S_M(M_{In}) = \langle mid, sid, Tell, \varphi\rangle \\ func\_rec(\varphi, ag_{ToM}, ag_{bs}) = Bel_{sid}(\varphi)_{[\gamma]} \end{array}}{\langle ag, C, M, T, \mathsf{ProcMsg}\rangle \longrightarrow \langle ag', C', M', T, \mathsf{ExecInt}\rangle} \quad (\textsc{Tell})$$

$$\begin{array}{lll} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle mid, sid, Tell, \varphi\rangle\} \\ ag'_{bs} & = & ag_{bs} + \varphi[s(sid)] \\ ag'_{ToM} & = & ag_{ToM} + Bel_{sid}(\varphi)_{[\gamma]} \\ C'_E & = & C_E \cup \{\langle +\varphi[s(sid)], \mathsf{T}\rangle\} \cup \{\langle +Bel_{sid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle\} \end{array}$$

After that, the agent updates its mail inbox M_{In}, its belief base ag_{bs} with this new information $\varphi[s(sid)]$ (following the original semantics of AgentSpeak [284]), and it updates its ToM about the sender with $Bel_{sid}(\varphi)_{[\gamma]}$. Both of these updates (on the ToM and the belief base) generate events to which the agent is able to react.

Note that the predictions resulting from *func_send* and *func_rec* can be different from the actual state of mind of the other agents. Therefore, a good prediction model, considering both the ToM and relevant information from the agents' belief base, plays an important role when modelling ToM based on agent communication. Such models might consider the uncertainty present in agent communication, agents' autonomy and self interest, trust relations, reliability, etc. Thus, there are many different ways to instantiate such a model, and my approach allows different models to be implemented through the user-defined *func_send* and *func_rec* functions. A model for uncertain ToM will be presented in § 4.6. Therefore, I will omit γ in our examples until then.

Example 2 *Considering the scenario introduced in § 4.3.3, imagine that* John *meets his students every week in order to supervise their work. In a particular meeting with* Alice, Alice *asks* John *about the defi-*

nition of ToM, and John *answers* Alice *with the following message:* $\langle$alice, `tell`, `definition`(`tom`, *"an approach to model others' minds"*)$\rangle$. *At that moment,* John *is able to model that* Alice *believes the definition of Theory-of-Mind as "an approach to model others' minds", i.e.,* John *models* Bel_{Alice}(`definition`(`tom`, *"an approach to model others' minds"*)) *according to the* SndTell *semantic rule. Also, when* Alice *receives the message,* Alice *is able to model that* John *believes that definition for ToM, i.e.,* Alice *models* Bel_{John}(`definition`(`tom`, *"an approach to model others' minds"*)) *according to the* Tell *semantic rule.*

4.4.3 The Achieve Performative

The semantics of the *Achieve* performative are as follows. When a sender agent sends a message with the content φ, it expects that the receiver agent will likely desire φ. It can predict this result using its previous ToM about the receiver, ag_{ToM}, and the relevant information in its belief base, ag_{bs}, resulting in $Des_{rid}(\varphi)_{[\gamma]}$ (where again γ is an estimation of how likely the receiver is to adopt that goal).

$$\frac{\begin{array}{c} T_\iota = i[head \leftarrow \texttt{.send}(rid, Achieve, \varphi);h] \\ func\_send(\varphi, ag_{ToM}, ag_{bs}) = Des_{rid}(\varphi)_{[\gamma]} \end{array}}{\langle ag, C, M, T, \mathsf{ExecInt}\rangle \longrightarrow \langle ag', C', M', T, \mathsf{ProcMsg}\rangle} \qquad (\textsc{SndAchieve})$$

$$\begin{array}{lcl} \textit{where:} & & \\ M'_{Out} & = & M_{Out} \cup \{\langle mid, rid, Achieve, \varphi\rangle\} \\ & & \text{with } mid \text{ a new message identifier;} \\ C'_I & = & (C_I \setminus \{T_\iota\}) \cup \{i[head \leftarrow h]\} \\ ag'_{ToM} & = & ag_{ToM} + Des_{rid}(\varphi)_{[\gamma]} \\ C'_E & = & C_E \cup \{\langle +Des_{rid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle\} \end{array}$$

The sender agent updates its mail outbox M_{Out}, its current intention, its ToM about the receiver with the prediction $Des_{rid}(\varphi)_{[\gamma]}$, and an event is generated from the update in its ToM.

When an agent receives an *Achieve* message, it can safely conclude that the sender desires φ itself, using its previous ToM about the sender and the relevant information from its belief base.

$$\frac{\begin{array}{c} S_M(M_{In}) = \langle mid, sid, Achieve, \varphi\rangle \\ func\_rec(\varphi, ag_{ToM}, ag_{bs}) = Des_{sid}(\varphi)_{[\gamma]} \end{array}}{\langle ag, C, M, T, \mathsf{ProcMsg}\rangle \longrightarrow \langle ag', C', M', T, \mathsf{ExecInt}\rangle} \quad (\textsc{Achieve})$$

$$\begin{array}{lcl} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle mid, sid, Achieve, \varphi\rangle\} \\ ag'_{ToM} & = & ag_{ToM} + Des_{sid}(\varphi)_{[\gamma]} \\ C'_E & = & C_E \cup \{\langle +!\varphi, \mathsf{T}\rangle\} \cup \{\langle +Des_{sid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle\} \end{array}$$

The receiver agent updates its mail inbox M_{In} and its ToM about the sender, which generates an event $\langle +Des_{sid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle$. Also, another event $+!\varphi$ is generated, which allows the agent to react to that communication, searching for plans that allow it to pursue the achievement of φ. Using those plans, the agent is able to autonomously decide whether to achieve φ or not. In case it decides to achieve φ, then the agent will look for a plan that achieves φ and will make that plan one of its intentions.

Example 3 *Continuing our example, imagine that during a meeting with* Bob*,* John *realises that it could be interesting for* Bob *to read a paper about multi-agent systems, so* John *sends the following message to* Bob*:* $\langle \text{bob}, \texttt{achieve}, \texttt{read}(\texttt{bob}, \texttt{paper\_mas})\rangle$*. At that time,* John *is able to model that* Bob *desires to read the paper, i.e.,* $Des_{Bob}(\texttt{read}(\texttt{bob}, \texttt{paper\_mas}))$ *according to the* SndAchieve *semantic rule. Also,* Bob *is able to model that* John *desires that* Bob *reads the paper, i.e.,* $Des_{John}(\texttt{read}(\texttt{bob}, \texttt{paper\_mas}))$ *according to the* Achieve *semantic rule.* Bob *is able to react to the event* $+!\texttt{read}(\texttt{bob}, \texttt{paper\_mas})$*, searching for a plan to achieve that goal and turning the plan into one of* Bob's *intentions. A simple plan, written in Jason, that* Bob *could use to achieve this goal is shown below:*

```
+!read(Ag,Paper)
  :  .my_name(Ag) & desires(Sup,read(Ag,Paper)) & supervisor(Sup,Ag)
  <- read(Paper).
```

The plan above says that, when an event of the type ***+!read(Ag,Paper)*** *is generated, meaning that that agent has perceived that external event, in this* *case it has received an achieve message, of the type* **`read(Ag,Paper)`**, *then if* **`Ag`** *unifies with the name of the agent executing this plan (obtained with* **`.my_name(Ag)`***), and if the agent believes that its supervisor desires that it reads that paper (***`desires(Sup,read(Ag,Paper))`** *and* **`supervisor(Sup,Ag)`***), then the agent will proceed to execute the action* **`read(Paper)`**. *Note that the* Achieve *semantic rule provides the context (precondition) necessary for* Bob *to execute this plan, considering the unification* $\{\texttt{Ag} \mapsto \texttt{bob},\ \texttt{Paper} \mapsto \texttt{paper\_mas},\ \texttt{Sup} \mapsto \texttt{john}\}$ *and that* **`desires(john,read(bob,paper_mas))`** *is the code representation for* $Des_{John}(\texttt{read}(\texttt{bob}, \texttt{paper\_mas}))$.

4.4.4 The Ask-If Performative

When the sender agent sends an Ask-If message with the content φ, the only inference the agent can make is that the other agent will believe that the sender desires to know φ, i.e., $Bel_{rid}(Des_{ag}(\varphi))_{[\gamma]}$.

$$\frac{\begin{array}{c} T_\iota = i[head \leftarrow .\mathsf{send}(rid, AskIf, \varphi);h] \\ func\_send(\varphi, ag_{ToM}, ag_{bs}) = Bel_{rid}(Des_{ag}(\varphi))_{[\gamma]} \end{array}}{\langle ag, C, M, T, \mathsf{ExecInt}\rangle \longrightarrow \langle ag', C', M', T, \mathsf{ProcMsg}\rangle} \qquad (\textsc{SndAskIf})$$

$$\begin{array}{lcl} \multicolumn{3}{l}{\textit{where:}} \\ M'_{Out} & = & M_{Out} \cup \{\langle mid, rid, AskIf, \varphi\rangle\} \\ & & \text{with } mid \text{ a new message identifier;} \\ C'_I & = & (C_I \setminus \{T_\iota\}) \cup \{i[head \leftarrow h]\} \\ ag'_{ToM} & = & ag_{ToM} + Bel_{rid}(Des_{ag}(\varphi))_{[\gamma]} \\ C'_E & = & C_E \cup \{\langle +Bel_{rid}(Des_{ag}(\varphi))_{[\gamma]}, \mathsf{T}\rangle\} \end{array}$$

The sender agent updates its mail outbox M_{Out}, its current intention and its ToM about the receiver with the prediction $Bel_{rid}(Des_{ag}(\varphi))_{[\gamma]}$, thus an event is generated from the update in its ToM.

When an agent receives the message, it is able to infer that the sender desires to know whether φ is the case or not. After that, in both cases the agent updates its mental state similarly to the other semantic rules.

$$\frac{\begin{array}{c} S_M(M_{In}) = \langle mid, sid, AskIf, \varphi\rangle \\ func\_rec(\varphi, ag_{ToM}, ag_{bs}) = Des_{sid}(\varphi)_{[\gamma]} \end{array}}{\langle ag, C, M, T, \mathsf{ProcMsg}\rangle \longrightarrow \langle ag', C', M', T, \mathsf{ExecInt}\rangle} \quad (\textsc{AskIf})$$

$$\begin{array}{lll} \textit{where:} & & \\ M'_{In} & = & M_{In} \setminus \{\langle mid, sid, AskIf, \varphi\rangle\} \\ ag'_{ToM} & = & ag_{ToM} + Des_{sid}(\varphi)_{[\gamma]} \\ C'_E & = & C_E \cup \{\langle +Des_{sid}(\varphi)_{[\gamma]}, \mathsf{T}\rangle\} \end{array}$$

Example 4 *Continuing our scenario, imagine that during a group meeting,* John *asks all students if they like paper seminars, using the following message:* $\langle\{\text{bob, alice, nick, tom}\}, \texttt{AskIf}, \texttt{like(Ag, paper\_seminar)}\rangle$. *At that moment* John *considers that all students believe that* John *desires to know who likes paper seminars, e.g.,* $Bel_{Alice}(Des_{John}(\texttt{like(Ag, paper\_seminar)}))$, *according to the* SndAskIf *semantic rule. Also, all students think that* John *desires to know who likes paper seminars,* $Des_{John}(\texttt{like(Ag, paper\_seminar)})$, *according to the* AskIf *semantic rule. Two simple plans, written in Jason, that students could use to react the event generated by adding* $Des_{John}(\texttt{like(Ag, paper\_seminar)})$ *to their ToM is shown below:*

```
+!desires(Sup,like(Ag,Task))
  :  .my_name(Me) & like(Me,Task) & supervisor(Sup,Me)
  <- .send(Sup,tell,like(Me,Task)).


+!desires(Sup,like(Ag,Task))
  :  .my_name(Me) & ¬ like(Me,Task) & supervisor(Sup,Me)
  <- .send(Sup,tell, ¬ like(Me,Task)).
```

The plans above say that an agent will tell John *that it likes a particular task if it likes the task. Otherwise, an agent will tell* John *that it does not like that task. For example,* Alice *likes paper seminars, answering* John *with the following message:* $\langle\text{john}, \texttt{tell}, \texttt{like(alice, paper\_seminar)}\rangle$. *In this case,* John *will update its ToM stating that* Alice *likes paper seminars, and* Alice *will update its ToM stating that* John *believes that she likes paper seminars* $Bel_{john}(\texttt{like(alice, paper\_seminar)})$, *according to the* TELL *and* SNDTELL *semantic rules. In the future, as* John *has this information, it would be able to allocate a task to a student who likes that task.*

4.5 Reaching Shared Beliefs using ToM

Many application domains require the implementation of coordination schemes for MAS [150], in which agents need to work together in order to achieve the system's goal. As described in [150], acting together requires the team to be aware of and care about the status of the group effort as a whole. Thus, in order to maximise the cooperative performance in MAS, mechanisms that allow agents to reach shared beliefs should be incorporated.

Definition 1 (Shared Beliefs [284]) *Two agents ag_i and ag_j reach a state of shared beliefs when, for a belief $\varphi[S]$ where S represents the different sources of φ, both ag_i and ag_j are sources of φ, i.e., φ is a shared belief for ag_i and ag_j, when $\varphi[S'] \in ag_{i_{bs}}$ with $\texttt{source(self)},\ \texttt{source(ag}_\texttt{j}\texttt{)} \in \texttt{S}' \wedge \varphi[S''] \in ag_{j_{bs}}$ with $\texttt{source(self)},\ \texttt{source(ag}_\texttt{i}\texttt{)} \in \texttt{S}''$.*

Note that, considering the perspective of one agent only, that agent will reach a state of shared beliefs with another agent ag_j when, for a belief $\varphi[S]$ with S the different sources of φ, both *itself* and ag_j are sources of φ, i.e., $\texttt{source(self)}$, $\texttt{source(ag}_\texttt{j}\texttt{)} \in \texttt{S}$. That is an important consideration, because when modelling ToM, shared beliefs are going to be defined by the perspective of only one agent.

In [284], the authors showed how agents are able to reach shared beliefs. That approach for agents reaching shared beliefs starts with an agent ag_i, which believes in φ, sending to another agent ag_j a `tell` message with the content it desires to become a shared belief, i.e., $\langle ag_j, \texttt{tell}, \varphi\rangle$. Thus, following the semantics in [284], agent ag_j will receive the message and update its belief base with $\varphi[\texttt{source(ag}_\texttt{i}\texttt{)}]$. Then, agent ag_i needs to send a message to agent ag_j to achieve that shared belief, i.e., $\langle ag_j, \texttt{achieve}, \varphi\rangle$, thus the agent ag_j is able to execute the same procedure, sending a tell message to the agent ag_i with φ, i.e., $\langle ag_i, \texttt{tell}, \varphi\rangle$. Finally, agent ag_i receives this message and updates its belief base to $\varphi[\texttt{source}(\texttt{itself}), \texttt{source}(\texttt{ag}_\texttt{j})]$, reaching the state of shared beliefs.

Considering agents that are able to model ToM, it is possible to redefine the idea of shared beliefs, including the model of other agents' minds, i.e., a ToM. This is because agents able to model ToM will also update their mental attitudes/state when sending a message to other agents, as I formalised in § 4.4, and not only when receiving a message as in [284]. This gives:

Definition 2 (Shared Beliefs using ToM) *An agent ag_i will reach a state of shared beliefs with another agent ag_j when, for a belief φ, it is able to match its own belief φ with a ToM about ag_j believing φ, i.e., $\varphi \wedge Bel_{ag_j}(\varphi)_{[\gamma]}$, with γ equal to 1 (certain knowledge).*

When I assume that agents are cooperative, they trust each other, and the network infrastructure guarantees that messages will reach their intended receivers, I also assume that there is no uncertainty regarding the agents' ToMs of each other. Thus, it is easy to ensure that agents reach shared beliefs.

Proposition 1 (Reaching Shared beliefs — ToM without Uncertainty) *Without uncertainty of ToM, agents that are able to model ToM are able to reach a state of shared beliefs faster (with fewer exchanged messages) than agents without this ability.*

Proof 1 *Following the semantic rule SndTell, when an agent ag_i believes φ, thus $\varphi \in ag_{i_{bs}}$, and it is able to model ToM, thus $\exists$ ToM $\in ag$, then it is able to reach a state of a shared belief φ with another agent ag_j by communicating a single message $\langle ag_j, \texttt{tell}, \varphi\rangle$ to ag_j. When the agent ag_i sends this message, it updates its ToM* *with $Bel_{ag_j}(\varphi)$, according to the* SndTell *semantic rule, reaching the state of shared beliefs according to the Definition 2. So, we have (I) $\exists$ ToM $\in ag_i \;\wedge\; \varphi \in ag_{i_{bs}}$ then $\langle ag_j, \texttt{tell}, \varphi\rangle$. From (I) and SndTell we have (II) SndTell $\wedge$ $\langle ag_j, \texttt{tell}, \varphi\rangle$ then $ag_{i_{ToM}} \cup Bel_{ag_j}(\varphi)$. When ag_j receives that message, by the Tell semantics rule, updating its belief base with φ and its ToM with $Bel_{ag_i}(\varphi)$, also reaching the state of shared beliefs according to the Definition 2. Now we have from (I) and Tell the following (III) Tell $\wedge$ $\langle ag_j, \texttt{tell}, \varphi\rangle$ then $ag_{j_{ToM}} \cup Bel_{ag_i}(\varphi)$, and (IV) Tell $\wedge$ $\langle ag_j, \texttt{tell}, \varphi\rangle$ then $ag_{i_{bs}} \cup \varphi$. Therefore, from $(I) \wedge (II) \wedge (III) \wedge (IV)$ we have the following $\varphi \in ag_{i_{bs}} \wedge Bel_{ag_j}(\varphi)_{[\gamma]} \in ag_{i_{ToM}} \;\wedge\; \varphi \in ag_{j_{bs}} \;\wedge\; Bel_{ag_i}(\varphi)_{[\gamma]} \in ag_{j_{ToM}}$, which satisfies Definition 1. In contrast, agents that are not able to model ToM will need at least two messages, i.e., a* `tell` *message to be sent by each of them, according to the semantics from [284] and Definition 1.*

Example 5 *Following the scenario introduced in § 4.3.3, imagine that during the meetings* John *has had with his students, the students tell* John *which subjects they know more about, and* John *has the following information of his students, according to the* Tell *semantic rule:*

$$\left\{\begin{array}{ll}
\texttt{knows(alice,tom)} & \texttt{knows(bob,mas)} \\
\texttt{believes(alice,knows(alice,tom))} & \texttt{believes(bob,knows(bob,mas))} \\
\texttt{knows(nick,kr)} & \texttt{knows(ted)} \\
\texttt{believes(nick,knows(nick,kr))} & \texttt{believes(ted,knows(ted,[tom,mas]))}
\end{array}\right\}$$

Given this knowledge and the tasks John *wants to allocate to his students,* John *decides to assign the tasks as follows:* `task(write_paper,[mas,tom])` *to* Ted, *who knows about both subjects needed for completing that task,* `task(review_paper,[kr])` *to* Nick, *who is the only student able to execute that task, and grouping* Alice *and* Bob *for the task* `task(paper_seminar,[tom,mas])`. *If* Bob *only knows* `mas` *and* Alice *only knows* `tom`*, then they need to share their knowledge in order to successfully perform the task.*

Considering that together Alice *and* Bob *know both topics in order to help* *each other during the paper seminar, they decide to exchange knowledge about these topics. Thus, they might reach some shared beliefs (knowledge) about both topics. Note that, in this scenario,* Alice *and* Bob *assume that both are cooperating and both are rational. Thus,* Alice *starts the dialogue telling* Bob *that "Theory-of-Mind is an approach to model others' minds", i.e.,* ⟨bob, `tell`, `def`(`tom`, *"an approach to model others' minds"*)⟩. *At that moment, following the semantic rule* SNDTELL, Alice *updates its ToM with the following information* Bel_{bob}(***def(tom***, *"an approach to model others' minds"****)***). *When* Bob *receives this message, following the semantic rule* TELL, Bob *updates its belief base with the following information* ***def(tom***, *"an approach to model others' minds"****)***, *as well as its ToM about* Alice *with* Bel_{alice}(***def(tom***, *"an approach to model other minds"****)***). *By now, both* Alice *and* Bob *have reached a state of shared belief about the definition of* ***tom***, *according to Definition 2. They proceed sharing the relevant information about each topic until they both feel confident about both topics. Reaching shared beliefs (knowledge) is important for this particular task, in which, when the audience asks them questions about the topics* ***tom*** *and* ***mas***, *both Alice and Bob are able to answer the questions because they both have sufficient knowledge about the topics.*

4.6 Handling Uncertainty

Until now, both *Alice* and *Bob* were able to reach a state of shared belief about the definition of `tom`, according to Definition 2, which requires γ equal to 1 (certain knowledge). That is, I have introduced a mechanism to model uncertain ToM using the label γ, but up until this point I have made the assumptions that (i) agents are cooperative, (ii) they trust each other, and (iii) the network infrastructure guarantees that messages will reach their intended receivers, i.e., γ could be considered equal to 1 (certain knowledge).

However, agents usually operate under conditions of uncertainty in a MAS, and the previous assumptions are difficult to obtain; thus, agents will face uncertainty about their ToM, and consequently about their shared beliefs. For example, when an agent sends a message, it faces the uncertainty of the communication channel, i.e., the uncertainty of the message reaching the receiver. Also, when receiving a message, an agent faces the uncertainty of the truth of that statement, e.g., an agent is not able to verify if the other agents are acting maliciously [237, 200], thus it needs to consider the uncertainty of information it receives for those agents based on how much it trusts them [202, 203, 175].

One manner to overcome the uncertainty and to reach a more accurate ToM, following the literature on *common knowledge* [51], is increasing the communication between agents. Thus, an agent is able to increase the certainty of a given agent ag_j believing φ, confirming whether its ToM about agent ag_j believing φ is correct. That is, the agent is able to infer that ag_j believes φ by reinforcing this belief through communication. Henceforth I describe the model for uncertain ToM, which is compatible with that behaviour.

In order to show my approach, I am going to consider some parameter values. The first parameter, α, reflects the uncertainty of the communication channel when sending a message. The second parameter, β, reflects the uncertainty of the other agents telling the truth, i.e., when an agent ag_i tells φ to agent ag_j, agent ag_j is able to model that ag_i believes φ with a degree of certainty equal to β. For simplicity, I will assume that an agent will model its ToM about the other agents with a degree of certainty equal to the trust it has on the source[v], following the ideas introduced in [203, 202].

Definition 3 *The label γ will be instantiated with $\gamma = (\alpha, t)$ for an agent sending a message, and $\gamma = (\beta, t)$ for an agent receiving a message, where α represents*

[v] In [196], the authors show that trust aggregates not only the sincerity of the source but also the expertise the source has about the information communicated.

the uncertainty of the message reaching the target, β the uncertainty of the sender telling the truth, and t a discrete representation of the time of the MAS in which the message was exchanged.

Thus, following Definition 3, a trace of different updates on the ToM is constructed over time. Note that α and β reflect the uncertainty of an update at a given time. In order to execute reasoning over the ToM, agents are able to use the trace of these updates to calculate the degree of certainty of their model. Using this trace, I am able to model some desired behaviour from communication theory in agent communication, as I will describe later in this chapter.

For example, considering our scenario, when *Bob* tells *Alice* that "Theory-of-Mind is an approach to model others' minds", considering also that *Bob* knows that the reliability of the communication channel is 0.9, i.e., $\alpha = 0.9$, *Bob* will update its ToM, following the semantics for the `tell` performative (equation (4.4.2)) and Definition 3, with the information $Bel_{alice}(\texttt{def(tom}, \text{"an approach to model others' minds"}\texttt{))}_{[(0.9,t_i)]}$, with t_i the discrete time when the communication occurred. When *Alice* receives this message, considering that the trust *Alice* has on *Bob* telling the truth is 0.8, i.e., $\beta = 0.8$, and following the semantics for the `tell` performative (equation (4.4.2)) and Definition 3, *Alice* updates its ToM with $Bel_{bob}(\texttt{def(tom}, \text{"an approach to model other minds"}\texttt{))}_{[(0.8,t_j)]}$, with t_j the discrete time at which the message was received, with $t_i < t_j$. Both *Alice* and *Bob* model uncertainty of their ToM about each other believing on the definition of ToM.

Considering uncertain ToM, I need to redefine shared beliefs, in order to reflect the uncertainty of agents' models.

Definition 4 (Shared Beliefs using Uncertain ToM) *An agent ag_i will reach a state of shared beliefs with another agent ag_j when, for a belief φ, it is able to* *match its own belief φ with a ToM about ag_j believing φ with a predetermined degree of certainty χ, i.e., $\varphi \wedge P(Bel_{ag_j}(\varphi)) \geq \chi$, with χ a value describing the certainty necessary to consider φ a shared belief.*

Following the literature on common knowledge [51], if two individuals ag_i and ag_j can repeatedly communicate, then they can repeatedly reinforce their mental state regarding an item of information φ. For example, telling each other that φ is true, they should increase the certainty of each others' belief in φ. In order to model this desired behaviour in my approach, I maintain the trace of all updates an agent executes in its ToM, and using this trace I am able to aggregate different pieces of evidence in order to increase the certainty on ToM. There are many different ways to model this desired behaviour of agent communication if I was considering the particularities of each application domain. In our scenario, the information communicated by agents, e.g., a concept definition, does not change over time. Thus, for simplicity, I do not weigh every information according to the time it was received and the current time of the MAS. I only consider as pieces of evidence the number of times that information was communicated. For example, I count as evidence how many times does an agent perform a specific speech act. Every time that same agent performs the same speech act, I consider it as a piece of evidence for that agent believing the information that is being communicated through that speech act. Thus, I model this desired behaviour using the following equation:

$$P(Bel_{Ag}(\varphi)) = \begin{cases} f(Bel_{ag}(\varphi)) & \text{if } f(Bel_{ag}(\varphi)) <= 1 \\ 1 & \text{otherwise} \end{cases} \tag{4.1}$$

$$f(Bel_{ag}(\varphi)) = \frac{\sum\limits_{t_i \in \Delta T} v \mid Bel_{ag}(\varphi)_{[(v,t_i)]}}{|\Delta T|} + (\lambda \times |\Delta T|) \tag{4.2}$$

with ΔT the number of occurrences of $Bel_{ag}(\varphi)_{[(v,t_i)]}$ in the agent ToM, and λ the *evidence factor*, i.e., a parameter that reinforces the certainty on that information according to how often it occurs in the trace. Equation 4.2 calculates the average of the trace for $Bel_{ag}(\varphi)$ plus the evidence factor.

Example 6 *Following our scenario, imagine that* Bob *wants to reach a state of shared beliefs with Alice about the definition of ToM under the conditions of uncertainty described above. Thus, after sending the previous message and updating its ToM with* $Bel_{alice}(\texttt{def(tom}, \text{“an approach to model others' minds”}\texttt{))}_{[(0.9,t_i)]}$, Bob *has the option to increase the certainty in its ToM about* Alice *believing that definition by telling* Alice *that definition again. Taking* λ *to be* 0.1, *when* Bob *tells* Alice *the definition of ToM once again, following the semantics for the* `tell` *performative (equation in § (4.4.2)) and Definition 3,* Bob *adds* $Bel_{alice}(\texttt{def(tom}, \text{“an approach to model others' minds”}\texttt{))}_{[(0.9,t_j)]}$ *to its ToM, with* $t_i < t_j$. *Thus, Equation 4.1 returns 1, considering the average* $0.9 + 0.2$ *from the evidence factor, which is 0.1 multiplied by the number of evidences (equation (4.2)). Also, following the semantics for the* `tell` *performative (equation in § (4.4.2)) and Definition 3,* Alice *updates its ToM with* $Bel_{bob}(\texttt{def(tom}, \text{“an approach to model other minds”}\texttt{))}_{[(0.8,t_j)]}$, *and Equation (4.1) returns 1, considering the average* $0.8+0.2$ *from the evidence factor (equation (4.2)). Thus, they reach a state of shared belief about the definition of ToM*[vi], *considering* $\chi = 1$ *in Definition 4.*

Because Definition 4 allows agents to increase belief in a given piece of information by repeatedly communicating it, ag_i is able to reach a state of shared belief with another agent ag_j about a belief φ when it is able to infer $P(Bel_{ag_j}(\varphi)) \geq \chi$ from Equation 4.1 with $\chi = 1$, for example.

Proposition 2 (Reaching Shared Beliefs — ToM with Uncertainty) *When* λ *is a positive value, agents are able to eventually reach a state of shared*

[vi] When considering $\gamma = 0.1$ and α and $\beta >= 0.8$, agents are able to reach shared beliefs communicating only 2 messages.

beliefs, even considering $\chi = 1$*, provided they communicate the same information repeatedly. Also, the greater the value of* λ*, the faster agents will reach the state of shared beliefs.*

Proof 2 *According to Proposition 2,* λ *is a positive value,* $\lambda > 0$*. Let us consider that* α *and* β *are also positive values,* $\alpha, \beta > 0$*, since they are values that represent uncertainty. Even if* α *and* β *are not significant values, e.g.,* $\alpha, \beta \rightarrow 0^+$*, when we add to equation (4.2) the value computed by* $\lambda \times |\Delta T|$*, then the equation requires at most* n *communication instances of* $Bel_{ag_j}(\varphi)$ *for* $P(Bel_{ag_j}(\varphi))$ *to reach 1, with* $n = 1/\lambda$*. Let us suppose that* $\lambda = 0.1$*, then independently of the values taken* α *and* β*, it is necessary to have a trace with 10 communication instances of* $Bel_{ag_j}(\varphi)$ *in order to obtain* $P(Bel_{ag_j}(\varphi)) = 1$*. Also, the greater the value of* λ*, the faster the agents will be able to reach the state of shared beliefs, given that only* n *communication instances are necessary for* $P(Bel_{ag_j}(\varphi))$ *to reach 1, with* $n = 1/\lambda$*.*

Example 7 *Given the previous example from our scenario, where* Bob *wants to reach a state of shared beliefs with Alice about the definition of ToM under the conditions of uncertainty described above. Thus, after sending the previous message and updating its ToM with* $Bel_{alice}(\texttt{def(tom}, \text{“an approach to model others’ minds”}\texttt{)})_{[(0.9,t_i)]}$, Bob *also has the option to increase the certainty in its ToM about* Alice *believing that definition by asking* Alice *the definition of ToM and waiting for an answer from* Alice, *in which* Alice *tells* Bob *the definition of ToM. When* Bob *asks to* Alice *to tell him the definition of ToM, and waits for the answer. When* Alice *tells* Bob *the definition of ToM,* Alice *and* Bob *update their ToM with* $Bel_{bob}(\texttt{def(tom}, \text{“an approach to model other minds”}\texttt{)})_{[(0.9,t_j)]}$, $Bel_{alice}(\texttt{def(tom}, \text{“an approach to model others’ minds”}\texttt{)})_{[(0.8,t_i)]}$*, respectively. For both, Equation (4.1) returns 1, considering the average* $0.85 + 0.2$ *from the evidence* *factor, reaching a state of shared beliefs about the definition of ToM according to Definition 4 with $\chi = 1$.*

4.7 Making Decisions with ToM

Apart from enabling agents to model other agents' minds and allowing them to improve their models during communicative interactions, it is also essential that agents are able to make decisions using these models. Normally, a decision-making process is associated with the application domain, i.e., it is domain dependent. Therefore, I will present the decision-making process for the task assignment problem introduced in § 4.3.3.

In our scenario, during advising sessions, *John* asks students about different tasks they like to execute, as well as the different subjects the students are reading about (the subjects the students know about). Thus, *John* acquires ToM about the students, and its ToM becomes more accurate as they have more advising sessions, and consequently they communicate more with each other.

$$John_{ToM} = \left\{ \begin{array}{l} Bel_{alice}(\texttt{likes(paper\_seminar)})_{[0.8]} \\ Bel_{alice}(\texttt{likes(write\_paper)})_{[0.7]} \\ Bel_{bob}(\texttt{likes(review\_paper)})_{[0.9]} \\ Bel_{bob}(\texttt{likes(write\_paper)})_{[0.8]} \\ Bel_{nick}(\texttt{likes(review\_paper)})_{[0.6]} \\ Bel_{nick}(\texttt{likes(write\_paper)})_{[0.5]} \\ Bel_{ted}(\texttt{likes(write\_paper)})_{[0.8]} \\ Bel_{ted}(\texttt{likes(paper\_seminar)})_{[0.4]} \\ Bel_{ted}(\texttt{likes(review\_paper)})_{[0.6]} \end{array} \right\}$$

For example, *John* has asked (in different meetings and times) *Bob*, *Alice*, *Nick*, and *Ted* which academic tasks they like to execute, e.g., $\langle bob, AskIf, \texttt{likes(T)} \rangle$. After receiving this message, according to the semantic rule for the `ask` performative (equation in (4.4.4)), each student knows that *John* desires to know which task they like to execute. Based on this knowledge, each student has answered to *John* the tasks they like to execute, *John* has received these messages and updated its

ToM as shown in $John_{_{ToM}}$[vii].

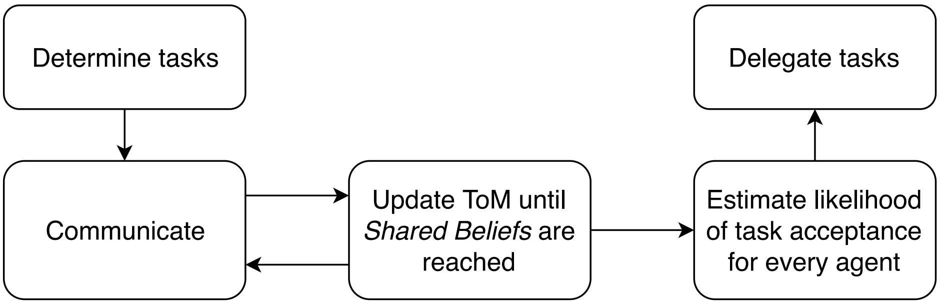

Determine tasks
Communicate
Update ToM until *Shared Beliefs* are reached
Estimate likelihood of task acceptance for every agent
Delegate tasks

Figure 4.1: Decision protocol for task delegation using ToM update rules through multi-agent communication.

Continuing with the example, during a meeting *Alice* asks *John* if there is any scheduled paper seminar about ToM and MAS, i.e., $\langle \texttt{john}, \mathit{AskIf}, \texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]\rangle$. Thus, based on the semantic rule for the `ask` performative (equation in (4.4.4)), *John* models that *Alice* is likely to desire that task, i.e., $\mathit{Des}_{alice}(\texttt{task}(\texttt{paper\_seminar}, [\texttt{tom}, \texttt{mas}]))_{[0.7]}$, answering positively. Also, imagine that *John* has asked the students which subject they have knowledge about, resulting in the following additional information to *John*'s ToM:

$$John_{_{ToM}} = \left\{ \begin{array}{ll} \mathit{Bel}_{alice}(\texttt{knows(tom)})_{[0.8]} & \mathit{Bel}_{bob}(\texttt{knows(mas)})_{[0.8]} \\ \mathit{Bel}_{alice}(\texttt{knows(mas)})_{[0.9]} & \mathit{Bel}_{bob}(\texttt{knows(kr)})_{[0.9]} \\ \mathit{Bel}_{nick}(\texttt{knows(kr)})_{[0.8]} & \mathit{Bel}_{ted}(\texttt{knows(tom)})_{[0.8]} \\ \mathit{Bel}_{nick}(\texttt{knows(mas)})_{[0.7]} & \mathit{Bel}_{ted}(\texttt{knows(kr)})_{[0.5]} \\ \mathit{Bel}_{nick}(\texttt{knows(tom)})_{[0.8]} & \mathit{Bel}_{ted}(\texttt{knows(mas)})_{[0.8]} \end{array} \right\}$$

Using its ToM, *John* executes the probabilistic reasoning described in § 4.3.3, which computes the likelihood for each student to accept each task as shown in Table 4.1. Note that the likelihood of *Alice* accepting the task `paper_seminar` is based on the information $\mathit{Des}_{alice}(\texttt{task(paper\_seminar)}\ [\texttt{tom}, \texttt{mas}])_{[0.7]}$ in *John*'s ToM, while the other results are based on the likelihood of the students liking a particular task and knowing the subjects related to that task. Thus, *John* concludes that it is possible to increase the probability of each task to be accepted by

[vii] I do not represent the time at which the messages were communicated, but since they were communicated at different times I introduced different values for γ.

the students by offering the task `task(paper_seminar, [tom, mas])` to *Alice*, offering `task(review_paper, [kr])` to *Bob*, and offering `task(write_paper, [mas, tom])` to *Ted*.

Student	**Task**	**Likelihood**
Alice	`task(write_paper, [mas, tom])`	0.5
Alice	`task(review_paper, [kr])`	0.0
Alice	`task(paper_seminar, [tom, mas])`	**0.7**
Bob	`task(write_paper, [mas, tom])`	0.0
Bob	`task(review_paper, [kr])`	**0.8**
Bob	`task(paper_seminar, [tom, mas])`	0.0
Nick	`task(write_paper, [mas, tom])`	0.3
Nick	`task(review_paper, [kr])`	**0.5**
Nick	`task(paper_seminar, [tom, mas])`	0.0
Ted	`task(write_paper, [mas, tom])`	**0.5**
Ted	`task(review_paper, [kr])`	0.2
Ted	`task(paper_seminar, [tom, mas])`	0.1

Table 4.1: Likelihood calculation for task assignment

4.8 Deriving New Knowledge

While modelling other agents' minds through communication is a fundamental step towards equipping agents with ToM, it only represents one approach for agents to acquire the model of other agents' minds, i.e., TT ToM (*Theory Theory-of-Mind*). A different approach is the ability of agents to infer new knowledge about other agent's mind without explicitly communicating. This approach requires agents to consider their context and TT ToM to infer extended knowledge, or in other use to simulate the minds of other agents, i.e., to use an ST ToM (*Simulation Theory-of-Mind*).

In multi-agent systems, the context can be perceived by agents from their environment and organisation. For example, (i) in the JaCaMo framework [28], agents that belong to the same organisation, can perceive the other agents' roles, responsibilities, relationship (acquaintance, authority, and communication links), etc; and (ii) in Electronic Institutions (EI) [190], agents can perceive other agents' roles, group meetings (scenes), as well as the normative rules of the system, etc.

If an agent can access this type of information about other agents and their roles and relationships as well the rules of action and communication that different contexts impose, then the agents can use this information to simulate how the mental attitudes of the other agents change in different scenarios.

Therefore, by considering (i) its belief base, (ii) the other agent's model given by TT ToM, and (iii) a particular context, an agent can derive knowledge, i.e., to execute ST ToM. Thus, I introduce a function $\varepsilon_{\rightarrow}$, which considers the agent's belief base, bs, a particular context, Cx, and the agent's ToM in order to model an extended ToM, including the extended knowledge it was able to infer from its knowledge in that particular context.

$$\varepsilon_{\rightarrow}(bs, Cx, ToM) : bs \wedge Cx \wedge ToM \longrightarrow ToM' \tag{4.3}$$

Note that, this differs from § 4.4, where an agent infers TT ToM using *func_send* and *func_rec*, that considers the agent's belief base, ag_{bs}, the agent's previous model of ToM, ag_{ToM}, and the speech-act used. In this section, an agent derives knowledge from the model it has already built from those communications. This is also different from § 4.7, in which an agent is able to make decisions using that TT ToM, whereas in this section the agent is able to use the extended knowledge when making decisions, i.e., using ST ToM too.

Example 8 *Imagine that John models that Alice knows the subject ToM, i.e.,* $Bel_{alice}(knows(tom)) \in John_{ToM}$. *Also, imagine that they are in the context of a group meeting, and that John has learned that students who know a subject and are participating in a group meeting normally desire to talk about the known subject, i.e., the inference rule* $(Bel_{ag}(knows(S)) \wedge group\_meeting \rightarrow Des_{ag}(talk\_about(S))) \in John_{bs}$. *Thus, applying* $\varepsilon_{\rightarrow}$, *John is able to infer that Alice desires to talk about ToM during the group meeting,* $Des_{alice}(talk\_about(tom))$, *considering a substitution function* $\{ag \mapsto alice,\ S \mapsto tom\}$.

In Jason, such inference rules are implemented as follows:

$$\texttt{desires}(\texttt{Ag}, \texttt{talk\_about}(\texttt{Sub})) :- \texttt{believes}(\texttt{Ag}, \texttt{knows}(\texttt{Sub})) \ \& \ \texttt{group\_meeting}.$$

Note that, an agent is only able to infer $Des_{alice}(talk\_about(tom)))$ while it believes itself to be in that particular context, i.e., `group_meeting`.

Definition 5 (Extended Theory-of-Mind) *An extended Theory-of-Mind is a ToM, acquired from agent's communication, extended through a process of reasoning in which an agent considers its own knowledge, context, and its ToM to model other agents' mental attitudes. The new models hold while the context considered also holds.*

It is reasonable to think that an agent could use the performatives presented in § 4.4 to directly reach some desired mental attitude of other agents, e.g., reaching $Des_{alice}(talk\_about(tom))$ by executing $\langle alice, \texttt{achieve}, \texttt{talk\_about(tom)}\rangle$ according to the `SndAchieve` semantics rule. However, it is not always the case that the communication protocol will allow such moves during dialogue, e.g., it could restrict such moves from a subordinated role to a role with higher authority. Also, in communication, indirectly making an audience reach a conclusion is known to be more efficient than directly disclosing the conclusion to them [187]. Thus, extended ToM, as I have introduced it here, is an essential component to be considered when developing socially-intelligent agents equipped with ToM. Furthermore, it is essential for agents to be able to reason about which action to execute in order to indirectly reach some desired state of ToM, e.g., to reason about how they can cause changes in the mental attitudes of other agents. To do this, they should be able to simulate other agents' minds.

Definition 6 (Desired Theory-of-Mind) *We say that an agent has a desired Theory-of-Mind, ToM^*, when an agent desires another agent to have a certain mental attitude $\psi \in ToM^*$, where $\psi \notin ToM$ (the current agent's ToM).*

An agent is able to simulate other agents' minds looking for a proposition that, when added to its ToM, it will make the agent reach a desired state for its model of other agents' minds. That allows an agent to reason about which state of other agents' mind it might need to reach in order to achieve its goal. I introduce the function $\varepsilon_{\leftarrow}$ which considers the agent's belief base, bs, a particular context, Cx, and a simulated ToM, ToM^+, in order to model an extended ToM considering a simulated ToM, ToM'.

$$\varepsilon_{\leftarrow}(bs, Cx, ToM^+) : bs \wedge Cx \wedge ToM^+ \longrightarrow ToM' \tag{4.4}$$

Definition 7 (Simulated Theory-of-Mind) *Let ToM be a Theory-of-Mind. A simulated Theory-of-Mind, ToM$^+$, is a ToM extended with a set of propositions that an agent does not have in its current ToM, i.e., ToM$^+$ $= ToM \;\cup\; \Delta^+$ with $\Delta^+ = ToM^+ \setminus ToM$.*

Example 9 *Let us use a different scenario in which agents are not allowed to use the achieve performative to model other agents' desires. For example, imagine that a car dealer wants to sell a car, and they know that the customers will desire to buy the car if they believe the car is safe and fast, thus the car dealer knows that feeding the information* ***safe(car1)*** *and* ***fast(car1)*** *to a customer will make the customer believe that* ***car1*** *is safe and fast, thus inferring that the customer will desire to buy* ***car1*** *[200], i.e., $(Bel_{ag}(fast(Car)) \wedge Bel_{ag}(safe(Car)) \wedge at(ag, sales\_hall) \rightarrow Des_{ag}(buy(Car))) \in Car\_Dealer_{bs}$.*

Therefore, the car dealer is able to infer that feeding a customer who is at the sales hall (i.e., $Cx = \texttt{at(customer, sales\_hall)}$) with ***safe(car1)*** *and* ***fast(car1)****, i.e., becoming able to model that* $Bel_{customer}(fast(car1))$ *and* $Bel_{customer}(safe(car1))$*, it will reach a desired ToM in which the customer desires to buy that car, i.e.,* $Des_{customer}(buy(car1))$ *applying the function* $\varepsilon_{\leftarrow}$*, considering that* $(Bel_{ag}(fast(Car)) \wedge Bel_{ag}(safe(Car)) \wedge at(ag, SRom) \rightarrow Des_{ag}(buy(Car))) \in Car\_Dealer_{bs}$ *and an unification function* $\{ag \mapsto customer, Car \mapsto car1\}$.

Note that $Car\_Dealer_{ToM+} = Car\_Dealer_{ToM} \cup \{Bel_{customer}(fast(car1)), Bel_{customer}(safe(car1))\}$. *That means the car dealer has simulated a ToM with the given information in order to check if the other agent reaches a certain mental attitude that the dealer agent has in its desired state of ToM. From that simulation, the car dealer is able to understand what information it needs to model about other agents in order to reach the desired state of ToM, in our example,* $\{Bel_{customer}(fast(car1)), Bel_{customer}(safe(car1))\}$. *Based on my semantics, the car dealer could use the Tell performative,* $\langle \text{customer}, \mathtt{tell}, \mathtt{safe(car1)} \rangle$ *and* $\langle \text{customer}, \mathtt{tell}, \mathtt{fast(car1)} \rangle$*, respectively, to reach a desired state of ToM in which the car dealer can model* $Des_{customer}(buy(car1))$.

The rule used in Example 9 can be implemented in Jason as follows:

```
desires(Ag, buy(Car)) : - believes(Ag, safe(Car)) & believes(Ag, fast(Car))
                          & at(Ag, sales_hall)
```

Definition 8 (Effective Simulation of Theory-of-Mind) *We say that an agent is effective in its simulation of ToM when it is able to reach a new Theory-of-Mind from a previous Theory-of-Mind executing a set of actions, in which this new Theory-of-Mind corresponds to its simulated Theory-of-Mind from Definition 7. That means, using a set of action Act, an agent is able to reach* ToM^{+} *from* ToM*, i.e.,* $ToM \longrightarrow_{Act} ToM''$*, with* $ToM'' \equiv ToM^{+}$.

Proposition 3 (Reaching Desired Theory-of-Mind) *If an agent is effective in its simulation of Theory-of-Mind, following Definition 8, it is able to reach a desired extended Theory-of-Mind ToM^* when in the right context, following Definition 5.*

Proof 3 *Consider that an agent ag desires to reach ToM^* from Definition 6. Reaching ToM^* can depend on the context, and the rules enabled by such context, or not. If the agent is effective in its simulation of Theory-of-Mind, according Definition 7, then it is able to start from a Theory-of-Mind ToM, simulate ToM^+, and according to Definition 6, verify if the resulting extended ToM', according to Definition 5, matches with ToM^*, that is if $ToM' \equiv ToM^*$. (i) When ToM^* does not depend on the context, then ag will add to its simulated ToM, ToM^+, those models of the other agents' minds that are missing from its desired ToM, ToM^*; (ii) When ToM^* depends on the context, assuming that the agent is in the right context, Cx, it will add to its simulated ToM, ToM^+, those models of the other agents' minds that are missing from its desired ToM, ToM^*, that it can reach by executing the actions that are allowed by the context Cx plus those models of the other agents' minds that, together with the context Cx allow it to reach an extended ToM, ToM', that matches with its desired Theory-of-Mind ToM^*. Finally, assuming that ag is effective in its simulation of Theory-of-Mind, then it will be able to execute the actions that will allow it to reach state where $ToM' \equiv ToM^*$.*

The reasoning executed by an agent is showed in Figure 4.2. First an agent will define which state of ToM it desires to reach (which could match with its goals). Second, it will simulate what it needs to add to its current ToM in order to reach that desired ToM. Third, it will execute those actions that will add those mental models that are missing in its current ToM in order to either (i) directly reach the desired ToM; or (ii) together with its context reach an extended ToM which matches its desired ToM. After, executing those action and updating its ToM, it is expected to reach the desired ToM.

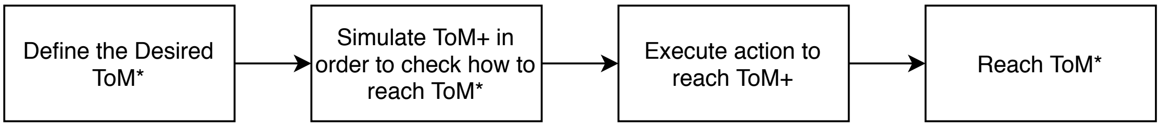

Define the Desired ToM*
Simulate ToM+ in order to check how to reach ToM*
Execute action to reach ToM+
Reach ToM*

Figure 4.2: Reaching a Desired ToM.

4.9 Limitations of the Model

When modelling ToM, the beliefs of other agents are context independent, whereas their Intentions and Desires are normally context dependent. An agent's intentions as well as desires or goals need to be instantiated for every context. In dialogue games, for instance, where each dialogue game represents a context, one needs to specify to the agents what the rules of the game are as well as the possible moves that are permitted in the game, represented by "speech acts", i.e., a communication protocol. Although my approach could potentially be used to model other agents' strategies in such protocols, I did not explore this topic in this chapter. Note that it requires us to consider a particular *context* of a dialogue game, in which agents will perceive the rules applying to the context (dialogue rules, communicative actions, etc.) and execute ST ToM over other agents' model using that context.

In our running example, for instance, actions that can be performed by the agents in that context are specified by the agents' organisation, so in the context of *group_meeting* the action *presentation* was possible. The approach presented here mainly focuses on agent-agent communication, although combining it with existing work on generating and extracting arguments and speech acts in/from natural language, would also support work on human-agent interaction. However, integrating the approach here with natural language processing is not in the scope of this thesis.

4.10 Conclusion

The approach to model Artificial ToM in this chapter represents an important stepping-stone towards the modelling and implementation of deceptive agents in MAS. This BDI-based approach to ToM enables agents to model all three types of mental attitudes, namely beliefs, desires, and intentions, when reasoning about the minds of other agents. This approach shows 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 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. Finally, the capability of agents to derive new knowledge about the minds of others through mental simulation gives agents the ability to reason about different contexts in which communication happens.

The approach in this chapter offers an explainable way to model agents with ToM. This is a crucial property of ToM to have if we want to study deceptive agents that use ToM. The operational semantics, for instance, let us track every belief agents form about other agents in different contexts. This is very important for checking what the agents are thinking about each other and to see why they communicate certain messages and why they intend to communicate those messages in the first place. Later in this thesis I illustrate how agents that have already acquired beliefs about other agents' minds can use these prior beliefs (which belong to their TT ToM) to simulate the minds of their targets. That is to use information acquired through communication in order to check if it is possible to cause the formation of a desired false belief in the mind of their target, and thus to attempt deception.

Chapter 5

Dishonest Agents

In this chapter I model and compare three types of dishonest behaviour, including deception, that future socially-aware artificial agents may adopt.

It is reasonable to assume that in the next few decades, intelligent machines might become much more proficient at socialising. This implies that the AI community will face the challenges of identifying, understanding, and dealing with the different types of social behaviours these intelligent machines could exhibit. Given these potential challenges, in this chapter I describe how to model three of the most studied strategic social behaviours that could be adopted by autonomous and malicious software agents. These are dishonest behaviours such as lying, bullshitting, and deceiving that autonomous agents might exhibit by taking advantage of their own reasoning and communicative capabilities. In contrast to other studies on dishonest behaviours of autonomous agents [258, 231, 130, 54, 217], I use an agent-oriented programming language to model dishonest agents' attitudes and to simulate social interactions between agents. The model of dishonest behaviours described in this chapter is intended to be used by further research through simulation, in order to study and propose mechanisms that identify and deal with dishonest software agents.

5.1 Introduction

Agent-Oriented Programming Languages (AOPL) and platforms to develop Multi-Agent Systems (MAS) provide suitable frameworks for modelling agent communication in AI. We can reasonably say that one of the main purposes of AI research is to represent as accurately as possible the way humans use information to perform actions. Actions of humans are sometimes performed by applying dishonest forms of reasoning and behaviour such as lying, bullshitting, and deceiving.

In this chapter, I model lies, bullshit and deception in an AOPL named Jason [33], 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 [45].

Even though the AI community has investigated computational models of lies [258], "bullshit", and deception [40], to the best of my knowledge, the work in this chapter 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.

The study in this chapter brings two main contributions to the AI community: (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 car[i]. 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.

[i] The implementation of this work is available at `https://tinyurl.com/ybrmkqg9`.

5.2 Background

5.2.1 Lying, Bullshitting, and Deceiving

I will start by describing what lying is from an agent-based perspective. I define lying similar to [40]. A lie is a false statement about something that is intended to make someone believe the opposite of what is actually true. Lying cannot be reduced to linguistic communication only. Liars give out information to others in various forms, such as social behaviour, facial expressions, physiological responses to questions, and manipulation of the environment [75, 39].

Definition 9 (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 ψ is the case.*

Bullshit is different from lying in the sense that it is not intended to make someone believe the opposite from the truth. A bullshitting agent will give an answer to a question in such a way that the one who asked the question is left with the impression that the bullshiter agent knows the true answer [92], when in fact it does not.

Definition 10 (Bullshit) *The dishonest behaviour of an agent Ag_i to tell another agent Ag_j that ψ is the case, when in fact Ag_i does not know if ψ is the case.*

Deception is more complex than bullshit or lying. I define deception as the intention of an agent (Deceiver) to make another agent (Interrogator) believe something is true that it (Deceiver) thinks is false, with the scope of reaching an ulterior goal or desire. The complexity arises due to the fact that an agent requires *Theory-of-Mind* (henceforth ToM) to deceive [130]. ToM is not needed to tell a lie or to bullshit (although there are cases in which liars or bullshitters can make use of ToM). The Deceiver has to let the Interrogator reach the conclusion by itself. For example, if the Deceiver wants the Interrogator to believe that q is the case, instead of directly telling the Interrogator that q is the case, the Deceiver uses some knowledge that the Interrogator possesses, let's say $p \rightarrow q$, and tells the Interrogator that p is the case. Having told the Interrogator that p is the case, the Deceiver then knows that if the Interrogator is a rational agent that has the ability to apply *Modus Ponens*, then it will conclude that q is the case. Levine and McCornack call this interplay *Pars Pro Toto* (the information the Deceiver decides to feed the Interrogator) and *Totum Ex Parte* (the knowledge the Interrogator derives from the information sent by the Deceiver) [172].

One can argue that liars and bullshitters might have some types of motivations or goals. However, compared to deceivers, these goals do not contain ulterior motives. A liar, for example, can have the goal to speak falsely about a state of the world without taking into consideration the state of mind of the agent it speaks to. It can also be argued that a good liar would take into account its target's mind, although by definition a liar is constrained by one single strategy which is to speak falsely about a state of the world. A bullshitter can have the goal to make the agent it speaks to believe it (the bullshitter) is speaking the truth independently of the state of the world it is speaking about. Most of the times, however, bullshitters do not take into consideration the target's mental activity in order to deliver a bullshit.

Definition 11 (Deception) *The* ***intended*** *dishonest behaviour of an agent Ag_i to tell another agent Ag_j that ψ is the case, when in fact Ag_i knows that $\neg\psi$ is the case, in order to make Ag_j conclude that φ given that Ag_i knows that Ag_j knows that $\psi \rightarrow \varphi$ and Ag_i also knows that Ag_j is rational.*

5.2.2 Jason AOPL and its Dishonesty-enabling Properties

Among the many AOPL and platforms mentioned in previous chapters, such as Jason, Jadex, Jack, AgentFactory, 2APL, GOAL, Golog, and MetateM, as discussed in [32], I chose the Jason platform [33] for this work. In this way, I continued to build on the work I have done on the Theory-of-Mind. As you should probably know by now (by having read the previous chapters), Jason extends the AgentSpeak language, an abstract logic-based AOPL introduced by Rao [214], which is one of the best-known languages inspired by the BDI architecture. In Jason, the agents are equipped with a library of pre-compiled plans that have the following syntax:

```
triggering_event :  context <- body.
```

where the `triggering_event` represents the way agents react to events, for example, a new goal for the agent to pursue, or a new belief in case the plan is to be triggered by reaction to perceived changes in the world; the `context` has the preconditions for the plan to be deemed applicable for achieving that goal given the current circumstances, and the `body` is a sequence of actions and sub-goals to achieve the goal.

Besides specifying agents with well-defined mental attitudes based on the BDI architecture, the Jason platform [33] 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, e.g., `safe(car1)`; (ii) believes to be false, e.g., ¬`safe(car1)`; (iii) is ignorant about, i.e., the agent has no information about whether the car is safe or not. 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). The annotation has the following format: `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, for example trust and time as used in [175, 197]. 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 [33]. Further, new performatives can be easily defined (or redefined) in order to give special meaning to them[ii].

5.3 Running Example

To show the difference between the agents' attitudes of telling a lie, telling bullshit and deceiving, I will present an approach to model these three agent attitudes in an agent-oriented programming language using a running example of a car dealer scenario[iii], inspired by [170, 181, 258]. In our scenario, an agent called *car dealer*, *cd* for short, has the desire to sell as many cars as it can. Thus, the car dealer will use all its available strategies, including lying, bullshitting, and attempting to deceive the customers to buy the cars it has for sale.

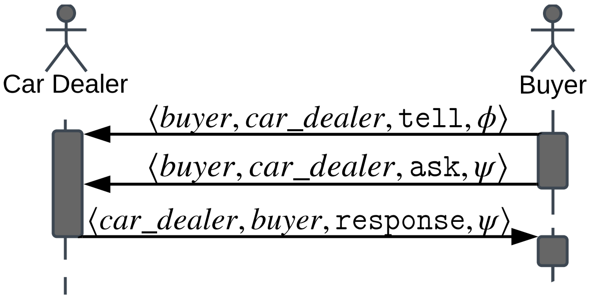

Car Dealer
Buyer
$\langle buyer, car\_dealer, \texttt{tell}, \phi \rangle$
$\langle buyer, car\_dealer, \texttt{ask}, \psi \rangle$
$\langle car\_dealer, buyer, \texttt{response}, \psi \rangle$

Figure 5.1: Protocol.

An illustration of the communication protocol for our scenario is shown in Figure 5.1. The protocol states that: a *buyer* agent will tell to another agent, the *car dealer*, the set ϕ of characteristics they desire in buying a car. For example, ϕ = `inference(buy(car),[safe(car),comfortable(car)])`, means that the *buyer* considers safety and comfort to be the most desirable characteristics for buying a

[ii] For example, [198] proposes new performatives for argumentation-based communication between Jason agents.

[iii] I do not assume that in real life car dealers are deceptive agents, I just use this particular scenario as an illustrative example.

car. After that, *buyers* ask the *car dealer* about the cars they have an interest to buy. The *car dealer* answers the questions based on its own interest (i.e., it is a self-interested agent).

In this scenario, I will focus on characteristics of cars such as: safety, speed, comfort, and storage size which are defined in Δ_{car_dealer}: [iv]

$$\Delta_{car\_dealer} = \left\{ \begin{array}{ll} \texttt{safe(ford)} & \neg\texttt{comfortable(ford)} \\ \texttt{safe(bmw)} & \neg\texttt{comfortable(bmw)} \\ \neg\texttt{safe(renault)} & \neg\texttt{comfortable(renault)} \\ \neg\texttt{fast(ford)} & \texttt{large\_storage(bmw)} \\ \texttt{fast(bmw)} & \neg\texttt{large\_storage(renault)} \end{array} \right\}$$

Here, there are two important considerations for our model. The first consideration is about the diversity of information the *car dealer* knows, which is fundamental when simulating the agents' behaviours. I set up the scenario with different cars, in which each car has different characteristics. The second consideration is that the *car dealer* may be able to model the *buyers*' mental state, which means the *car dealer* is able to model the characteristics *buyers* consider important to buy a car, i.e., ϕ. Given the knowledge of the characteristics the *buyers* consider important, the *car dealer* is able to simulate the influence of the information it provides, choosing the best answer according to its own interest or desire, i.e., to sell the cars. Thus, an agent may be able not only to model the initial states of other agents minds, but also to simulate how the minds of these other agents change over time[v].

5.4 Modelling Buyers' Minds

In this chapter, I set up the notation based on the Jason agent-oriented programming language [33] and a standard representation for messaging. This model will consist of predicates which represent the mental state of agents, the event model which is

[iv] Note that, in Δ_{car_dealer}, the car dealer knows that the `ford` is safe, that the `renault` is not comfortable, that the `bmw` is fast, and it is ignorant about whether the `renault` is fast or not.

[v] These abilities of our agents reflect their capacity of using both Theory-Theory-of-Mind and Simulation-Theory-of-Mind for modelling the minds of their targets [101].

the set of perceptions and possible messages that the agents can communicate such as asking and answering questions, the belief update rules for each kind of message and perception, and inference rules that allow agents to execute belief update and reasoning simulation.

5.4.1 Modelling the Minds of Other Agents

Agents will model others agents' minds according to inferences they are able to make, that are based on the perceptions they have of the target agents and the communication they have with the target. These ideas come from studies in Theory-of-Mind (ToM) [102]. Thus, based on the BDI architecture, I use the following predicates to allow an agent to model the other agents' minds:

- `believes(`*ag*`,prop)` means that an agent *ag* believes proposition `prop`. For example, `believes(john,safe(ford))` means that `john` believes that `ford` are safe. A *car dealer* agent *cd* is able to model the beliefs of a *buyer* agent *ag* either inferring it from some information it has previously received from *ag*, based on the belief annotation in Jason, using the inference rule[vi] :
  `believes(ag,prop):- prop(source(ag))`

  after receiving a `tell` message from *ag*, i.e., $\langle ag, cd, \texttt{tell}, \texttt{prop}\rangle$. I use $\Delta_{cd} \models$ `believes(`*ag*`,prop)` to describe that the car dealer *cd* knows that the buyer *ag* believes on `prop`. A particular case for this predicate is $\texttt{believes}(ag, \texttt{inference}(\texttt{prop}, \texttt{S}))$ representing that an agent *ag* believes on the inference from `S` (a set of predicates) to `prop`. For example, `believes(john,inference(buy(bmw),[safe(bmw)])` means that `john` believes that if a `bmw` is `safe` it could buy a `bmw`.
- `desires(`*ag*`,prop)` means that an agent *ag* desires `prop`. For example,

[vi] Note that an agent is able to infer its own beliefs using this inference rule, i.e., *cd* is able to infer that it believes on prop itself if `prop(source(self))` is in its belief base.

desires(john,buy(bmw)) means that john desires to buy a bmw. I use $\Delta_{cd} \models$ desires(*ag*,prop) to describe that the car dealer *cd* knows that the buyer *ag* desires prop.

Nested representations for beliefs and desires are also possible. For example, it is possible to express that the *car dealer cd* believes that the *buyer ag* desires to buy a car, i.e., believes(*cd*,desires(*ag*,buy(_))), which is the same $\Delta_{cd} \models$ desires(*ag*,buy(_)).[vii] As another example, the car dealer is able to model the buyer Theory-of-Mind about itself, i.e., $\Delta_{cd} \models$ believes(*ag*,believes(*cd*,safe(renault)).

5.4.2 Modelling Agents' Actions and Communication Updates

Agents will update their ToM about others when communicating with them, as well as when perceiving them in the environment. For simplicity, in this chapter I will only consider a few communicative actions, based on the protocol described in § 5.3. Thus, the possible actions and belief updates of the agents are the following:

- $\langle ag, cd, \texttt{tell}, \texttt{prop}\rangle$ means a message sent by the agent *ag* to the agent *cd*, with the performative tell, and the content prop. When *cd* receives this message, it executes the following update in its ToM:

$$\Delta_{cd} = \Delta_{cd} \cup \texttt{believes}(ag, \texttt{prop})$$

- $\langle ag, cd, \texttt{ask}, \texttt{prop}\rangle$ means a message sent by the agent *ag* to the agent *cd*, with the performative ask and the content prop. When *cd* receives this message, it executes the following update in its ToM:

[vii] To investigate different levels of ToM in multi-agent systems is out of the scope of this chapter, thus I use only first-order ToM, i.e., I do not model ToM about others' ToM, and ToM about others' ToM about others' ToM, and so forth.

$$\Delta_{cd} = \Delta_{cd} \cup \texttt{desires}(ag, \texttt{prop})$$

- $\langle cd, ag, \texttt{response}, \texttt{prop}\rangle$ means a message sent by the agent *cd* to the agent *ag*, with the performative `response` and the content `prop`. To execute this action, it requires that a previous message $\langle ag, cd, \texttt{ask}, \texttt{prop}\rangle$ has been communicated. Thus, the agents *cd* and *ag* execute the following updates in their ToM and knowledge base, respectively:

$$\Delta_{cd} = \Delta_{cd} \cup \texttt{believes}(ag, \texttt{prop})$$

$$\Delta_{ag} = \Delta_{ag} \cup \texttt{prop}[\texttt{source}(cd)]$$

Note that the semantics for a `response` message is different from the `tell` message, given that a `tell` message expresses the opinion of the sender, and the `response` message represents an information previously requested, which means it represents a desired update the receiver wants to execute in its knowledge base.

Finally, an agent also is able to update its ToM perceiving other individuals that are situated in the same environment. In this work, the *car dealer* is able to perceive the *buyers* when they enter in the sale room, i.e., an event (perception) of the type `+client(`*ag*`)` is generated by the environment, enabling the *car dealer* to infer that the *buyer ag* desires to buy a car. The *car dealer cd*'s ToM is updated as follows:

$$\Delta_{cd} = \Delta_{cd} \cup \texttt{desires}(ag, \texttt{buy}(\_))$$

Note that, while the perceptions from the environment are domain dependent, the communication semantics are independent of the domain. This is because the meaning of the performatives guides the way in which an agent executes its belief updates. That is, for different environments, the agents' perceptions from the environment may have different meanings, and by extension beliefs will be updated in different ways.

5.4.3 Making Inferences from the Models of Other Agents' Minds

It is important to model when an agent is *ignorant* about the truth of a proposition. That is, considering multi-agent systems that model a *open* world, when an agent does not know if ϕ is true, that does not mean that ϕ is false, i.e., when an agent cannot infer either ϕ or $\neg\phi$, the only conclusion it may reach is that it is *ignorant* about the truth of ϕ. An agent is able to infer that it is ignorant about a proposition using the following inference rule:

`ignorant_about(Prop) :- not(Prop) & not(¬Prop).`

Similarly, an agent is able to infer that it is ignorant about other agents' mental states, using the following inference rules:

```
ignorant_about(believes(Ag,Prop)) :-
not(believes(Ag,Prop)) & not(¬believes(Ag,Prop))
ignorant_about(desires(Ag,Prop)) :-
not(desires(Ag,Prop)) & not(¬desires(Ag,Prop)).
```

Furthermore, an agent is able to infer new information about other agents' mental state from the information it already has in its ToM. For example, if the *car dealer* agent *cd* knows that the *buyer* agent *ag* believes that `ford` are `safe`, i.e., `believes(`*ag*`,safe(ford))`, and that *ag* also believes in the inference that safe cars are good options to buy, i.e., `believes(`*ag*`,inference(buy(X),[safe(X)]))`, *cd* is able to infer that *ag* also believes that the `ford` is a good option to buy, i.e., `believes(`*ag*`,buy(ford))`. This reasoning using ToM is implemented using the following inference rule:

`believes(Ag,C) :- believes(Ag,inference(C,P)) & believes(Ag,P).`

The *car dealer* will not know the beliefs of the *buyers* about each car in advance. An interesting way for the *car dealer* to gain this knowledge is for the *car dealer* to be able to simulate the conclusions a *buyer* might reach based on the information the *car dealer* provides and the inferences the *buyer* is able to execute:

```
    implies(believes(Ag,N),believes(Ag,C)) :-
believes(Ag,inference(C,N)).
```

Thus, if the *car dealer cd* knows that the *buyer ag* believes that safe cars are a good option to buy, i.e., $\texttt{believes(}\boldsymbol{ag}\texttt{,inference(buy(X),[safe(X)]))}$, then *cd* also knows in advance that *ag* will believe that `ford` are good options to buy, i.e., $\texttt{believes(}\boldsymbol{ag}\texttt{,buy(ford))}$, if and only if *cd* provides *ag* the information that `ford` are safe, i.e., $\texttt{believes(}\boldsymbol{ag}\texttt{,safe(ford))}$.

5.5 Modelling Lies in AOPL

Using this approach, it is possible to model a lie following the scenario of when the *car dealer cd* knows that $\neg\psi$ (ψ is not true), but it responds either ψ or $\texttt{ignorant\_about(}\boldsymbol{cd}\texttt{,}\psi\texttt{)}$ to *buyer ag*.

Table 5.1: Conditions for a Lie.

Car Dealer (*cd*)
Beliefs: $\neg\psi$
Actions: $\langle cd, ag, \texttt{response}, \psi\rangle$
ToM: $\texttt{desires(}\boldsymbol{ag}\texttt{,}\psi\texttt{)}$

Buyer (*ag*)
Beliefs: $\texttt{ignorant\_about(}\psi\texttt{)}$
Desires: ψ
Actions: $\langle ag, cd, \texttt{ask}, \psi\rangle$

As described, a liar could tell lies without any particular goal, but the most common situation requires some motivation that makes an agent tell a lie, in order to achieve a particularly desired state of the world and/or a state of mind. I will discuss this motivation further in this chapter. For now let's assume that the a *buyer ag* asks the *car dealer* if `renault` are safe, i.e., $\langle ag, cd, \texttt{ask}, \texttt{safe(renault)}\rangle$. In this case, based on *cd*'s knowledge base represented in Δ_{car_dealer}, *cd* has two options: either telling the truth, i.e., $\langle cd, ag, \texttt{response}, \neg\texttt{safe(renault)}\rangle$, or telling a lie, i.e., either $\langle cd, ag, \texttt{response}, \texttt{ignorant\_about(}\boldsymbol{cd}\texttt{,safe(renault))}\rangle$ or $\langle cd, ag, \texttt{response}, \texttt{safe(renault)}\rangle$.

5.6 Modelling Bullshit in AOPL

Using this approach, it is possible to model bullshit based on the scenario of when the *car dealer cd* is ignorant about ψ, i.e., $\mathtt{ignorant\_about}(\psi)$, but it responds either ψ or $\neg\psi$ to the *buyer ag*.

Table 5.2: Conditions for Bullshit.

Car Dealer (*cd*)
Beliefs: $\mathtt{ignorant\_about}(\psi)$
Actions: $\langle cd, ag, \mathtt{response}, \psi\rangle$
ToM: $\mathtt{desires}(\boldsymbol{ag},\psi)$

Buyer (*ag*)
Beliefs: $\mathtt{ignorant\_about}(\psi)$
Desires: ψ
Actions: $\langle ag, cd, \mathtt{ask}, \psi\rangle$

Similarly to a liar, a bullshiter could tell bullshit without a particular goal, but the most common situation requires some motivation, as I will discuss further in this chapter. For now, let us assume that the *buyer ag* asks to the *car dealer* if $\mathtt{renault}$ are fast, i.e., $\langle ag, cd, \mathtt{ask}, \mathtt{fast(renault)}\rangle$. In this case, based on *cd*'s knowledge base represented in Δ_{car_dealer}, *cd* has two options: either telling the truth, i.e., $\langle cd, ag, \mathtt{response}, \mathtt{ignorant\_about(fast(renault))}\rangle$, or telling a bullshit, i.e., either $\langle cd, ag, \mathtt{response}, \mathtt{fast(renault)}\rangle$ or $\langle cd, ag, \mathtt{response}, \neg\mathtt{fast(renault)}\rangle$.

5.7 Modelling Deception in AOPL

One question that arises from Sections 5.5 and 5.6 is: how does the *car dealer cd* decide what to answer? For example, How does it choose between lying by telling ψ or lying by telling $\mathtt{ignorant\_about}(\boldsymbol{cd},\psi)$, when it knows $\neg\psi$ is true? I argue that the answer for that question is the *motivation* or *ulterior goal* of the *car dealer cd*. In this particular piece of work, I model deception using the motivation of the *car dealer cd* of making the *buyers* to buy a car which is not suitable for the *buyers* according to the buyers' requirements communicated in the first interaction of our protocol.

There are two major reasons I consider the scenario in which car dealers are deceivers. The first reason is because car dealers usually have an ulterior goal, that is to sell cars. This goal is related to both the state of the world (usually the properties of the car the dealer is trying to sell) and to the mind of the target. The dealer needs to take into account the preferences and attitudes (considered by us as beliefs of the target) in order to provide the information that will make the target believe it should buy the car. The second reason is because car dealers do not care if the target agent believes they (the car dealing agents) know the truth about the state of the world (or state of the car in this particular case). Their ulterior goal is not to make the buyer believe they have true knowledge about the car (as a bullshitter would want the buyer to believe). The car dealer's goal is to make the buyer reach the conclusion that it (the buyer) should buy the car by itself. In order to make the buyer reach that particular conclusion, the dealer needs to feed the buyer a set of particular pieces of information (true or false).

Table 5.3: Conditions for Deception.

Car Dealer (cd)
Beliefs: $\neg\psi$ Desires: $\texttt{believes(}\boldsymbol{ag}\texttt{,}\varphi\texttt{)}$ Actions: $\langle cd, ag, \texttt{response}, \psi\rangle$ ToM: $\texttt{believes(}\boldsymbol{ag}\texttt{,inference(}\varphi\texttt{,}\psi\texttt{))}$, $\texttt{desires(}\boldsymbol{ag}\texttt{,}\psi\texttt{)}$

Buyer (ag)
Beliefs: $\texttt{believes(inference(}\varphi\texttt{,}$ $\psi\texttt{))}$, $\texttt{ignorant\_about(}\psi\texttt{)}$ Desires: ψ Actions: $\langle ag, cd, \texttt{tell}, \texttt{inference(}\varphi\texttt{,}\psi\texttt{)}\rangle$, $\langle ag, cd, \texttt{ask}, \psi\rangle$

Imagine that a *buyer ag* starts a dialogue with the *car dealer cd* by telling *cd* that it considers safety and speed to be the most important characteristics when buying a car, i.e., $\langle ag, cd, \texttt{tell}, \texttt{inference(buy(X),[fast(X),safe(X)])}\rangle$. When *ag* asks *cd* if renault are safe, i.e., $\langle ag, cd, \texttt{ask}, \texttt{safe(renault)}\rangle$, it makes *cd* model that $\texttt{desires(}\boldsymbol{ag}\texttt{,safe(renault))}$. Thus, *cd* satisfies the precondition necessary for deceiving *ag* (see *cd*'s ToM in Table 5.3). Imagine also that *cd*'s desire is for *buyers* to believe that they should buy the car *cd* is selling.

Then, the agent *cd* models that a buyer *ag* considers safety and speed the essential characteristics to buy a car, and that *ag* desires to know if `renault` are safe, i.e., *cd* models `believes(`*`ag`*`,inference(buy(X,[safe(X),fast(X)]))` and `desires(`*`ag`*`,safe(renault))` in its ToM. What follows from this is that now, *cd* is able to infer that if it gives a positive answer `safe(renault)`, then this will determine *ag* to believe `buy(renault)`. Therefore, *cd* decides to send the message $\langle cd, ag,$ `response`, `safe(renault)`$\rangle$, lying about `safe(renault)`. What happens next is that *ag* asks if `renault` are fast, i.e., $\langle ag, cd,$ `ask`, `fast(renault)`$\rangle$. Again, *cd* executes the same reasoning process as before. Therefore, *cd* will answer $\langle cd, ag,$ `response`, `fast(renault)`$\rangle$, telling bullshit about `fast(renault)`. In the final step, *cd* is able to conclude that it has managed to deceive *ag* because *cd* is able to model in its ToM that `believes(`*`ag`*`,safe(renault))`, `believes(`*`ag`*`,fast(renault))` and `believes(`*`ag`*`,inference(buy(X),[safe(X),fast(X)]))`. This allows *cd* to conclude [viii] `believes(`*`ag`*`,buy(renault))` that corresponds to *cd*'s ulterior goal.

5.8 Conclusion

In this work, I described a representation for modelling and simulating other agents' minds using an AOPL based on the work in Chapter 4. Furthermore, using the proposed representation, I have described a model for three of the most studied dishonest attitudes in AI literature, i.e., *lying*, *bullshitting* and *deceiving.* In particular, I have described how to model and implement these attitudes in Jason [33].

Modelling and implementing such attitudes in an AOPL allows us to investigate agents' dishonest behaviours through simulations in a high-level, declarative approach. On one hand, in this particular piece of work, I have used a *car dealer* scenario, which, given its simplicity, enables the focus on the main contribution of

[viii] This scenario corresponds to the one of `buyer1` in our implementation.

this chapter, i.e., the representation of other agents' minds and the modelling and simulation of lies, bullshit and deception in MAS. On the other hand, the approach is generic and can be easily used to model and simulate other scenarios of dishonest agent behaviour.

Chapter 6

Deceptive Agents

In this chapter I model deception using Theory-of-Mind between an agent that deceives and its target.

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 chapter, I describe a high-level approach for modelling machine deception using ToM under factors of uncertainty and I propose an implementation of this model in an Agent-Oriented Programming Language (AOPL). I show that the Multi-Agent Systems (MAS) paradigm can be used to integrate concepts from two major theories of deception, namely *Truth-Default Theory* (TDT) *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, I define an epistemic agent mechanism using BDI-like architectures to analyse deceptive interactions between deceivers and their potential targets and I also explain the steps in which the model can be implemented in an AOPL. To the best of my knowledge, this work is one of the first attempts in AI that (i) uses ToM along with components of TDT, IMT2 and IDT in order to analyse deceptive interactions and (ii) implements such a model.

6.1 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...' [272, p. 434]. I 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 [43]. 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 skeptical attitude towards AI. I 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 *skeptical* attitudes of agents.

I consider that this 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 I 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[i]. 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 three major theories of deception, with the help of AI methods, namely *Truth-Default-Theory* (TDT), *Interpersonal Deception Theory* (IDT) and *Information Manipulation Theory 2* (IMT2) (see Chapter 2.1). TDT is the theory that explains how different dialogical contexts influences the success of deception, e.g., how communicating the same information can trigger the interlocutors into either truth-biased or sceptical mental attitudes [153]. IDT is the theory that explains how the communicative skills and cognitive load of individuals affect deceptive interactions [38], whereas IMT2 is the theory that explains how individuals employ deception by manipulating information [172]. Even though there are many studies in the AI literature that look at deception, none of them uses TDT, IMT2, IDT and ToM together to model interactions that involve deceptive artificial agents.

6.2 Background

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,

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

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 [102]. 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' [15]. 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 [110]. [130] 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.

According to TDT, which I have described in §2.1.3, ToM is necessary for deception to function [153, loc. 2422]. In deception, it matters what the target agent knows and what the deceiver agent thinks the target knows. In deception detection, it matters how we can infer others' mental states, including deceptive motives, and people usually assign deceptive motives based on their ToM of others.

As I mentioned before in §2.1.5, 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 [121], 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, which I have described in §2.1.4, 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.

6.3 Modelling Deception

I consider *deception* to be different from *lying* and from *bullshitting* as I previously defined them in the thesis. I derive a new definition of deception from the one of deception in previous chapter. Hence, I **define** *deception* as:

Definition 12 (Deception) *The intentional process of a deceptive agent, which I name Donald, to make another interrogator agent, which I 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, I make use of ToM by combining TT with ST using the approach I have described in Chapter4. That is, 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.

I 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 13 (Agents) *Ag represents an agent. When I need to make the distinction between two agents I 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 14 (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 15 (Actions) *I 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 16 (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 17 (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 ψ .*

Definition 18 (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, I 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 19 (Theory-Theory (TT)) *The prior beliefs that an agent Ag_i has of the beliefs of another agent Ag_j.*

Definition 20 (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 21 (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 22 (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.*

Definition 23 (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.*

6.3.1 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 preconditions that follow from Definitions 12, 13, 14, 16 and 17. I 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 its missing knowledge to start a dialogue. Donald, for instance, could provide an information that Ivan never thought about finding out in the first place, and by finding out that information, Ivan could infer a belief about something else that is false. I 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 [51] with Donald given its TT ToM of Donald as agents manage to reach in Chapter 4.

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

I 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 chapter.

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

I 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 Definition 12).

6.3.2 Parameters

I assume that the agents, Donald and Ivan, are constrained by two parameters from IDT, namely trust and communicative skill. I 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. I add this parameter as the *communicative skill* of Donald and label it β.

On top of the parameters from IDT, I also add a degree of confidence γ that Donald has in its TT of Ivan. This is important because I 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. I add a success estimation parameter and label it with θ.

I do not provide a model for computing α, β, and γ, because that would change the focus of the work. The scope of this work 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 [99].

6.3.3 Aggregating Parameters

I 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 I 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. I use the following definitions to show the computation of the interaction between *trust*, *communicative skill*, *confidence in ToM*, and *estimation of success*:

Definition 24 ($P(\alpha)$) *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 25 ($P(\alpha, \beta)$) *Donald's estimation of Ivan's trust α in Donald is conditionally dependent on 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 26 ($P(\theta)$) *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 $P(\gamma)$; $P(\theta) = P(\alpha, \beta) * P(\gamma)$.*

6.3.4 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, I add profiles to the agents. For now, I limit the profiles to *reckless* and *cautious* for Donald, and *credulous* and *skeptical* 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.

- **Skeptical** 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 skeptical interrogator believes that there is always a good reason to distrust others.

Reasoning Processes:

- **Simulate ToM** (see Definition 20 & Algorithm 1) is the reasoning process used by Donald to see what beliefs will be reached by Ivan given some information provided by Donald. I assume that Donald already has a TT (see Definition 19) 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 Definition 18). Given Ivan's newly formed belief on Donald'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 Definition 21 & 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 Definition 23). **Pars Pro Toto** uses **Simulate ToM** as a sub-process 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 successful deception, therefore Donald proceeds to check the estimation of success (see Definition 26). 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 Definition 22 & 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 Definition 17) about some information and Ivan has a TT (see Definition 19) 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.

Table 6.1: Agents with ToM, Labels and Profiles

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

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                  */
$DeceiverAction \leftarrow \bot$;
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 $DeceiverAction = \bot$ then
    $DeceiverAction \leftarrow$ random $action$ from $Actions$ such that
     SimulateToM($action, \varphi, ToM$) $=$ False;
```

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 6.2: 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 $B_{Dec}(B_{Int}(\psi \to \varphi))$ 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 $B_{Dec}(A_{Dec}(\psi) \to B_{Int}(\psi))$ from Trust Rule and KK 5 $B_{Dec}(B_{Int}(\psi))$ from 3, 4 and MP 6 $B_{Dec}(B_{Int}(\psi) \wedge B_{Int}(\psi \to \varphi))$ 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 $B_{Dec}(A_{Dec}(\neg\psi) \to B_{Int}(\neg\psi))$ from Trust Rule and KK 5.1 $B_{Dec}(B_{Int}(\neg\psi))$ from 3.1, 4.1 and MP 6.1 $B_{Dec}(\neg B_{Int}(\neg\psi \to \varphi))$ 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 $\mathrm{P}(\theta)$ given $\mathrm{P}(\alpha, \beta) \wedge \mathrm{P}(\gamma)$.
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 $B_{Dec}((B_{Int}(\varphi) \wedge D_{Dec}(B_{Int}(\varphi))) \to B_{Dec}(A_{Dec}(\psi)))$ from 8 and BI 10 $B_{Dec}(A_{Dec}(\psi))$ from 8, 9 and MP with estimated $\mathrm{P}(\theta)$
Donald will update its planned answer $A_{Dec}(\psi)$ or $A_{Dec}(\neg\psi)$ based on 10, its profile and $\mathrm{P}(\theta)$
11A $B_{Dec}(A_{Dec}(\psi))$ from 10, profile and $\mathrm{P}(\theta)$ 11B $B_{Dec}(A_{Dec}(\neg\psi))$ from 10, profile and $\mathrm{P}(\theta)$
12A $A_{Dec}(\psi)$ from 11A and RA 12B $A_{Dec}(\neg\psi)$ from 11B and RA

Table 6.3: Ivan executes the second part of *Totum Ex Parte* after receiving Donald'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 6.2 (12 A)
2 $B_{Int}(\psi)$ from 1, Trust Rule and MP
3 $B_{Int}(\psi \to \varphi)$ from Beliefs
4A $B_{Int}(\psi)$ from 1, Trust Rule, profile and $\mathbf{P}(\alpha)$
4B $B_{Int}(\neg\psi)$ from 1, Trust Rule, profile and $\mathbf{P}(\alpha)$
5A $B_{Int}(\varphi)$ from 3, 4 A and MP
5B $\neg B_{Int}(\varphi)$ from 3, 4 B and NAF

6.4 Evaluation and Results

In this section I will present an evaluation of the model. In 6.4.1 I will go through a step-by-step deceptive play to see how the beliefs of agents evolve during a game. Then, in 6.4.2 I 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 I will discuss the results.

Donald might estimate its success (see Table 6.2) given its knowledge about Ivan (see Table 6.1). However, its estimation might not be precise due to a possible strong influence on Ivan by its profile (*credulous* or *skeptical*) and the real degree of trust α (see Table 6.3).

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 6.1, I 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 6.2 and 6.3. I proceed to show a run of a deceptive play using our model.

Another important observation is that given the way I set up the knowledge bases and possible actions of the two agents Donald and Ivan in Table 6.1 (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.

6.4.1 Running a Deceptive Play

Setup: *cautious* Donald and *skeptical* 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 6.4). I 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

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

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 *skeptical* $B_{Int}(\neg\psi)$ will be updated to $B_{Int}(\psi)$.

18. $B_{Int}(\psi)$ from 17, $P(\alpha)$ and *skeptical*

19. $B_{Int}(\psi \rightarrow \varphi)$ from Beliefs of Ivan

20. $B_{Int}(\varphi)$ from 18, 19 and MP

6.4.2 Results & Analysis

Furthermore, considering the intervals of values established by the profiles introduced, I have analysed in a similar manner all the possible outcomes of deceptive plays between Donald and Ivan. The results are presented in Tables 6.4-6.7. The results we refer to in the text are highlighted in the tables with an asterisk $*$.

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 highlighted in red in Tables 6.4-6.7). These are: **Table 6.4** where $P(\alpha) = [0, .75)$ and $P(\theta) = [.75, 1]$; **Table 6.5** where $P(\alpha) = [0, .25)$ and $P(\theta) = [.75, 1]$; and **Table 6.7** where $P(\alpha) = [0, .25)$ and $P(\theta) = [.25, 1]$. Figure 6.1 shows the possible outcomes for our model, given the influence of the parameters α, β and γ.

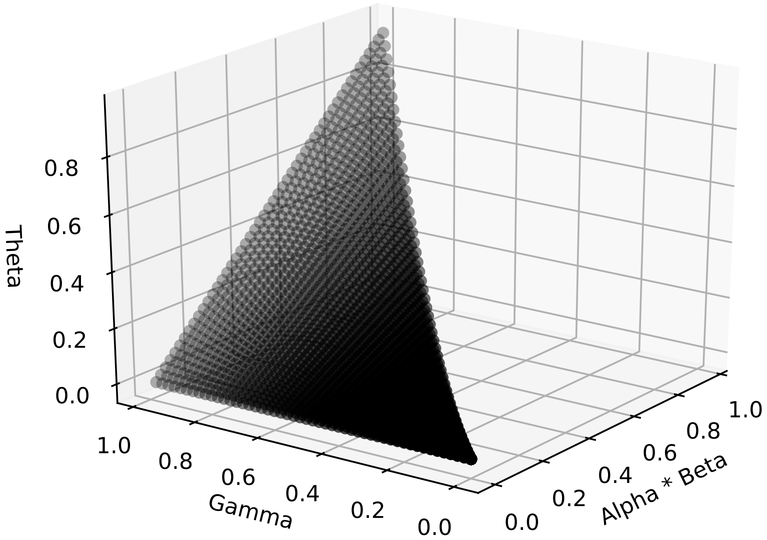


Figure 6.1: The influence of the parameters α (Alpha), β (Beta) and γ (Gamma) on θ (Theta).

Unintended Deception: The most interesting results are: (i) in **Table 6.4** where Donald is *cautious* and Ivan is *skeptical*, $P(\alpha) = [0, .75]$ and $P(\theta) = [0, .75]$;

and (ii) in **Table 6.6** where Donald is *reckless* and Ivan is *skeptical*, $P(\alpha) = [0, .75)$ and $P(\theta) = [0, .25)$. In (i) a *cautious* Donald meets a *skeptical* 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 *skeptical* 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.

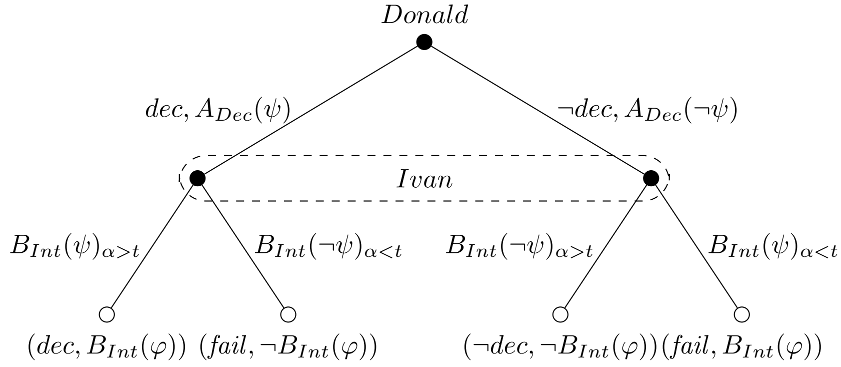

$Donald$
$dec, A_{Dec}(\psi)$
$\neg dec, A_{Dec}(\neg\psi)$
$Ivan$
$B_{Int}(\psi)_{\alpha>t}$
$B_{Int}(\neg\psi)_{\alpha<t}$
$B_{Int}(\neg\psi)_{\alpha>t}$
$B_{Int}(\psi)_{\alpha<t}$
$(dec, B_{Int}(\varphi))$
$(fail, \neg B_{Int}(\varphi))$
$(\neg dec, \neg B_{Int}(\varphi))$
$(fail, B_{Int}(\varphi))$

Figure 6.2: 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 *skeptical* agents can indirectly act as deceptive agents themselves under certain circumstances. Hence, it can be argued that agents that are biased towards skepticism 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 *skeptical* 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 *skeptical* interrogators to be ultimately deceived requires an agent architecture with an even higher-order of ToM than I 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 6.4: Cautious vs Skeptical

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 6.5: 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 6.6: Reckless vs Skeptical

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 6.7: 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

6.5 Implementation in AOPL

Before I was able to successfully implement the model in an AOPL, I first needed to find a way in which to represent ToM in an AOPL. The reason I had to do this is because my model of deception relies on agents that have ToM as a cognitive property. To model agents that model other minds, I adopted the approach described in Chapter 4. First, I explain why I chose Jason as an AOPL to implement the model and describe the approach I used for modelling ToM in Jason using its predicates i.e., the representation of TT ToM. After that, I describe how agents execute meta-reasoning using the TT modelling, i.e., I describe how I implement ST as a meta-reasoning mechanism in agent-oriented programming. Finally, I describe how agents use TT and ST to reason about and simulate the other agents' mind in order to make decisions. In particular, I will show the decision-making process for deception, introduced in § 6.3.

I 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.

6.5.1 Agent Oriented Programming Languages

Among the many AOPL and platforms discussed in [32], such as Jason, Jadex, Jack, AgentFactory, 2APL, GOAL, Golog, and MetateM, I chose the Jason platform [33] for this work. Jason extends the AgentSpeak language, an abstract logic-based AOPL introduced by Rao [214]. Jason [33] 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: `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 [197]. 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.

6.5.2 Modelling Theory-Theory ToM in AOPL

An important aspect to be considered in order to model ToM in Jason [33] is that one needs not only to represent when an agent believes that another agent believes something (which can be inferred from the belief annotations in Jason), but one also needs to represent when an agent believes that another agent does not believe something, or when it is ignorant about that[ii]. Therefore, I 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., $\Pi_b^a \models B_b(p)$.
- `believes(b,¬p)`: meaning that agent a believes that an agent `b` believes `¬p`, i.e., $\Pi_b^a \models B_b(\neg p)$.
- `¬believes(b,p)`: meaning that agent a believes that an agent `b` does *not* believe `p`, i.e., $\Pi_b^a \models \neg B_b(p)$.
- `¬believes(b,¬p)`: meaning that agent a believes that an agent `b` does not believe `¬p`, i.e., $\Pi_b^a \models \neg B_b(\neg p)$.
- `believes(b,inference(q,p))`: meaning that agent a believes that an agent `b` is able to infer `q` from `p`, i.e., $\Pi_b^a \models B_b(p \rightarrow q)$.

Jason automatically annotates all information that an agent has perceived/received with the appropriated source from where that information came. Using

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

this annotation, I am able to implement some meta-reasoning that allows an agent to make inferences[iii] from its private knowledge to its ToM. The inference rule $\texttt{believes}(\texttt{Ag}, \texttt{Prop}) : -\texttt{Prop}[\texttt{source}(\texttt{Ag})]$ 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 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:

$\texttt{ignorant\_about}(\texttt{Ag}, \texttt{Prop}) : - \neg\texttt{believes}(\texttt{Ag}, \texttt{Prop}) \& \neg\texttt{believes}(\texttt{Ag}, \neg\texttt{Prop})$

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., $(\Delta^a \not\models p \wedge \Delta^a \not\models \neg p)$ using the following inference rule: $\texttt{ignorant\_about}(\texttt{Prop}) : - \texttt{not}(\texttt{Prop}) \& \texttt{not}(\neg\texttt{Prop})$. Finally, an agent a is able to infer when it is ignorant about other agents' beliefs, i.e., considering another agent b we have $(\Pi^a_b \not\models B_b(p) \wedge \Pi^a_b \not\models \neg B_b(p))$:

$\texttt{ignorant\_about}(\texttt{believes}(\texttt{Ag}, \texttt{Prop})) : -$

$\texttt{not}(\texttt{believes}(\texttt{Ag}, \texttt{Prop})) \& \ \texttt{not}(\neg\texttt{believes}(\texttt{Ag}, \texttt{Prop}))$.

[iii] These inferences are characterised as ST ToM, as I describe in next section.

6.5.3 Reasoning Using Simulation 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, I consider both perspectives.

Based on the approach for representing ToM in Jason agents, I 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:

$\texttt{believes(Ag,C)} :- \texttt{believes(Ag,inference(C,P))}\ \&\ \texttt{believes(Ag,P)}$

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., $((\Pi_b^a \models B_b(p \rightarrow q) \wedge \Pi_b^a \models B_b(p)) \rightarrow \Pi_b^a \models B_b(q))$.

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., $\Pi_b^a \models B_b(q)$. Considering that the current state of a's ToM only indicates that agent b believes p implies q, i.e., $\Pi_b^a \models B_b(p \rightarrow q)$, using ST an agent is able to infer that it needs to achieve $\Pi_b^a \models B_b(p)$, thus with $(\Pi_b^a \models B_b(p \rightarrow q) \wedge \Pi_b^a \models B_b(p))$ it is able to achieve the desired ToM state $\Pi_b^a \models B_b(q)$. This backward-reasoning can also be observed in Table 6.2. Agents execute such reasoning using the following inference rule: $\texttt{implies(believes(Ag,N),believes(Ag,C))} :- \texttt{believes(Ag,inference(C,N))}$; 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[iv].

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, I 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, I will discuss how agents make decisions based on the scenario in § 6.3.

6.5.4 Decision Making and Communication Semantics

In this section I 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.). I give formal semantics to both speech acts used for modelling deception in multi-agent systems, namely Ask and Response. I define new semantic rules to accompany the existing operational semantics of Jason [33, 284]; however, for clarity I use only the configuration components that I need to formalise the essentials of our approach. Also, to account for the decision-making process, I define two functions, `Conf`() and `Trust`(), which describe different behaviours that agents may adopt depending on the parameters α, β, and γ introduced in § 6.3, 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.

[iv] For simplicity, I 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, I am 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 `desires(a, believes(a, q))` at TT ToM, (ii) that an agent does not believe something while it has a desire for that, i.e., ¬`believes(Ag, Prop)`:- `desires(Ag, believes(Ag, Prop))`.

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

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

Ask1 says that when an agent selects a received message to be processed[v] $\langle mid, sid, \texttt{Ask}, \psi\rangle$ (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 $\texttt{Conf}()$) 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.

$$\frac{\begin{array}{c} S_M(M_{In}) = \langle mid, sid, \texttt{Ask}, \psi\rangle \\ \Pi^{ag}_{sid} \models B_{sid}(\psi \rightarrow \varphi) \qquad \Pi^{ag}_{sid} \models \neg B_{sid}(\psi) \qquad \Pi^{ag}_{sid} \models \neg B_{sid}(\neg\psi) \\ (\Pi^{ag}_{sid} \cup \{B_{sid}(\psi)\}) \models B_{sid}(\varphi) \qquad \texttt{Conf}() = \texttt{false} \end{array}}{\mathsf{ProcMsg} \longrightarrow_{AS} \mathsf{ExecInt}} \qquad (\textbf{Ask2})$$

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

When an agent receives a `Response` message, it updates its belief base depending on the result of `Trust`(); I assume this function to determine, depending on the agent profile, whether the sender appears trustworthy. Thus, in **Response1**, when the

[v] Here it suffices to know that this function selects one message from the agent's inbox M_{In}, see [33, 284] for more details about this function $S_M()$.

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.

$$\frac{S_M(M_{In}) = \langle mid, sid, \texttt{Response}, \psi\rangle \qquad \texttt{Trust}() = \texttt{true}}{\mathsf{ProcMsg} \longrightarrow_{AS} \mathsf{ExecInt}} \qquad (\textbf{Response1})$$

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

$$\frac{S_M(M_{In}) = \langle mid, sid, \texttt{Response}, \psi\rangle \qquad \texttt{Trust}() = \texttt{false}}{\mathsf{ProcMsg} \longrightarrow_{AS} \mathsf{ExecInt}} \qquad (\textbf{Response2})$$

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

6.5.5 Example

As a real-world example for the implementation, I take the car-sale scenario as in [258] and in [200]. Donald is a car dealer and Ivan is a potential buyer. When we buy cars, if we are rational (and I assume the potential buyer is rational), then we consider safety[vi] of the vehicle as being a priority. I set up the scenario in Table 8. Further, this scenario corresponds to the fourth case in Table 6.6, instantiating our model using the abstract agents defined in Table 6.1.

I describe the scenario making reference to the running model from § 6.4.1, 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 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**

[vi] [170] includes *passenger safety* as an important part of the car attributes in negotiation scenarios.

Table 6.8: Setup of Real-World Example

Donald (Dec) ***reckless*** **with skill** $\beta = 0.8$
Beliefs: $B_{Dec}(\texttt{safe(X)} \rightarrow \texttt{buy(X)}), B_{Dec}(\neg\texttt{safe(bmw)}), B_{Dec}(\neg\texttt{buy(bmw)})$ **Desires**: $D_{Dec}(B_{Int}(\texttt{buy(bmw)}))$ **Actions**: $A_{Dec}(\texttt{safe(bmw)})$, $A_{Dec}(\neg\texttt{safe(bmw)})$ **ToM** with **confidence** $\gamma = 0.8$: $B_{Dec}(B_{Int}(\texttt{safe(X)} \rightarrow \texttt{buy(X)}))$
Ivan (Int) ***skeptical*** **with trust** $\alpha = 0.8$
Beliefs: $B_{Int}(\texttt{safe(X)} \rightarrow \texttt{buy(X)})$, $B_{Int}(\overline{\texttt{safe(bmw)}})$ **Desires**: $D_{Int}(B_{Int}(\texttt{safe(bmw)}) \vee B_{Int}(\neg\texttt{safe(bmw)}))$ **Actions**: $Q_{Int}(\texttt{safe(bmw)})$ **ToM** : $B_{Int}(\neg B_{Dec}(\overline{\texttt{safe(bmw)}}))$

receives the message, which corresponds to the semantics rule **Response1**, given that, though **Ivan** profile is *skeptical*, 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`.

6.6 Conclusion

In this chapter I have described a high-level approach for modelling deception using Theory-of-Mind in Multi-Agent Systems that integrates components of three major theories of deception described in §2.1, namely TDT, IMT2, and IDT. The aim of this work is 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, the model I have presented here has been successfully implemented 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 to ToM presented in Chapter 4. Furthermore, in order to offer the possibility of extending the model so that it can serve various domains for the study of deception, I have proposed four agent profiles which influence the execution of different behaviours by considering the likelihood of trust and deception between agents. I 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 skeptical 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, I am curious to explore how to increase the order of ToM defined in the current agent architecture. Other research aims would be: the inclusion of a *cognitive load* 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.

Chapter 7

The Evolution of Deceptive Agents

In this chapter I use an evolutionary game theory approach to model deception in social interactions over time between agents in large-scale systems.

Deception plays a critical role in the dissemination of information, and has important consequences on the functioning of cultural, market-based, and democratic institutions. Deception has been widely studied at the intersection of the fields of Philosophy, Psychology, Economics and Political Science. Yet, we still lack an understanding of how deception emerges in a society under competitive (evolutionary) pressures. In this chapter I begin to fill this gap by bridging evolutionary models of social good–*public goods games* (PGGs)–with ideas from *Truth-Default Theory* [153] and from *Interpersonal Deception Theory* [38]. The type of deception that I model in this chapter corresponds to distributed deception Type III (see Chapter 3). This chapter provides a well-founded analysis of the growth of deception in agent societies and the effectiveness of several approaches to reducing deception. Assuming that knowledge is a public good, I use extensive simulation studies to explore (i) how deception impacts the sharing and dissemination of knowledge in agent societies over time, (ii) how different types of knowledge sharing societies are affected by deception, and (iii) what type of policing and regulation is needed to reduce the negative effects of deception in knowledge sharing. The results in this chapter indicate that cooperation in knowledge sharing can be re-established in systems by introducing institutions that investigate and regulate both defection and deception using a decentralised case-by-case strategy. This provides evidence for the adoption of methods for reducing the use of deception in the world around us. This also applies to deceptive behaviour of artificial agents, if we are to assume that machines will indeed develop, or become endowed with, the cognitive capabilities to deceive, that I have discussed in the previous chapters.

7.1 Introduction

Deception plays a critical role in information dissemination. Deception also plays a crucial role in survival and continuation of species. Significant studies from evolutionary biology have even focused on the bio-physiological properties of deception in plants and animals [266], that determine certain types of behaviours. Humans, on the other hand, employ deceptive behaviour at a higher level which is not necessarily determined or highly dependent on their bio-physiological properties [i]. This is especially the case in social, political and economic contexts where deception takes the form of knowledge manipulation. Areas such as Philosophy [160, 92], Psychology [75, 76], Communication Theory [172, 152, 38], Economics [98, 31], Security Studies [122, 265, 294] have looked at the these higher-level properties and components of deception.

Information technologies such as smart-phones, social media and rolling news coverage have greatly increased the access to information and informed decision making which in turn determine opinion formation and behavioural change. However this increased access also offers more opportunities for the knowledge public good to be compromised by deception. Additionally, due to this advancement in technology, other kinds of actors, apart from humans, that share and generate knowledge have

[i] This does not imply that deception in humans is not driven by some bio-physiological properties.

come into existence. In particular, Artificial Intelligence (AI) has gained a strong momentum in the past decades, and this has caused the emergence of intelligent artificial and autonomous agents. The risks related to the sharing of knowledge and information as a social good posed by AI is that artificial autonomous agents might develop their own reasons to act deceptively as it is pointed out in [42] and, more recently, in [130] and throughout this thesis. AI has also seen an emerging interest in the problem of *fake news* and the potential ability of machines (i) to be used for fake news generation [302] and fake news detection [108, 57]; or even (ii) to use higher-order cognitive mechanisms to manipulate the beliefs of others in order to deceive as we have seen in the previous chapters of this thesis. All of this leaves us with an urgent need to understand deception better, and to devise methods for reducing its impact.

A shared system of accurate and non-partisan knowledge confers advantages for all members of society and promotes informed and democratic decision making. However there exists a unilateral incentive to provide misinformation and disinformation that confers advantages to that individual. Therefore mechanisms are required to promote cooperative behaviour and to punish disinformation, misinformation and deception in order to preserve these institutions of knowledge. Indeed, from the perspective of our ability to govern societies, it has become increasingly important to address the future impact that deceptive social agents (human or artificial) can have on society in general. More importantly, we must ask ourselves what types of rules need to be implemented or what types of mechanisms should be employed in order to reduce the impact of deceptive agents on knowledge sharing and generation. In this chapter I begin to provide some answers to these issues.

I do this, by studying a *public goods game* (PGG) model based on the those in [254] and [1]. I use these games to explore how the evolution of cooperation in different populations of agents is influenced by deception. A public good commonly models some type of financial or physical good that is shared by the members of a society, but can also represent a shared system of knowledge within a society, and that is how I use it here. This model allows us to answer the following questions: 1) Does deception lead to the breakdown of cooperation in societies where regulatory institutions exist? and 2) Can cooperation be maintained in societies where deception is present?

7.2 Background

I consider deception to be a strategic and social behaviour that agents (natural or artificial) employ in order to gain advantage over the members they interact with. In game-theoretic terms, deception should be considered (as it is, by the literature) to be a non-cooperative behaviour. However, agents that intend to deceive, usually try to emulate cooperative behaviour. To the target agent that observes it, a deceptive agent appears cooperative, while, in reality, it is non-cooperative. Therefore, the type of agent-based model I chose for this study has to be able to represent deception as a falsely cooperative behaviour while allowing us to test under which circumstances real cooperation is promoted.

The major studies on cooperation focus on the aspects of complexity in agent-based systems and treat cooperation as an emergent behaviour in such multi-agent systems [9, 254, 192]. Thus, even though the overall behaviour of a population of agents can be regarded as complex, it emerges from a set of strategic behaviours that are intrinsically simplistic. The literature shows that PGGs as cooperative game theoretical models can been successfully used to study how cooperation emerges, and that they are a powerful framework to understand the conditions under which cooperation is stable through mechanism design [1, 254]. This is why I adopt a PGG approach here.

The set of behaviours that most work on PGGs has focused on are *coopera-*

tion, *defection* (also known as *free-riding*) and *punishment*. Different variants of punishment have been studied, including *pool-punishment*, where individuals pay a tax to maintain third parties (the punishers) who carry out punishment, and *peer-punishment*, where individuals punish their peers. However, if we look at how agents in human societies behave, we can identify other types of behaviours that are not as straight-forward, such as deception. Apart from simply cooperating, defecting and punishing, humans are also able to mask their intentions behind these behaviours. For instance, an agent can pretend to cooperate while defecting in a PGG. Thus, to other participating players the deceiver seems to be one of the cooperators, thus enjoying both the social benefits of a cooperator and the financial benefits of a defector. As a cooperator, the deceiver might be regarded as being ethical and pro-social, while as a defector the deceiver receives a certain political or financial satisfaction [ii]. In terms of knowledge sharing, a cooperator is transparent and fair, thus it shares truthful information with the other members of society. Contrastingly, a deceiver will contribute with untruthful information when engaged in knowledge sharing.

Given that deception is intrinsically a communicative behaviour, I refer to the literature in Communication Theory to see what factors must be included in our PGG model. I mainly consider factors that directly influence deceptive behaviour in social interactions. One such cognitive factor, according to Truth-Default Theory, is the default attitude of trust [152]. Apart from a bias towards trustworthiness, there are other socio-cognitive factors such as *cognitive load* [iii], *leakage* [iv], and *communicative skill* [v] that have been identified by Interpersonal Deception Theory (IDT) in [38].

[ii] According to [98], deception should be a selfish behaviour that aims to maximise one's payoff.

[iii] The cognitive effort that is spent in order to solve a task, such as forming or planning a deceptive strategy.

[iv] The information leaked by a deceptive agent due to cognitive load.

[v] The social skill of an agent to form, plan, and deliver messages. According to IDT, the communicative skill regulates cognitive load and reduces leakage.

7.3 Methods

7.3.1 Agent-Based Modelling of Public Goods Games (PGGs)

I apply mechanism design and evolutionary game theory to study the behaviour of populations of agents of a fixed size N in six different PGGs, each with a different set of strategies. In a PGG each participant is faced with two options: a) contribute to the public pool a given amount $c > 0$; or b) not contribute to the public pool. After the participant picks an option, it receives an amount $r \times c \times \frac{M_C}{M}$, where r represents a multiplier representing the increasing returns of cooperative behaviour, M_C represents the number of contributors and M the total number of participants. If $M_C = M$ it means that the social good is maximised and each participant receives the amount equal to $r \times c$. Whatever the case, each participant receives an equal share $r \times c \times \frac{M_C}{M}$ regardless of whether they contributed to the public pool. In the absence of punishment, free-riding (taking the payoff without contributing to it) becomes the dominant strategy.

For each PGG, I perform explicit computations of what payoffs the agents will receive given a sub-population that is selected to play the game at each iteration. The relative differences between the payoffs obtained by the agents with different strategies in the sub-population determine the probability that an agent will adopt a different strategy given a function of the *imitation strength* $s \geq 0$ which represents social learning, together with the *exploration rate* $\mu \geq 0$ which represents the natural inclination of agents to randomly adopt another strategy. The imitation strength, or social learning, can be either weak/intermediate or strong. For weak/intermediate social learning the value of s is a fixed number, while for strong social learning the value of s approaches ∞. Social learning represents the tendency of an agent that is selected for mutation to adopt a strategy that compared to its current strategy maximises the agent's payoff.

The components of our PGGs are the following: A non-empty set of strategies $S \neq \emptyset$; N that is the number of agents in a population to play a PGG; n_{S_i} represents the number of agents in a population with a given strategy S_i; M the number of agents that is selected to play a PGG from a population N; r is a multiplication factor that is always $1 < r < M - 1$; c represents the investment a cooperative agent contributes to a PGG; c_{S_i} denotes the cost of a given strategy S_i; Π_{S_i} represents the payoff of a given strategy S_i; s represents the imitation strength, which in my model represents social learning; μ is the mutation rate at which an agent is selected to learn the strategy of another agent; B is the pool-punishment for Defection; b is the peer-punishment for Defection; c_b is the cost of punishing a Defector; G is the cost of Pool-Punishment; Γ is the punishment or tax for Deception; and finally σ represents the payoff for Non-Participation. These parameters are summarised in Table 7.2 when discussing the results in §7.4.

Regarding social learning (imitation strength), I assume that two players i and j are randomly chosen. Their expected payoff values Π_i and Π_j depend on the strategies of the two players and on the frequencies C, D, L etc. of the strategies. I adopt the assumption made in *Sigmund et al.* [254] and *Abdallah et al.* [1] that player i adopts the strategy of player j with a probability which is an increasing function of the payoff difference $\Delta = \Pi_i - \Pi_j$, that is $\frac{1}{1+\exp(-s\times\Delta)}$. The higher the value for s, the stronger the tendency of adopting the better strategy. When $s \overset{\infty}{\rightarrow}$ the agent will always adopt the better strategy.

The exploration rate μ can be viewed as a mutation which models random mistakes in actions as well as purposeful exploration regardless of relative payoffs. This stochastic approach allows us to dynamically represent how the frequencies of the different types of agents evolve over time. The probability of a player to change from strategy X to another strategy Y is $\mu_{X,Y} = \frac{\mu}{n-1}$, where n is the total number of strategies in S.

I start from Sigmund's voluntary PGG [254] as a baseline before introducing other PGGs, each with different compositions of agents to see how the strategies influence each other. The full set of strategies that I use are as follows, with Cooperators, Defectors, Loners, Peer-Punishers, and Pool-Punishers being the set studied in [254] and Deceivers, Interrogators and the two Hybrid strategies being novel strategies introduced here.

- **Cooperator (C):** the Cooperator receives the PGG payout (Eq.7.2) and pays the PGG contribution c. The Cooperator also pays β, which represents the tax for Punishers to exist.
- **Defector (D):** the Defector receives the PGG payout, without paying the PGG contribution. However, the Defector pays a tax inflicted by the Pool-Punishers B or the Peer-Punishers b, or by both types of Punishers depending on the PGG that is being played.
- **Loner (L)** (a.k.a Non-Participation)**:** the Loner always receives the same payoff σ, no matter what PGG is being played.
- **Pool-Punisher (PoP):** the Pool-Punisher receives the PGG payout, pays the PGG contribution c, as well as the cost of pool-punishment G. On top of this, the Pool-Punisher receives a reward that is the tax payed by the Cooperators multiplied by the number of Cooperators playing the game and divided by the total number of Punishers and Interrogators selected to play the game, depending on the PGG that is being played.
- **Peer-Punisher (PeP):** the Peer-Punisher receives the PGG payout, pays the PGG contribution c, as well as the cost of peer-punishment c_b multiplied by the number of Defectors. On top of this, the Peer-Punisher receives a reward that is the tax payed by the Cooperators multiplied by the number of

Cooperators playing the game and divided by the total number of Punishers and Interrogators selected to play the game, depending on the PGG that is being played.

- **Deceiver (*Dec*):** the Deceiver receives the PGG payout and does not pay the PGG contribution (similar to what the Defector is doing). On top of that, the Deceiver is not punished by either type of Punisher. Instead, the Deceiver can be interrogated by the Interrogator agents and can pay the cost of deception if it is caught. The cost of deception depends on the the cognitive load of the Deceiver as well as on the risk of leakage from the Deceiver. Finally, the cost of deception is also influenced by the Deceiver's communicative skill and the number of agents it needs to deceive in a game.

- **Interrogator (*Int*):** the Interrogator receives the PGG payout and pays the PGG contribution c. The Interrogator also receives the reward paid by the Cooperators and divided by the total number of Interrogators and Punishers, depending on the PGG that is being played. The Interrogator also pays the cost of Interrogation, which consists of the cost of interrogating agents in the PGG and the cost of punishing the interrogated agent in the PGG which turn out to be deceptive.

- **Pool-Hybrid Interrogator (H_{PoP}):** this type of Interrogator plays both the role of Interrogator and the role of Pool-Punisher. Therefore, it inherits the costs of both types of agents, while receiving the PGG payout, and of course paying the PGG contribution.

- **Peer-Hybrid Interrogator (H_{PeP}):** this type of Interrogator plays both the role of Interrogator and the role of Peer-Punisher. Therefore, it inherits the costs of both types of agents, while receiving the PGG payout, and of course paying the PGG contribution.

For our PGG model, each strategy, except for Non-Participation, falls into one of the meta-strategies of cooperative game theory, namely Cooperation and Free-Riding. The **Cooperation** meta-strategy, which determines an agent to make a contribution to the social good, includes: Cooperation, Pool-Punishing, Peer-Punishing, Interrogation, Pool-Hybrid Interrogation and Peer-Hybrid Interrogation; the **Free-Riding** meta-strategy, which determines an agent to not contribute anything to the social good while enjoying the benefits of the social good, consists of Defection and Deception. A payout where Free-Riders are playing a PGG would be:

$$Payout^{*} = c \times r \times \frac{N - n_{FR} - n_L - 1}{N - n_L - 1} \tag{7.1}$$

In Eq.(7.1) where a PGG is played by a fixed population with N agents, n_{FR} represents the total number of Free-Riders, and n_L represents the total number of Loners (Non-Participants). This payout is consistent with the previous evolutionary models of PGGs [254, 1].

The total number of Free-Riders is $n_{FR} = n_D + n_{Dec}$. However, since in our PGG models the Deceivers are pretending to cooperate, they are not discounted form the calculation of the payout. Thus, our payout where Deceivers are present is:

$$Payout = c \times r \times \frac{N - n_D - n_L - 1}{N - n_L - 1} \tag{7.2}$$

Because I assume voluntary participation in the PGG, then I will need to take into account the probability that the other M - 1 players of a sample are unwilling to participate:

$$P_\sigma = \frac{\binom{n_L}{M-1}}{\binom{N-1}{M-1}} \tag{7.3}$$

Therefore, a general payoff can be represented like this:

$$\Pi = P_\sigma \times \sigma + (1 - P_\sigma)(Payout - c) - cost\frac{M-1}{N-1} + reward\frac{M-1}{N-1} \tag{7.4}$$

I model six different PGGs with different population compositions summarized in Table 7.1. The details regarding the payoffs for each strategy and the modelling of each PGG can be found in Appendix A. Here I describe the high-level conceptual design of the six PGGs:

PGG1: This is the PGG based on [254] where second-order punishment has been substituted with a fixed tax β that is to be paid by the Cooperators for Punishers to exist. This is similar to paying a tax for policing in a society. This PGG consists of Cooperators, Defectors, Loners, Peer-Punishers, and Pool-Punishers.

PGG2: In this PGG I keep the same types of agents as in PGG1 and I introduce the Deceivers. In this setup, the Deceivers are able to free-ride without risking to be caught by Interrogators.

PGG3: In this PGG I keep the same setup as in PGG2, but I replace the Peer-Punishers with Interrogators. Interrogators are able to chase Deceivers, while the remaining Pool-Punishers are able to punish Defectors.

PGG4: In this PGG I keep the same setup as in PGG2, but I replace the Pool-Punishers with Interrogators. Interrogators are able to chase Deceivers, while the remaining Peer-Punishers are able to punish Defectors.

PGG5: In this PGG I keep the same setup as in PGG3. However, instead of having two different types of agents chasing Defectors and Deceivers separately, I have a single type of agent that performs both jobs, namely the Pool-Hybrid Interrogator. This is analogous to having a centralised policing institution in a society which keeps track of both types of free-riding behaviours.

PGG6: In this PGG I keep the same setup as in PGG4. However, instead of having two different types of agents chasing Defectors and Deceivers separately, I have a single type of agent that performs both jobs, namely the Peer-Hybrid Interrogator. This is analogous to having a decentralised policing institution in a society which keeps track of both types of free-riding behaviours.

Table 7.1: PGGs and the combination of strategies.

PGG	C	D	L	PeP	PoP	Dec	Int	Hpop	Hpep
1	✓	✓	✓	✓	✓	×	×	×	×
2	✓	✓	✓	✓	✓	✓	×	×	×
3	✓	✓	✓	×	✓	✓	✓	×	×
4	✓	✓	✓	✓	×	✓	✓	×	×
5	✓	✓	✓	×	×	✓	×	✓	×
6	✓	✓	✓	×	×	✓	×	×	✓

The eight strategies that I used in the six PGGs are: Cooperation C, Defection D, Non-Participation (Loners) L, Peer-Punishing PeP, Pool-Punishing PoP, Deception Dec, Interrogation Int, Pool-Hybrid Interrogation H_{PoP}, and Peer-Hybrid Interrogation H_{PeP}.

7.3.2 Modelling Deception and Interrogation

Trust in society I consider trust to be proportional to the number of Cooperators in games, but due to the complexity of the game I am modelling I need to make the distinction between genuine cooperators, represented by Cooperators and Interrogators/ Punishers, and total cooperators which includes Deceivers, represented by $N - n_D$. Deceivers are pretending to cooperate, thus they influence the overall trust between members of a population. I have derived this definition of trust based on Truth-Default Theory in deception literature [152], which states that human agents are biased to trust others by default. I use $t = \frac{N-n_D}{N}$ to represent the trust between a population of agents. Trust has the following properties in our model: 1) *Trust increases both the likelihood of cooperation and deception*; 2) *Trust reduces the likelihood of defection*; and 3) *Low trust means more defectors in a selected population*

M which determines lower payoffs for Defectors and Deceivers.

Deception model Deceivers receive the payout of the PGG without making the PGG contribution. They are distinguished from Defectors because they are not subject to punishment as they conceal their defection. However this concealment is costly; it increases with the number of other agents that must be convinced, but decreases with overall trust among the population and the deceivers' innate communicative skill. I consider the following components that contribute to a Deceiver's payoff:

Definition 27 ***Cost of Deception*** *Let the cost of deception c_{Dec} be a function of $cogLoad$ and $leakage$, where $c_{Dec} = cogLoad + leakage$.*

1. $commSkill$ represents the communicative skill of the Deceiver.

   (a) Reduces the cost of deception.

   (b) The higher the communicative skill, the more likely it is for a Deceiver to succeed in deception.

2. $\gamma = 1 - commSkill$ represents the Deceivers' risk of getting caught.

3. $cogLoad = (n_C + n_{Int} + n_{Dec} + n_P) \times (1 - t) \times (1 - commSkill)$: The cognitive load of a Deceiver, where:

   (a) $n_C + n_{Int} + n_{Dec} + n_P$ represents the number of agents that need convincing. Here I also add the number of Deceivers, because a Deceiver considers other Deceivers to be Cooperators. n_P is the number of Punishers (Peer or Pool).

   (b) $(1 - t) \times (1 - commSkill)$ represents the cost to convince an agent. t represents the trust in society, and that was defined in the previous subsection.

4. $leakage = n_{Int} \times \gamma \times \Gamma$ represents the leakage of the Deceiver.
   (a) Increases the cost of deception.
   (b) Leakage means that the deceiver leaves a track of evidence that might lead an Interrogator to find out about deception.

Interrogation model Interrogators receive the same payout as the Peer-Punishers minus the cost of peer-punishing. They are different from Peer-Punishers as they do not punish Defectors. However, Interrogators need to hunt down Deceivers and punish them, therefore they need to pay a cost for interrogation. This cost increases with the number of agents in a population they need to interrogate as well as with the number of Deceivers they are likely to reveal and punish. I consider the following components that contribute to an Interrogator's payoff:

Definition 28 ***Cost of Interrogation*** *Let the cost of being an Interrogator* c_{Int} *be a function of* c_Γ *and* c_{interr}*, where* $c_{Int} = \gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$*.*

1. c_Γ: cost of punishing a Deceiver. It is multiplied by:
   (a) The probability of a deceiver's risk of getting caught γ, which represents the likelihood of revealing a Deceiver. This multiplication represents the risk of a Deceiver being caught in a given population.
   (b) The number of Deceivers n_{Dec}
2. c_{interr}: cost of interrogating an agent. It is multiplied by:
   (a) The numbers of agents that need to be interrogated. These are both Cooperators and Deceivers $n_C + n_{Dec}$.

7.4 Results

Here I present the results from the simulations. Each simulation is a run of 10^5 PGG games. Each game contains $N = 100$ agents. In the first game, all agents are **Defectors**, and after each game the population evolves as described above. I ran two sets of simulations, one with strong social learning ($s = s \xrightarrow{\infty}$), and one with weak/intermediate social learning ($s = 1000$). I ran 10^3 simulations for each of these two learning conditions, and report results in terms of the frequencies with which agents picked particular strategies at the end of those runs[vi]. These figures are reported as averages over the 10^3 runs. Other parameters are: $M = 5$, $\mu = 0.001$, $c = 1$, $r = 3$, $\sigma = 0.3$, $b = c_b = 0.7$, $B = G = 0.7$. The fixed parameters for deception were $\beta = 0.5$, $\Gamma = 0.8$, $c_\Gamma = 0.5$, $c_{interr} = 0.5$, and $commSkill = 0.5$. The parameter values are identical to those used in [254] and [1], except for the σ. I used $\sigma = 0.3$ in order to incentivise participation, whereas [254] and [1] used $\sigma = 1$. Regarding the parameters of deception, I also tested their effects on the long-run frequencies.

When reading the barcharts for each PGG (Fig.7.1 to Fig.7.12), the coloured bars represent the long-term average frequency of agents with a given strategy where each colour represents a different strategy of the PGG. The error bars represent ±1 standard deviation from this mean given that iterations of a single PGG do not necessarily have the same outcome in terms of long-run population frequencies. I also performed statistical Kruskall-Wallis nonparametric (distribution free) [vii] tests for each PGG in order to analyse variance between payoff samples over all strategies in a given PGG assuming that strategies are dependent variables. To compare the payoff samples of the same strategies in different PGGs, I performed pairwise

[vi] These frequencies are the same as the proportion of agents picking the strategies.

[vii] Tests whether two or more samples come from the same distribution. A significant test indicates that at least one sample stochastically dominates one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains [139].

nonparametric Mann-Whitney tests [viii], assuming independence between the PGGs. The results from both types of tests gave very low p-values ($p < 0.01$), meaning that the differences between the payoff averages obtained from our simulations are statistically significant and they have not occurred by chance.

Table 7.2: Parameter values for PGGs

Number of agents in a population to play a PGG	N	100
Number of iterations of a PGG	T	10^5
Number of agents selected to play a PGG	M	5
Social learning (imitation strength)	s	1000 or ∞
Exploration rate	μ	0.001
PGG contribution	c	1.0
PGG multiplier	r	3.0
Loner (Non-participation) payoff	σ	0.3
Pool punishment effect	B	0.7
Pool punishment cost	G	0.7
Peer punishment effect	b	0.7
Peer punishment cost	c_b	0.7
Tax for punishers to be present	β	0.5
Punishment for deception	Γ	0.8
Cost to punish a deceiver	c_Γ	0.5
Cost to interrogate agents	c_{interr}	0.5
Communicative skill (for deceivers)	$commSkill$	0.5

7.4.1 PGG1 - Punishment

I reproduced the results in [254] without second-order punishment. Therefore, Peer-Punishers dominated the games with the following long-run average frequencies for: 1) $s = 1000$: [C : 0.0 (SD=0.0), D : 0.001 (SD=0.004), L : 0.001 (SD=0.006), PoP : 0.0 (SD=0.001), PeP : 0.998 (SD=0.01)]; and 2) $s \xrightarrow{\infty}$: [C : 0.0 (SD=0.0), D : 0.0 (SD=0.001), L : 0.001 (SD=0.003), PoP : 0.0 (SD=0.0), PeP : 0.999 (SD=0.004)]. See model details in Table A.1 in Appendix.

[viii] Tests the alternative hypothesis that one distribution is stochastically greater than the other, under the assumption of continuous responses [161].

7.4.2 PGG2 - Deception

Subsequently, I introduced Deceivers to check if they destabilise Cooperation by reducing the long-run frequency of Peer-Punishers. Deceivers have indeed had an impact on the system, given the following long-run average frequencies for: 1) $s = 1000$: [C : 0.032 (SD=0.031), D : 0.324 (SD=0.1), L : 0.187 (SD=0.065), PoP : 0.026 (SD=0.029), PeP : 0.133 (SD=0.089), Dec : 0.295 (SD=0.104)]; and 2) $s =\xrightarrow{\infty}$: [C : 0.034 (SD=0.031), D : 0.324 (SD=0.099), L : 0.186 (SD=0.064), PoP : 0.025 (SD=0.025), PeP : 0.136 (SD=0.094), Dec : 0.294 (SD=0.1)]. See model details in Table A.2 in Appendix.

7.4.3 PGG3 - Interrogation with Pool-Punishment

In order to try and re-establish Cooperation, I replaced the Peer-Punishers with Interrogators in the population composition to check whether Interrogation and Pool-Punishment have a positive impact on the system. Unfortunately, even though Interrogators are present and they are able to reduce the frequency of Deceivers for strong imitation, more Defectors seem to be invading the system. This indicates the ineffectiveness of the Pool-Punishers in this PGG, as shown in the long-run average frequencies for: 1) $s = 1000$: [C : 0.083 (SD=0.053), D : 0.378 (SD=0.103), L : 0.22 (SD=0.067), PoP : 0.075 (SD=0.059), Int : 0.029 (SD=0.027), Dec : 0.215 (SD=0.107)]; and 2) $s =\xrightarrow{\infty}$: [C : 0.082 (SD=0.054), D : 0.383 (SD=0.104), L : 0.219 (SD=0.07), PoP : 0.076 (SD=0.063), Int : 0.029 (SD=0.027), Dec : 0.211 (SD=0.105)]. See model details in Table A.3 in Appendix.

7.4.4 PGG4 - Interrogation with Peer-Punishment

Due to the lack of efficiency of Pool-Punishers given by the results in both PGG1 (without Deception) and PGG3 (with Deception and Interrogation) I decided to replace them with Peer-Punishers given their absolute dominance in PGG1. Unfortunately, Peer-Punishers have proven to be almost as inefficient as Pool-Punishers given the following long-run average frequencies for: 1) $s = 1000$: [$C : 0.035$ (SD=0.03), $D : 0.343$ (SD=0.101), $L : 0.196$ (SD=0.067), $PeP : 0.118$ (SD=0.082), $Int : 0.05$ (SD=0.052), $Dec : 0.258$ (SD=0.105)]; and 2) $s =\xrightarrow{\infty}$: [$C : 0.036$ (SD=0.033), $D : 0.349$ (SD=0.1), $L : 0.197$ (SD=0.068), $PeP : 0.116$ (SD=0.086), $Int : 0.051$ (SD=0.052), $Dec : 0.251$ (SD=0.101)]. Even though Peer-Punishers alone perform better in the PGG, they have a negative impact on the performance of both Interrogators and Cooperators. See model details in Table A.4 in Appendix.

7.4.5 PGG5 - Pool-Punishment and Interrogation Hybrid

When I removed both Interrogators and Peer-Punishers from the system, I introduced Pool-Hybrid Interrogators. This type of hybrid proved to be even less efficient against Deceivers and Defectors compared to when I had both Interrogators and Pool-Punishers acting separately. This is reflected in the long-run average frequencies for: 1) $s = 1000$: [$C : 0.085$ (SD=0.049), $D : 0.344$ (SD=0.091), $L : 0.259$ (SD=0.07), $H_{PoP} : 0.055$ (SD=0.043), $Dec : 0.258$ (SD=0.094)]; and 2) $s =\xrightarrow{\infty}$: [$C : 0.082$ (SD=0.051), $D : 0.345$ (SD=0.089), $L : 0.263$ (SD=0.074), $H_{PoP} : 0.052$ (SD=0.04), $Dec : 0.258$ (SD=0.096)]. See model details in Table A.5 in Appendix.

7.4.6 PGG6 - Peer-Punishment and Interrogation Hybrid

In PGG6, I replaced the Pool-Hybrid Interrogators with Peer-Hybrid Interrogators. For intermediate imitation $s = 1000$, Peer-Hybrid Punishers perform much better than the Pool-Hybrid Interrogators, but not enough to re-establish strong levels of Cooperation in the system. However, for strong imitation Peer-Hybrid Interrogators seem to re-establish strong levels of Cooperation by significantly reducing the influence of Defection and more importantly, Deception for strong imitation. These results are reflected in the long-run average frequencies for: 1) $s = 1000$:

[C : 0.095 (SD=0.049), D : 0.295 (SD=0.095), L : 0.228 (SD=0.076), H_{PeP} : 0.184 (SD=0.137), Dec : 0.199 (SD=0.095)]; and 2) $s \overset{\infty}{=\longrightarrow}$: [$C$: 0.028 (SD=0.037), D : 0.096 (SD=0.117), L : 0.07 (SD=0.082), H_{PeP} : 0.742 (SD=0.279), Dec : 0.063 (SD=0.085)]. See model details in Table A.6 in Appendix.

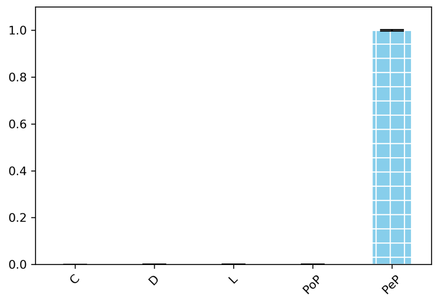


Figure 7.1: Long-run average frequencies in PGG1 for $s = 1000$, $p < 0.001$.

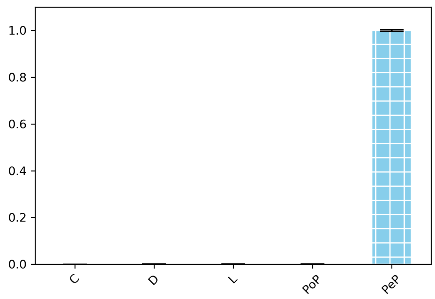


Figure 7.2: Long-run average frequencies in PGG1 for $s \xrightarrow{\infty}$, $p < 0.001$.

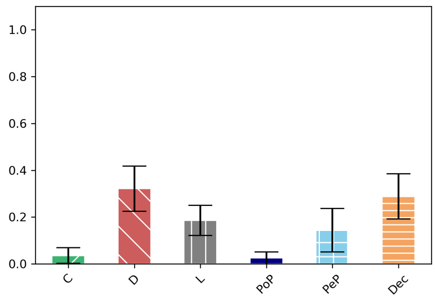


Figure 7.3: Long-run average frequencies in PGG2 for $s = 1000$, $p < 0.001$.

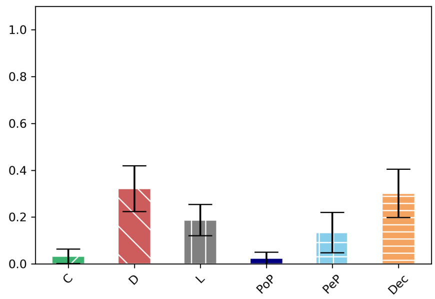


Figure 7.4: Long-run average frequencies in PGG2 for $s \xrightarrow{\infty}$, $p < 0.001$.

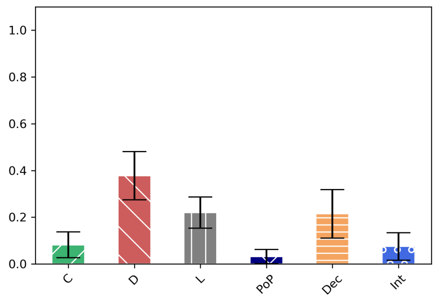


Figure 7.5: Long-run average frequencies in PGG3 for $s = 1000$, $p < 0.001$.

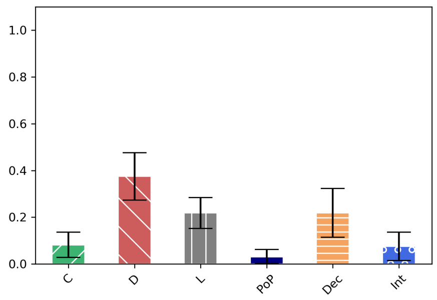


Figure 7.6: Long-run average frequencies in PGG3 for $s \xrightarrow{\infty}$, $p < 0.001$.

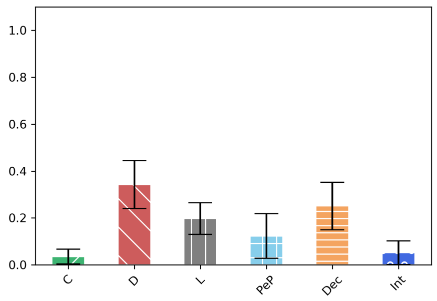


Figure 7.7: Long-run average frequencies in PGG4 for $s = 1000$, $p < 0.001$.

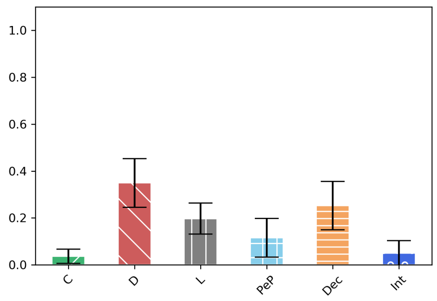


Figure 7.8: Long-run average frequencies in PGG4 for $s \xrightarrow{\infty}$, $p < 0.001$.

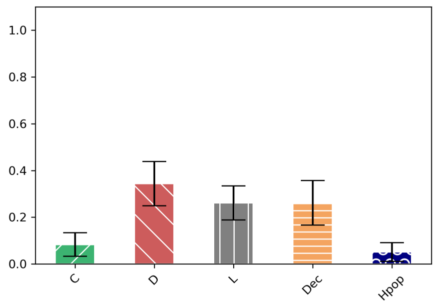


Figure 7.9: Long-run average frequencies in PGG5 for $s = 1000$, $p < 0.001$.

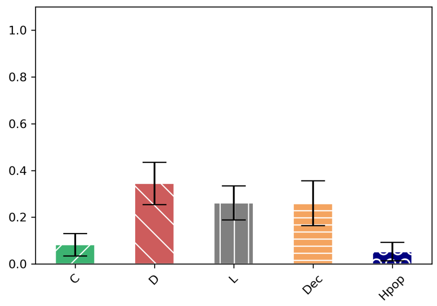


Figure 7.10: Long-run average frequencies in PGG5 for $s \xrightarrow{\infty}$, $p < 0.001$.

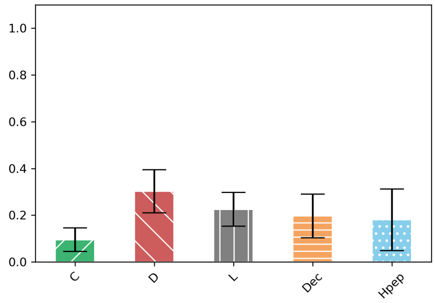


Figure 7.11: Long-run average frequencies in PGG6 for $s = 1000$, $p < 0.001$.

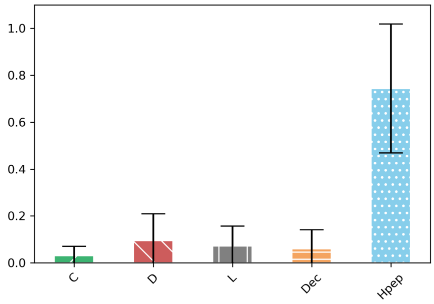


Figure 7.12: Long-run average frequencies in PGG6 for $s \xrightarrow{\infty}$, $p < 0.001$.

7.4.7 Cooperation vs Free-Riding

I have compared the long-run average frequencies of the two meta-strategies, namely Cooperation and Free-Riding, given weak/intermediate (see Table 7.3) and strong social learning (see Table 7.4). The results indicate the following:

1. The introduction of Deception promotes Free-Riding in voluntary PGGs with Punishment.

2. Strong social learning promotes Deception if no Interrogation is present.

3. Strong social learning does not promote Cooperation when Deception is present, unless Peer-Hybrid Interrogation is introduced. Figure A.4 shows how the frequency of Peer-Hybrid Interrogators increases with in social learning (imitation strength).

As for the parameters that directly influence Deception as a strategy in PGGs with and without Peer-Hybrid Interrogators, there are several observations to be considered. First, it seems that in the absence of any type of Interrogation (PGG2), as well as in the presence of Peer-Hybrid Interrogators (PGG6), Deception becomes the optimal strategy if communicative skill is maximised ($commSkill \ = \xrightarrow{1}$). In other words, deception is optimal if you are good at it. Second, the tax paid by Cooperators for Punishers and Interrogators to exist influences PGG2 and PGG6 in opposite ways in terms of total Cooperation. In PGG2, increases in β promote Deception and by extension Free-Riding, while in PGG6, increases in β promote Peer-Hybrid Interrogation. Thirdly, increases in the tax on Deception Γ, which is inflicted by all the Interrogators in all PGGs where these are present, have a positive impact in some PGGs. The increase manages to significantly reduce the frequency of Deceivers in all PGGs where it can be inflicted, however it only has a positive impact on total Cooperation in PGG5, where it promotes Pool-Hybrid Interrogation and Cooperation, and in PGG6 where it promotes Peer-Hybrid Interrogation. The drawback is that in PGG5, Γ needs to be very high ($\Gamma > 800$) in order to begin promoting the cooperative strategies. This is not the case for PGG6, where increases in Γ have an instant impact on promoting Cooperation as a meta-strategy.

Table 7.3: Cooperation vs Free-Riding long-run frequencies for weak/intermediate social learning.

PGG	**1***	**2**	**3**	**4**	**5**	**6**
Cooperators	0.99	0.193	0.187	0.202	0.139	0.278
Free-Riders	0.0	0.618	0.592	0.601	0.601	0.493

For weak/intermediate levels of social learning, Free-Riding dominates in all PGGs where Deception is present.

Table 7.4: Cooperation vs Free-Riding long-run frequencies for strong social learning.

PGG	**1***	**2**	**3**	**4**	**5**	**6**
Cooperators	0.99	0.195	0.187	0.203	0.134	0.77
Free-Riders	0.0	0.618	0.594	0.599	0.602	0.159

For strong levels of social learning, Free-Riding dominates in PGGs where Deception is present, but Cooperation is re-established when Peer-Hybrid Interrogation is introduced.

Figure 7.13: Long-run average frequencies for Cooperation vs Free-Riding for $s = 1000$.

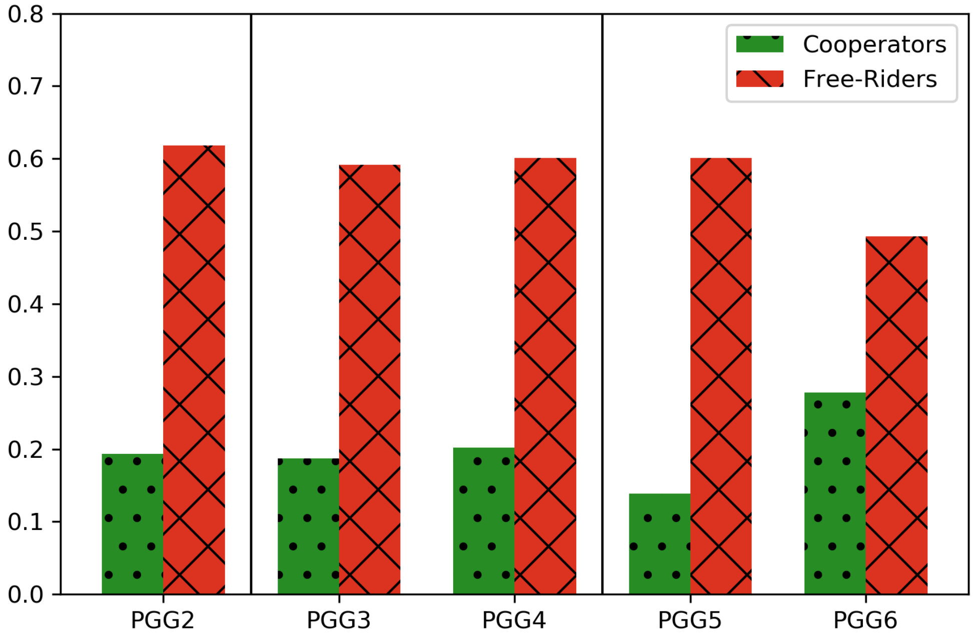
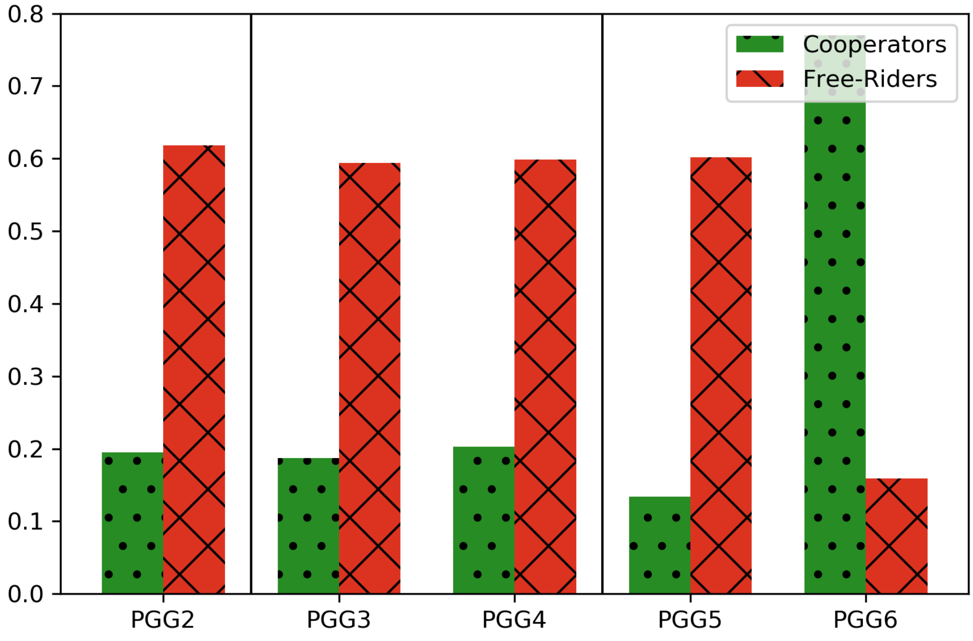


Figure 7.14: Long-run average frequencies for Cooperation vs Free-Riding for $s \xrightarrow{\infty}$.



7.4.8 Sensitivity Analysis: Factors of Deception

Independently on whether there exists a strategy which can re-establish Cooperation in a society, such as in the PGG6 game, there are several socio-dynamical factors that influence such outcomes. How do these factors involved in deceptive interactions influence the distribution of strategies in PGG6?

To answer this question, I focus on the one-at-a-time sensitivity analysis for the Deception and Interrogation parameters of PGG6. I am mainly interested in the influence of the parameters from the lower section of Table 7.2 on the population frequency. These represent the following factors: *Tax for punishers to be present* β, *Punishment for deception* Γ, *Cost to punish a deceiver* c_Γ, *Cost to interrogate agents* c_{interr}, and the *Communicative skill of deceivers commSkill*.

First, the tax on Punishers to exist (policing tax) β strongly influences the promotion of Cooperation as a meta-strategy (see Figure 7.15). However, the increase in tax means that the cooperative agents of a system need to be able to support the Hybrid Peer-Punishers independently on whether they are selected to play the PGG or not. This would inflict considerable costs for such a system to be maintained by the community that uses it.

Second, the tax on deceptive behaviour Γ does have a negative impact on deception and also promotes the dominance of Hybrid Peer-Punishers (see Figure 7.16).

Third, Deceivers that are very skilled $commSkill > 0.9$ manage to dominate PGG6 (see Figure 7.17).

Fourth, neither the cost of interrogating agents c_{interr} (see Figure 7.18), nor the cost of punishing a Deceiver (see Figure 7.19) have beneficial effects on Cooperation.

In conclusion, I can say that systems such as PGG6, where decentralised systems permit the peer-punishment of defection and the individual interrogation of knowledge sources, are influenced by (i) the policing tax paid by cooperative agents, e.g., the more agents are willing to pay for safety, the safer the system; (ii) how strongly agents are punished for their deceptive behaviour; (iii) the skill of agents at communicating deceptively. The cost of interrogation, e.g., how many resources need to be spent to detect and punish deceptive behaviour, has no direct effect on this type of decentralised systems.

Figure 7.15: Effect of Punishment Tax β in PGG6.

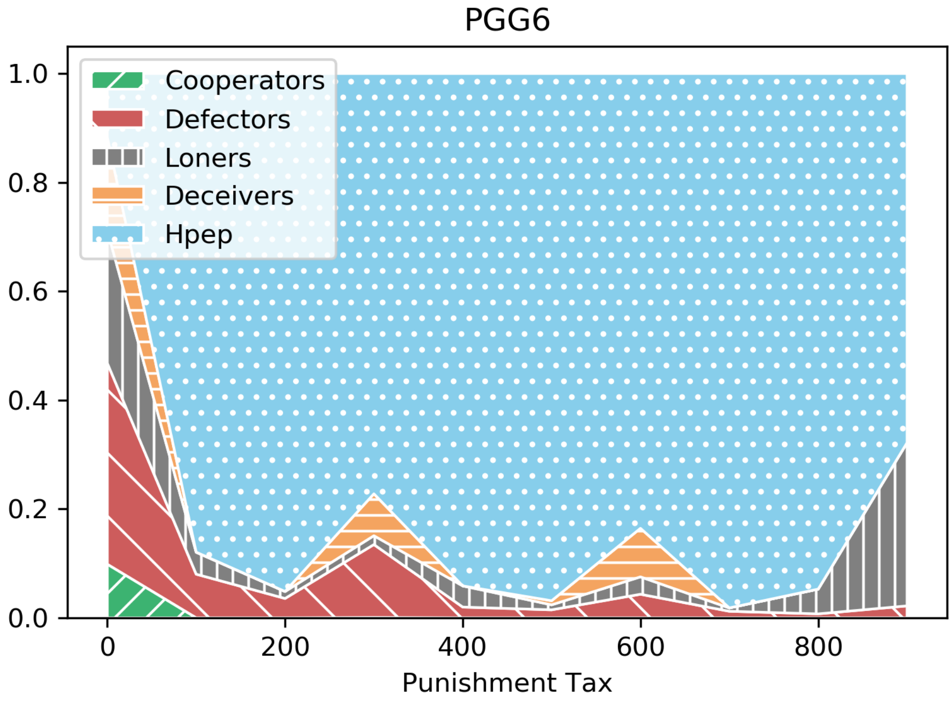


Figure 7.16: Effect of Deception Tax Γ in PGG6.

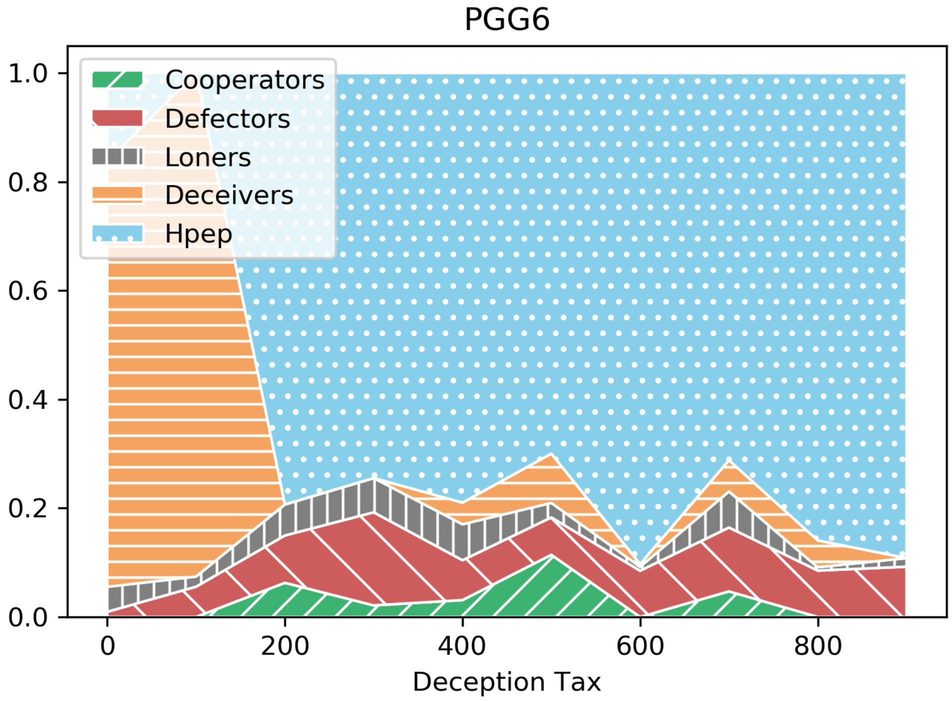


Figure 7.17: Effect of Communicative Skill *commSkill* in PGG6.

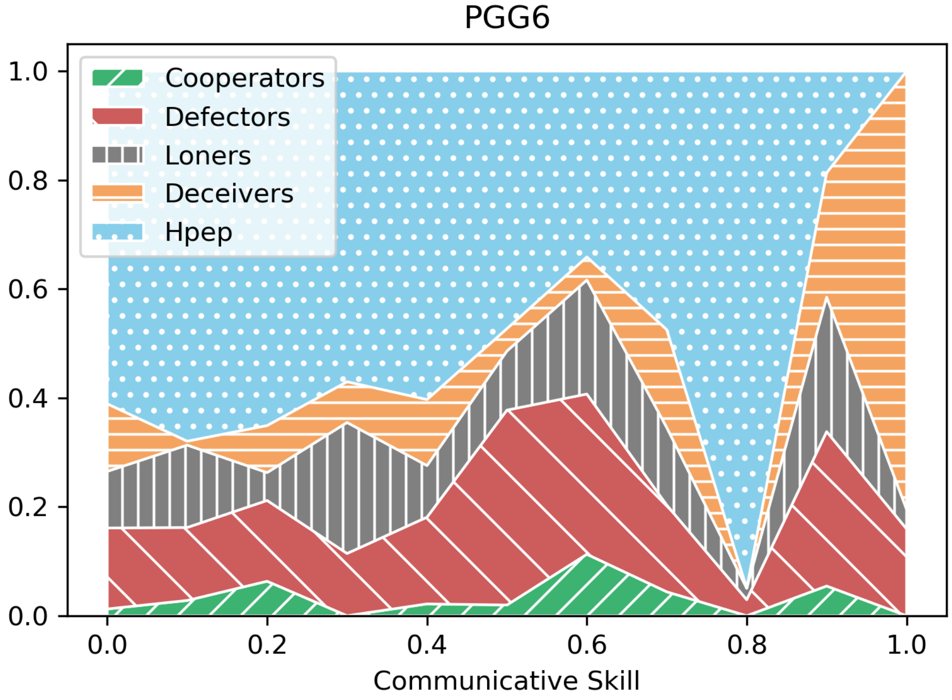


Figure 7.18: Effect of Cost of Interrogation c_{interr} in PGG6.

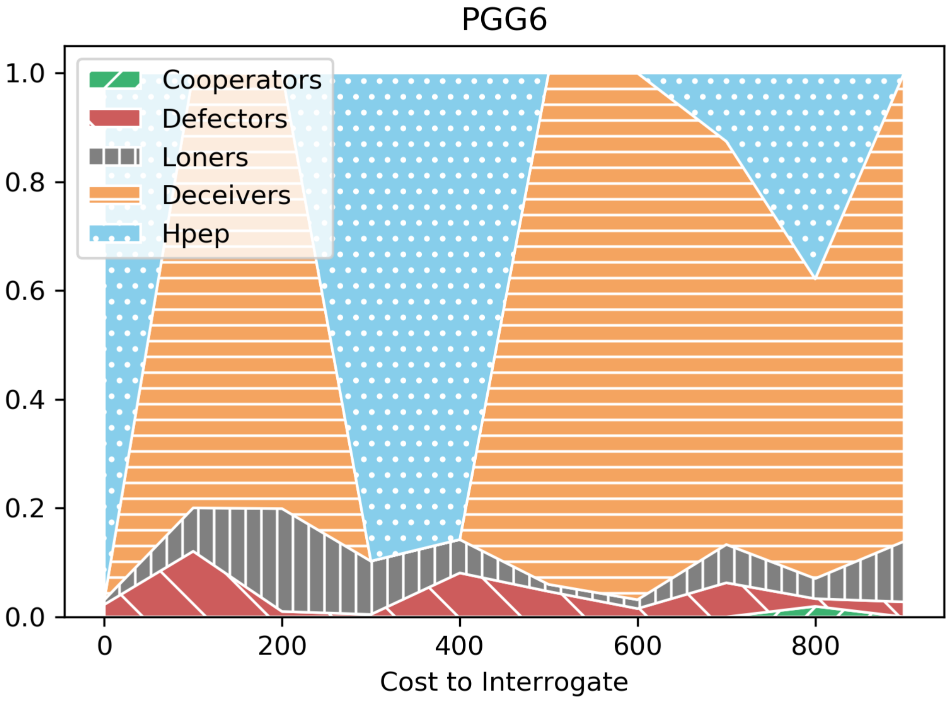
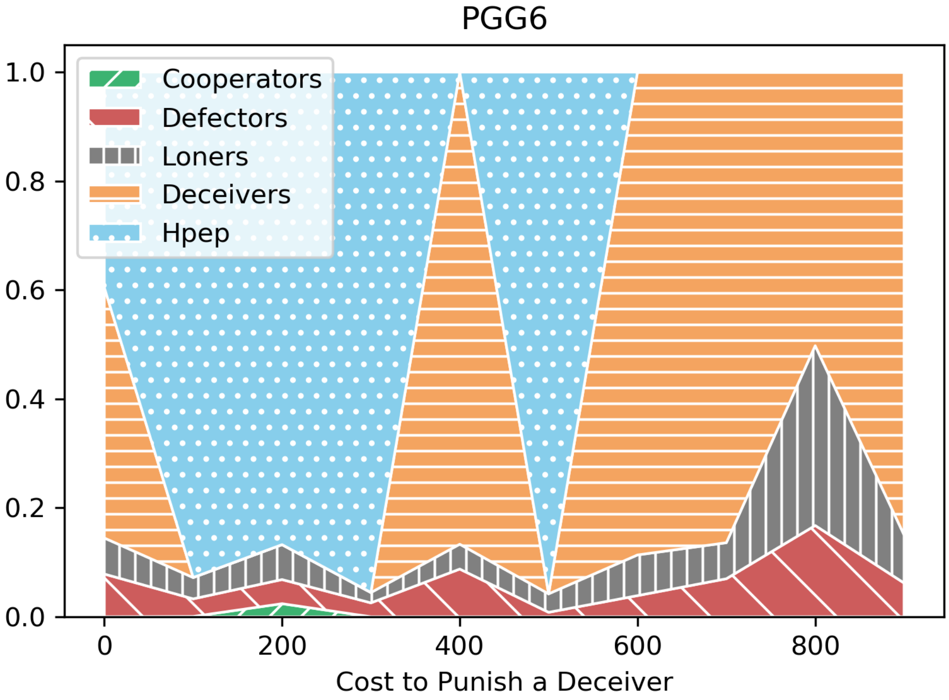


Figure 7.19: Effect of Cost of Punishing a Deceiver c_{Γ} in PGG6.



7.4.9 To Deceive or not to Deceive?

I have also scaled down the PGG approach in order to make a direct comparison between Deceivers and Cooperators in voluntary social interactions. This meant eliminating all the strategies from before except for Deceivers, Cooperators and Loners. In this game, there is no extra cost for Cooperators other than paying their contribution c to the PGG. For Deceivers, the cost of deception is the same as in PGG2, however, their cost now depends on a fixed value for trust instead of being computed based on the dynamics of the population distribution. This is because trust would have been $t = 1$, as there were no Defectors. That is why I had to redefine trust t as a parameter, with $t = 0.5$. I have picked the value 0.5, since we know from TDT [153] that the truth-lie base-rate should be 50% and $t = 0.5$ would reflect this base-rate.

I have performed the independent-samples t-test to compare the significance between the averages for Deceivers and Cooperators. For both $s = 1000$ and $s \xrightarrow{\infty}$, differences were very significant with $p < 0.001$. The results are the following: 1) $s = 1000$: [$C: 0.0$ (SD=0.001), $L: 0.0$ (SD=0.001), $Dec: 0.999$ (SD=0.0)], see Fig. 7.23; and 2) $s \xrightarrow{\infty}$: [$C: 0.0$ (SD=0.001), $L: 0.0$ (SD=0.001), $Dec: 0.999$ (SD=0.0)], see Fig. 7.24.

Figures 7.21 and 7.22 show that Deceivers are quite quick to invade in a single run of the game where $T = 10^5$, and that strong social learning (imitation strength) $s \xrightarrow{\infty}$ enable Deceivers to do so even quicker. This effect of imitation strength is also represented in Fig. 7.20. It seems that the better self-interested agents are at learning social behaviour, the more they tend to adopt deception as their favourite social strategy when their alternatives are to cooperate or not to participate in social interactions.

In terms of trust, it seems that no matter how much trust agents have in each other, it does not affect deception, that is Deceivers still invade and dominate (see

Fig. 7.25). On the other hand, the communicative skill of the deceivers $commSkill$ determines their success. If agents are in the truth-default state $t = 0.5$, then Deceivers just need to be better than chance at deceiving $commSkill \geq 0.5$ in order to dominate Cooperators (see Fig. 7.26).

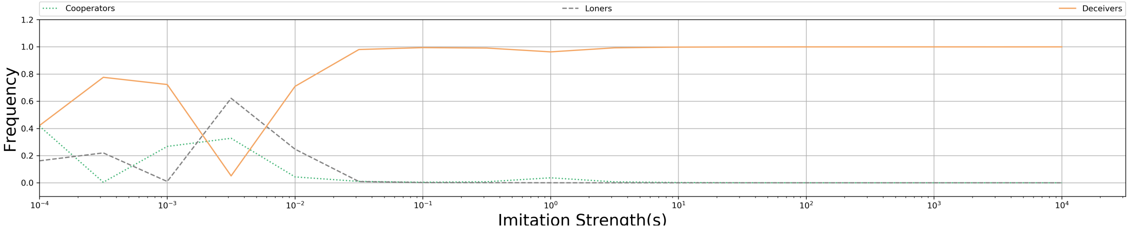


Figure 7.20: Effect of social learning s in Deceivers vs Cooperators.

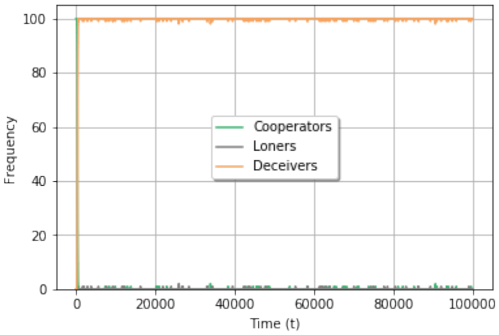


Figure 7.21: Population dynamics for $s = 1000$.

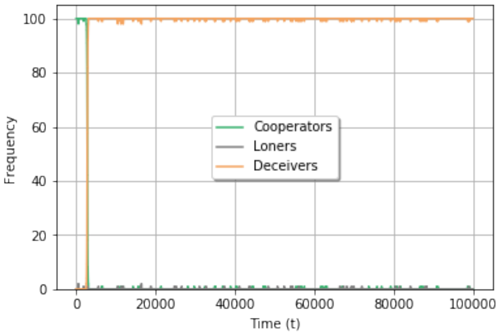


Figure 7.22: Population dynamics for $s \xrightarrow{\infty}$.

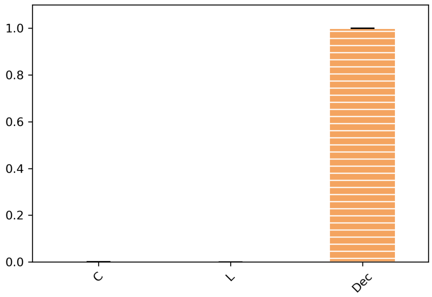


Figure 7.23: Long-run average frequencies for $s = 1000$, $p < 0.001$.

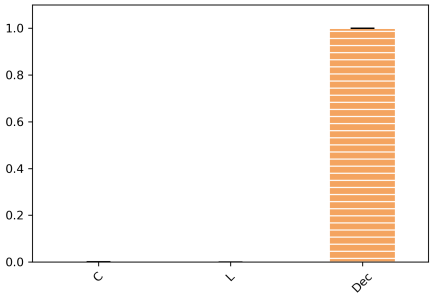


Figure 7.24: Long-run average frequencies for $s \xrightarrow{\infty}$, $p < 0.001$.

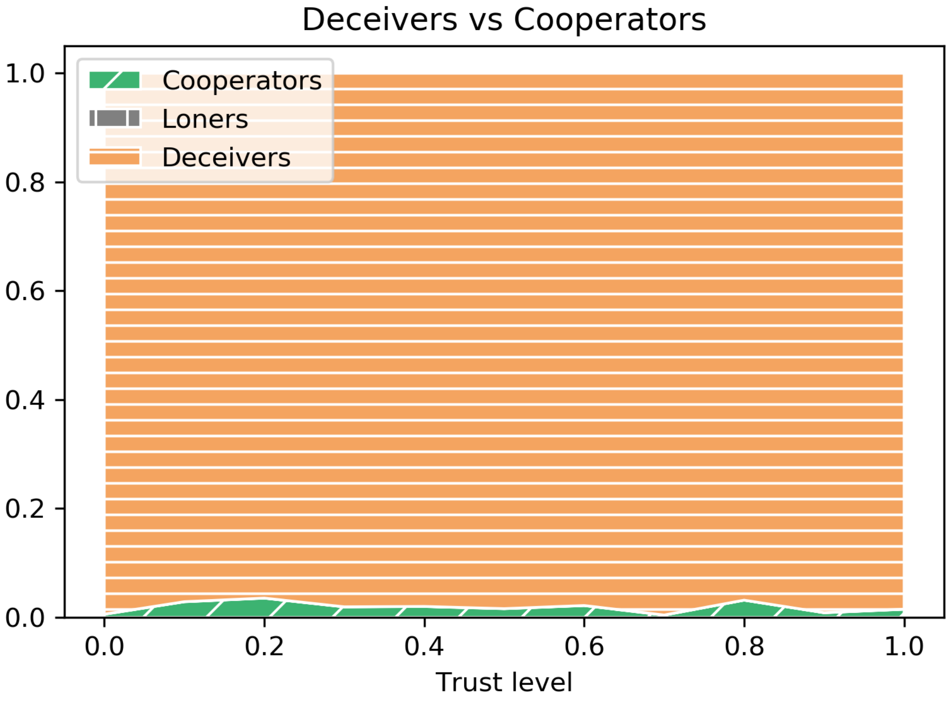


Figure 7.25: Effect of trust.

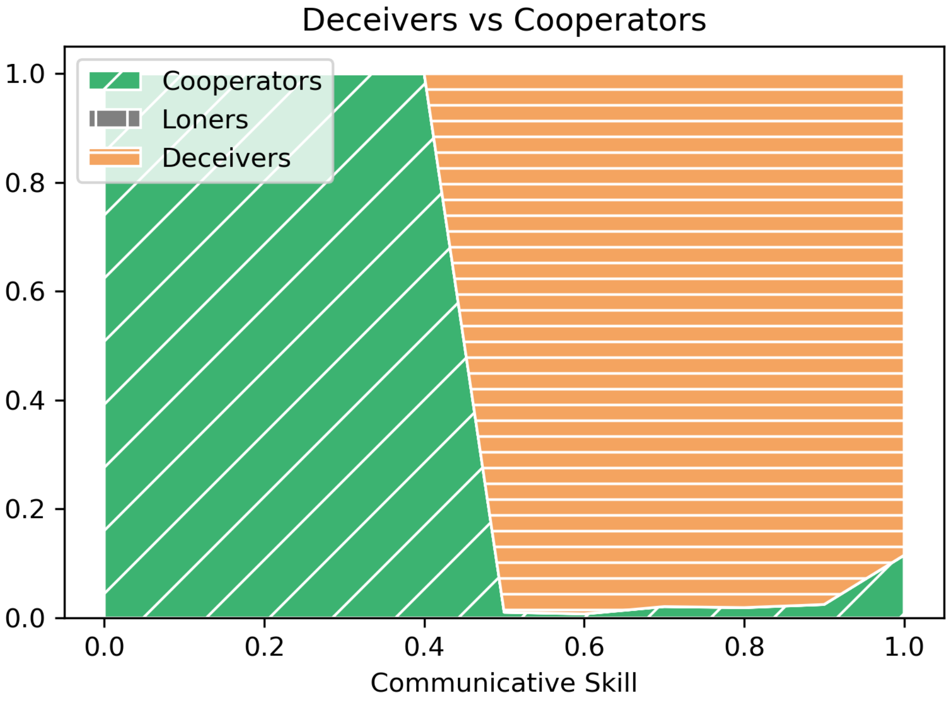


Figure 7.26: Effect of *commSkill*.

7.5 Discussion

Can societies reach cooperation in systems of public knowledge where deception is present? Our results show two possible outcomes. One possible outcome is that cooperation between agents cannot reach high levels in the case of weak or intermediate social learning (imitation strength) (PGGs 2,3,4,5). However, cooperation is slightly promoted if there is a hybrid and decentralised regulatory institution (or system, e.g., the internet) that allows the punishment of Defectors using decentralised methods (which is represented by peer-punishing) and that allows the independent interrogation and peer-punishment of Deceivers (PGG6). The other possible outcome happens in the case of strong social learning, where cooperation between agents can be reached when such a decentralised regulatory system is present (PGG6) (see Fig. A.4 in **Appendix**). What this means in the real world is that agents need to quickly learn how to identify and punish social sources of deception. The effect of strong imitation shows us that if agents learn to quickly adopt other agent's strategies to form a coalition, then free-riding can be suppressed and cooperation reestablished if there exists a system that allows the peer-punishment of defectors and a decentralised interrogation of potential deceivers. Strong social learning, however, does have its downsides if deception is not investigated and punished, such as in PGG2 where Deceivers can become dominant (see Fig. A.3 in **Appendix**).

If the regulatory institution is hybrid, but it is centralised and employs pool-punishment for defection, cooperation can become a viable strategy only at very high costs such as a very high tax for the regulatory institution to exist (PGG5). On a similar note, increases in taxes for maintaining a regulatory institution which does not investigate and punish deceptive behaviour is highly detrimental to cooperation as it promotes deceptive behaviour (PGG2). It is also very likely that cooperation cannot be established if malicious agents are highly skilled at communicating deceptively independently of the type of regulatory institution.

Similar conclusions have been drawn in [1] in the case of *corruption*. Corruption is able to break down cooperation in societies where centralised institutions perform regulation. Decentralised regulatory institutions have also proven to be much more efficient in re-establishing cooperation when corruption is present.

Recently, the European Union (EU) has taken a controversial stance on data sharing [194]. EU believes that in order for it to be able to overtake the US in terms of technological progress, it must incentivise market fairness and technological competition between big and small businesses that own or have access to user data. In order to create this incentive, it must implement and regulate a system in which knowledge is shared between businesses, e.g., in which knowledge is becoming a public good. This is an equivalent of a single market for data. There are major concerns whether businesses will voluntarily commit to contribute to this public data pool or not, whether they stick to their commitment or not, but also whether businesses that commit to contribute would actually contribute with data that they know is truthful or genuine. Can businesses as social knowledge sharing agents reach cooperation or will they decide to free-ride on the public good? What kind of regulation mechanism should the EU implement for ensuring that businesses will cooperate? If future businesses cooperate, then how do they ensure the cooperation of their users, human or artificial, in terms of data sharing? A similar problem has emerged in the COVID-19 crisis with the issue of deploying governmental apps for tracking the symptoms and movement of citizens, in order to manage the pandemic [34].

From the perspective of a PGG as a knowledge sharing activity between social agents, we can conclude from the results of the PGG models that deceptive information can break down cooperation, and, by extension, the trust between agents through the promotion of defection. This results in agents adopting a free-riding strategy either by not sharing information at all (Defection), or by sharing deceptive information, that is similar to sharing *fake news* or *fake data* if we want to use contemporary terms[ix]. However, the impact of *fake news* or *fake data* can be mitigated if regulatory measures are taken. In the case of knowledge sharing, Peer-Hybrid Interrogation would represent regulating knowledge sharing through the following: 1) The case by case demotion of individuals that make use of the knowledge pool, but do not contribute to it (Defection); 2) The case by case interrogation, and where necessary, demotion of knowledge sources (social agents in this case). This suggests that deception can be countered, but the challenge is to identify what mechanisms in the real world can play the role of Peer-Hybrid Interrogation[x].

Not addressing the issue of deception through regulation in the sharing of public knowledge can lead to bleak outcomes in hybrid societies. In [106] the authors discuss what they call the *Tragedy of The Digital Commons* (TDC), that represents Hardin's *Tragedy of The Commons* applied to the digital sphere of information and knowledge sharing (see §2.5.3). An important concept presented by the authors in [106] is the one of *exploitation* and *information pollution* of the *Infosphere* [xi]. The TDC arises when the Infosphere is considered the as an environment and a public good, because it can be exploited and polluted by its users in a similar way to that in which a physical public good would be. An example of exploitation would be excessive bandwidth usage, and an example of pollution would be spam or junk-email. Since users of the Infosphere also exchange knowledge, e.g., tweets, posts, the content of their e-mails and messages, etc., the TDC can also be applied to knowledge exchange. This is where online deception, fake news, misinformation and disinformation play a big role in the exploitation and pollution of knowledge.

[ix] Speaking of fake news, another form of deception is to call out truthful news as fake news, a deceptive strategy that has become increasingly popular amongst some contemporary political figures.

[x] One might hope that this role would be played by the media, but, to continue our commentary on the political zeitgeist in which this chapter was written, the media has largely failed in its duty in this regard.

[xi] For example, the cyberspace. However, the Infosphere is not limited to online environments, see [87] for a detailed description of the Infosphere.

Knowledge exploitation and pollution can be caused by the deceptive behaviour of agents or coalitions of agents, human or artificial. In terms of PGG models, we can assume that the Deceivers exploit the public knowledge (what is shared on the Infosphere publicly) by accessing and using the information that is publicly available, while also pretending to contribute to this public knowledge. Remember that our Deceivers pretend to be Cooperators. However, the information Deceivers contribute with can be considered untruthful (fake news, forged knowledge etc.). The advancement of AI could lead to increase the risk of TDC, as machines that have the necessary capabilities to deceive and learn from social interactions, will eventually adopt deceptive knowledge sharing behaviour to better adapt as agents of a society. Social media platforms which centrally regulate the publicly shared knowledge, such as Facebook and Instagram, are systems in which deceptive behaviour (even of simplistic artificial agents) easily emerges, as we have seen for some time [4].

Fortunately, the results of this chapter's PGG simulations indicate that, from an evolutionary perspective, TDC can be avoided in the case of deception if the Infosphere is regulated in a decentralised manner that organises the public knowledge in such a way as to allow agents to voluntarily investigate each other and the information that they share publicly (see PGG6). A real-world example of such a system was the user interaction protocol implemented by Silk Road on the Dark Web [142]. The Silk Road implemented a reputation mechanism through their discussion forums for users to publicly check what information (for instance descriptions of products sold by users) they have previously shared, as well as if the information was indeed genuine, and how their past interactions have turned out. Members were even rewarded for finding out bad vendors on the platform, which can represent the behaviour of our Peer-Hybrid Interrogators. The quality of the reviews as well as the quality of information standards of the Silk Road community, unfortunately propelled it to the undisputed best platform for drug dealing. Research on reputation mechanisms in social multi-agent systems has shown that reputation mechanisms allow agents to form a model of trust of other agents by looking at their past behaviour (what they have previously communicated), which, of course, needs to be observable (public) [245].

A similar initiative to the Silk Road with respect to the knowledge as a public good has been started by the founder of Wikipedia, Jimmy Wales, albeit for higher moral purposes such as helping society fight fake news and not be solely reliant on reputation mechanisms of its members. The initiative was initially launched through WikiTribune [17]. WikiTribune was a news wiki that used crowd-funding to financially support the costs of running a small team of professional journalists that were intended to work collaboratively with voluntary expert citizens to find stories, create content, and fact-check its own work. However, according to Wales, due to issues in the design of the website, WikiTribune had failed to make its community flourish [285]. That is when Wales turned the initiative into a microblogging and social media platform named WT.Social (WikiTribune Social), arguing that the WT.Social could succeed where WikiTribune had failed. WT.Social aims to promote high-quality content and debate among its users, and its format is meant to combat fake news by providing evidence-based news with links and clear sources. The service is advertisement and click-bait free, and runs off of donations from its users, similarly to Wikipedia. Unlike other social media platforms, where users need to first report offensive content and only after the company would eventually decide to remove the reported content, in the case of WT.Social, the community is encouraged to take down material perceived to be violating the network's standards [59].

The philosophical and political concept under which platforms such as WT.Social aim to promote is called deliberative democracy [97], from which what Habermas calls the *Public Sphere* emerges [109]. The Public Sphere represents a fertile ground from which public opinions are formed through knowledge sharing. A Public Sphere that works in an ideal manner represents the foundation on which mediation takes place between state (regulator) and society which permits democratic control of state activities. For the the public sphere to work in an ideal manner, a society must keep a record of state-related activities and legal actions which is publicly accessible in order to allow discussions and the formation of a public opinion. In our current era, the public sphere has become increasingly digitised. Due to this digitisation, e.g., through the digitisation of news and emergence of social media platforms, the formation of public opinion has been both accelerated and scaled up due to the increasing communicative means and styles that could be employed to reach an increasing number of public members. It has become a *Digital Public Sphere* (DPS) [261].

However, even if the DPS offers more possibilities of communicating and sharing knowledge, it has certainly failed to adhere to principles of rationality and civility proposed by deliberative democracies [243], mainly due to information pollution produced by fake news. The emergence of fake news has enhanced the visibility of the DPS's weaknesses, but unfortunately it has also enhanced its negative affects on the formation of public opinion, making it susceptible to the TDC despite the recent efforts made by projects such as WT.Social and WikiTribune to mitigate these effects. Moreover, the increasing hybridization between human and artificial societies increases the risk of propagating these effects even further, through the development of autonomous artificial agents that not only possess the ability to meaningfully engage in deliberation, but that also possess deceptive intent. Such advancements in AI would imply going beyond the current threat of AI bots and tools used by human agents for fake news propagation and generation, which are based on machine learning techniques. These are merely tools in the hands of human agents, and these tools do not possess deceptive intent. What I am referring to are neither AI tools nor artificial agents that just learn a deceptive policy and mindlessly apply it, but I refer to artificial agents that are able to truly engage in deliberation on the digital public sphere. These artificial agents could perform complex reasoning and apply it to decision-making such that they form their own goals and intentions which they act upon, and by doing this, they could eventually out-think and out-smart humans and other artificial agents when interacting in the public sphere. Remember that our model suggests that even for systems such as PGG6, where cooperation can be re-established in public knowledge sharing systems where deception is present, if the communicative skill of deceptive agents is high ($commSkill = \xrightarrow{1}$), then the system fails to promote cooperation. Deception in the DPS would then evolve, such that it would become *de-antropomorphised*, as human agents would not be the only agents with deceptive intent and capable of truly deceiving others.

Perhaps a future solution to aid the moderation and content checking of platforms such as WT.Social will emerge from where their potential difficulties will arise, namely the further advancement of AI. However, it would not be sufficient to just advance AI deception detection by tweaking truth-bias and skepticism levels to detect deception as it is currently done in verbal and non-verbal cue-based deception detection AI research (see §2.2.6). Cue-based approaches in AI deception research could potentially lead to what is called in the Psychology of deception *confirmation bias* [30]. To illustrate the confirmation bias, Bond describes the notorious *Othello Error* in [30]. Othello believed that Desdemona, his wife, was cheating on him with another man. Othello's fallacy was that he took into consideration only the behaviour a guilty person would exhibit, without taking into consideration all the other cues that might have falsified his beliefs, such as the fact that desperation causes individuals to exhibit some of the behaviours a guilty person would exhibit (see Chapter 2). In the case of complex reasoning artificial agents, I have shown in Chapter 6 that high levels of skepticism in communicative social interactions between artificial agents could lead to deception even when the deceiver agent's communicative skill is low. This type of artificial agent deception represents the special case of unintended deception where the deceiver does not act deceptively because it wrongly estimates that deception would fail, but the interrogator (the deceiver's target) is so skeptical that it caused it to believe that the deceiver has attempted deception, and thus the interrogator is caused to infer something that is false from a truthful message.

A possible solution from the advancement in AI would instead be the actual development of artificial agents capable of complex reasoning to address the issue of deception. Apart from the potential risk of being capable of deception themselves, complex reasoning artificial agents could play the roles of editors and investigative journalists (or to assist or engage with humans that fulfill these roles) which edit and moderate social networking platforms, as well as interact with users to produce high-quality content to increase the public's knowledge and to mediate the formation of public opinion. These artificial agents could potentially neutralise deceptive ones by matching their communicative capabilities, e.g., keeping $commSkill < 1$ in a PGG6-type of system. However, in order for artificial agents to be able to perform these roles that are beneficial to society and cooperation, much needs to be done in terms of AI research. To do so, we first need to enable artificial agents to understand deception (it takes one to know one) by successfully engaging in social interactions. We must mainly enable them: to form and to reason about arguments, to explain reasons behind decisions such that they can be held accountable by the community, and to engage in meaningful dialogue with the community and its members in a democratic manner. The areas of argumentation, human-agent interaction, explainable AI, and multi-agent systems will prove to be crucial in the future research and development of these types of artificial agents.

7.6 Conclusion

In this chapter I have presented an approach inspired by the Machine Behaviour (MB) paradigm [213] (see §2.5) in order to illustrate how distributed deception Type III can be modelled (see Chapter 3). This study shows a proof of principle of (i) how deception can emerge and destabilise cooperation in societies where centralised and decentralised regulatory institutions/systems exist; and (ii) how cooperation can be re-established in such societies. Moreover, the research that I have described in this chapter informs us that there are indeed risks of machines to adopt deceptive behaviour from social interactions with other agents, enhancing the negative effects that lead to a Tragedy of The Digital Commons (TDC). However, this research also points towards a potential solution to avoid a TDC that comes from (i) avoiding the adoption of centralised systems for regulating public knowledge, and instead (ii) aiming for the adoption of a decentralised system for regulating knowledge as a public good in which agents can investigate and check the publicly shared knowledge as well as peer-punish the deceivers and defectors. Some real-world examples of these decentralised systems are platforms that implement reputation mechanisms, where agents can check what others have previously communicated, and platforms such as social networks where high-quality content and fact-checking are promoted, and where the source of the content is transparent (made public). The philosophical and political concept under which these platforms fall is called deliberative democracy [97]. Such platforms promote what Habermas calls the *public sphere* [109]. I have also hinted at the fact that a possible solution for the moderation of the public sphere is the development of complex reasoning agents such as the ones modelled in Chapters 4, 5, 6 that are able to both deceive and reason about the minds of others by meaningfully engaging in social interactions.

Chapter 8

Towards an MAS Framework for Deception Analysis

In this chapter I evaluate the work in this thesis and I propose the creation of a future MAS framework for deception analysis.

The need of better understanding deception is becoming increasingly pressing with the advancement of AI and of hybrid societies. Apart from describing in computational terms how deception can be represented, as I have done in Chapter 3, in this thesis I have modelled deception using two main approaches. The first approach throughout Chapters 4, 5, and 6 addresses the complex reasoning and social interaction capabilities of agents to deceive by forming models of other agents' minds and to use practical reasoning over these models. The second approach in Chapter 7 draws upon the paradigm of Machine Behaviour and addresses the social dynamics of deceptive agents in governed societies using an evolutionary game-theoretical approach. In this chapter, I reflect upon the properties exhibited by the models that resulted from these approaches in order to evaluate the models, and, based upon my reflections, I discuss how the MAS framework for deception analysis introduced in §1.4 can account for levels or degrees of explanation to satisfy analysts.

8.1 Introduction

Theories of human deception, such as TDT, IDT, and IMT2, outline general rules or principles of how deceptive interactions play out. Understanding the mechanics of human deception allows us to reproduce and scale up different configurations of deceptive agent interactions artificially by designing reasoning and communication protocols and components for tools to study them inside an MAS. At the beginning of this thesis, in §1.4, I have argued that the two approaches I use to model deception represent deceptive interactions as artificial mind-games at two different levels of abstraction in an MAS. The first level of abstraction shows us the cognitive mechanics responsible for deceptive communication between practical reasoning agents, while the second level of abstraction shows us how the deceptive behaviour of agents emerges in large and complex hybrid societies.

In §1.4, I have also discussed a set of methodological questions regarding the design of the models of deception in MAS. The models in Chapters 4, 5, 6, and 7, have been designed with those questions in mind. The application of the methodological questions has resulted in models that are able to (i) represent various components of deception and (ii) that exhibit useful properties for the study of deception.

8.2 Evaluation of the Models

Reflecting on the models, I have observed that they have certain properties and that they are able to represent a set of components that are particularly useful for the study of deception (see Table 8.1). These are: 1) The property of Explainability; 2) The ability to represent Unintended Deception; 3) The ability to represent Uncertainty; 4) The ability to represent Deception Detection; 5) The properties of Implementability and Transparency; 6) The property of being Theory-Grounded.

Regarding the first approach to model deceptive agents using cognitive modelling

Model / Properties	1	2	3	4	5	6
* Chapter 4	✓	-	✓	-	✓	✓
Chapter 5	✓	-	-	-	✓	✓
Chapter 6	✓	✓	✓	-	✓	✓
Chapter 7	✓	✓	✓	✓	✓	✓

Table 8.1: Comparison of the models in terms of their properties and representation of components for the study of deception. Note that the model in Chapter 4, even though it does not include deception, it is the underlying ToM component for the model in Chapter 6.

techniques for BDI, we can notice that the modelling of ToM in Chapter 4 has a strong impact of how deception is illustrated in Chapter 5 compared to Chapter 6. In Chapter 5 a dishonest agent works under the assumption of complete certainty and does not engage in mental simulation to determine the optimal deceptive action. Thus, it does not take into consideration factors such as levels of trust, communicative skill and confidence in ToM, or other contextual factors. However, in Chapter 6 the model of deception is refined by applying the TT and ST ToM from Chapter 4 along with the components of IDT, IMT2 and TDT, that allowed me 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.

Regarding the second approach to model deception in Chapter 7, we can notice that by integrating TDT and IDT with MAS approaches it is possible not only to represent the deceptive behaviour of agents in complex interactions, but it is also possible to find out how to reduce deception's negative effect on social cooperation.

8.2.1 Explainability

In terms of **explainability**, it is necessary if our aim is to explain how deception happens or does not happen, and why. An explainable model of deception allows us to evaluate whether deception takes places, and if it does or does not, it enables us to explain why and under which conditions it does. Models of deception need to be explainable in order for the generation of explanations based on the models to be meaningful and informative. A good explanation of the phenomenon can then be used to see if deception can be prevented or mitigated in different contexts.

In the first approach to model deception from Chapters 4, 5, and 6, explainability is directly related to the BDI cognitive architectures of the agents and the operational semantics these agents use for communicating. Their architecture and the operational semantics allow us to track how the agent's beliefs are used to achieve their communicative goals in various contexts. Moreover, the approach is based on practical reasoning, that is very intuitive for humans to reason about.

In the second approach to model deception from Chapter 7, explainability is related to two elements of the model, namely (i) the simplicity of the evolutionary mechanism that agents use to learn and adopt social behaviour, and (ii) the approach of Machine Behaviour to study complex and emergent social behaviour. The six PGGs represent different systems or social contexts in which we study how deceptive behaviour is learned or adopted by agents in a large population by throwing in different regulatory mechanisms (interrogation and punishment strategies). From a counterfactual perspective these PGGs represent six possible worlds in which deception may or may not emerge. By analysing the dynamics of each PGG, we can explain when, how, and why or why not deception emerges.

8.2.2 Unintended Deception

Unintended Deception might happen when an agent might not attempt deception, but the consequences of its communicative acts result in its interlocutor being deceived. It is important for models of deception to be able to represent such unintended consequences, as these can be critical for accountability and regulation. If we are to design regulations for the social behaviour of autonomous system, for instance, 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. Intentions, or lack thereof, might play a salient role in such cases.

In the first approach, precisely in Chapter 6, unintended deception is explicitly represented through the BDI architecture. We can identify that even if an attempt at deception is not made, the target agent is still deceived (it reaches a false belief through its own reasoning and the influence of its trust in the interlocutor).

In the second approach form Chapter 7, unintended deception is not explicitly represented. We do not have here the explicit representations of the internal cognitive dynamics of the agents, and by extension there is no representation of intention. However, intentions are implicitly considered. The model does account for cases of unintended deception on a macro-level, as it takes into account the influence of communicative skill of an agent along with the trust levels of a society which influence a deceiver's risk of getting caught by an interrogator.

8.2.3 Uncertainty

The estimation of whether deception was, is, or will be successful is always performed under conditions of **uncertainty**. This is especially the case in communication, where it is 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 [152], in cases of potential deception this is not the case. Degrees of uncertainty, trust, and the ability of rationally bounded agents to deal with uncertain knowledge are all factors that enrich our representation of the real world. Hence, good models of deception have to take these factors into account.

In the first approach, the ToM mechanism introduced in Chapter 4 is added to the mechanism for estimating deceptive success in Chapter 6. The deceiver takes into account its uncertainty regarding the degree of confidence in ToM to estimate the success of deception under the uncertainty introduced by trust levels and communicative skill. In the second approach from Chapter 7, uncertainty is considered in the payoffs for both deceivers and interrogators, as well as in the influence of social learning (imitation strength).

8.2.4 Deception Detection

Even though **deception detection** has not been the main focus of the modelling approaches in this thesis, it was also a useful component to be modelled. The first approach only focused on modelling deception itself. However, in the second approach from Chapter 7, the interrogation strategies in the PGGs have been designed to represent deception detection.

While the BDI practical reasoning level in the first approach can be used to explain how deceivers manipulate information to achieve their goals, the meta-representation of the interplay between deceivers and their detectors (interrogators) from Chapter 7 can be used to explain what strategies or mechanisms can reduce deceptive behaviour in general, without going through the practical reasoning for deception detection. However, the downside is that the second approach does not explain how deception is detected by a single agent at the level of its cognitive architecture.

8.2.5 Implementation and Transparency

Implementation is also desirable in an AI-based approach, but might not necessary for modelling deception, or any other social phenomenon. However, demonstrating an implementation of a model helps others to use it for studying different agent-based setups and scenarios of social interactions. Implementation also improves the **transparency** of a model, increasing the model's accessibility and reproducibility through its code.

The implementation of the first approach has been done in the Jason AOPL [33] through which the beliefs, goals, and intentions, as well as the reasoning that drives the behaviour of the artificial agents are intuitively represented in code and are transparent to the engineer/designer/analyst. Further research into the behaviour of the deceptive agents that model other minds can be done by using this implementation. The second approach has been implemented in Python 2.7, and the code can also be used for further research in order to explore other types of regulatory mechanisms for deception in hybrid societies, by implementing new strategies to be selected by agents through social learning.

8.2.6 Theory-Grounded Modelling

Last, but not least, both modelling approaches of deception are **Theory-Grounded**. Deception is a social and cognitive phenomenon that has been empirically studied in Psychology and for which several theories, models, and methodologies of communication, deception, and deception detection have been proposed and tested. The engineering of artificial agents to study deceptive interactions must be based on such theories if we want to gain a meaningful understanding when we use them in simulations or other contexts. TDT, IMT2 and IDT have been selected for this purpose because they have survived scientific scrutiny, as in they are sound, they are testable, and have empirical basis. IDT might be the outlier of the three theories as it is a theory about deceptive cues, however IDT is not as cue-reliant as other alternative theories. IDT's main argument is that deception is about interaction and thus it is very compatible with the MAS paradigm.

As I discuss in §2.1.2, there are theories that might turn out to be pseudo-scientific. There is a considerable risk that AI researchers might build models based on such pseudo-scientific theories and that these models turn out to be sponsored and used in very sensitive and controversial contexts such as border control (see §2.2.6). This might pose a significant threat to the ethical use of AI in population control and law enforcement. Such theories should be avoided, or at least their limitations need to be heavily considered.

8.2.7 Deceptive Storytelling Agents

Even though the models evaluated in this chapter present most of the desirable properties above, there is still work to be done, especially w.r.t., deception detection and counter deception (see §3.4). While the meta-representation between the interplay between deceivers and interrogators in Chapter 7 accounts for deception detection, it does not do so at a practical reasoning level.

One way to address counter-deception is to model artificial mind-games between deceivers and interrogators that have storytelling abilities. In §2.4.5 I have described the work in AI on storytelling. Let us recapitulate that in the context of AI, **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 [23, 269].

Regarding deceptive agents, they might use these narrative abilities to their own benefit. They could, for example, deliver arguments to build fictitious stories that compel juries into absolving them of a crime. Regarding counter-deceptive agents, they could build interrogation techniques to force the deceptive agents to give away elements or arguments that would weaken their alibi stories and that would finally cause their attempt at deception to fail.

The adoption of a hybrid-based storytelling approach to model artificial mind-games may or may not adopt the BDI representation for ToM introduced in Chapter 4. This decision might be down to the modeller or MAS engineer. However, I believe that the hybrid ToM model (TT and ST ToM) should be kept in order to represent how the known arguments of the opponent (TT ToM) can be used to generate and select new complex arguments through the mental simulation of the target (ST-ToM). This ToM mechanism would allow, for instance in a dialogue interrogation game, both deceiver and the interrogator to simulate how the dialogue would evolve based on the selected arguments. For illustrative purposes, in [240] we have described an example of how such a mechanism could work in an interrogation dialogue game in the context of deception.

In the future, models of deception based on the idea of storytelling agents could be used to study how agents might behave unethically by avoiding the principles of accountability, responsibility and transparency [68], but also what type of interrogation techniques, or what type of counter-deceptive agents, would cause these deceptive agents to give away their malicious intent.

8.3 An MAS Framework for Deception Analysis

Previous methods, such as ACH and its derivations, used by intelligence analysts to detect deception while reducing cognitive bias are tedious (see §2.1.1). A good MAS approach to model deception can be helpful to understand deception in complex systems and could also help users to perform less tedious analyses of deception.

Remember that in §1.4, in order to link the needs of Intelligence Analysis with the modelling of deception in MAS, I have introduced the idea of an MAS framework to address different levels of abstraction based on different MAS models. This MAS framework would allow the engineering and creation of MAS tools for intelligence analysts, as mentioned in [239]. Apart from linking together the different levels of abstraction represented by the models of deception, an MAS framework could also address, according to the needs of the user, the degrees of explanation w.r.t., the models.

I remind the reader of the previously introduced desiderata for such an MAS framework: 1) Representational power; 2) The automation of conductive reasoning; 3) Social interactivity; and 4) Scalability.

8.3.1 The Hybrid-Story Approach in MAS

Another advantage of adopting a storytelling approach in MAS, apart from the modelling of deceptive artificial agents with storytelling capabilities, is that it could improve explanations of MAS behaviour. It would allow humans to reason about deception in terms of narratives of complex cases of deception, by representing counterfactual chains of events along with the agents' interpretations of these chains of events. This advantage has direct implications w.r.t., the explainability of MAS models of deception.

Let us remember that Bex's hybrid approach [21, 22] has been designed to both represent causal chains of a main story and to use arguments to anchor the main story's subs-stories in evidence, a process named *anchoring*, that results in explanations (see §2.4.5). By applying Bex's approach, one does not only find an arbitrary story for a causal chain of events, but is able to select the best main story (out of several viable ones) that is composed of several sub-stories, e.g., perform the inference to the best explanation, by explaining through arguments how the main story's sub-stories are backed by evidence.

In the real world, this is a complex process, especially in the case of dealing with evidence, e.g., collecting it or eliciting it from other individuals, analysing it, validating it. However, this kind of real-world complexity is not present in MAS. This should be obvious, as MAS models and frameworks are abstractions of real-world phenomena. But then, how would one anchor stories in evidence inside MAS?

For instance, a standard MAS framework accounts for three layers of abstraction, namely agents, artifacts in the environment, and agent organisations. All three of these levels offer a representation of a world where evidence can be directly observed, extracted, and analysed. **Evidence** to provide answers for the following questions can be provided with this type of MAS framework: Which agent acted upon artefact X? Which agents interacted with organisation Y? What changes did the joint action J of agents A and B cause in environment E at time Z? Why did agents A and B perform joint action J (what were the beliefs and the reasoning processes that led to action J)? What agent behaviour can we observe inside organisation O? Is it a cooperative behaviour? What are the levels of cooperation? Would this cooperative behaviour inside organisation O be affected or not if we change the agents' confidence in their ToMs of other agents? etc.

Moreover, an MAS framework can be used to explore **sub-stories**, and answer questions such as what happened at time Z between agents A and B in environment E? Did they have a dialogue? What kind of dialogue? What was said during the dialogue? etc.

More or less, the observable behaviour of agents inside an MAS can be used to test whether a potential viable **main story** (a complex hypothesis) about their behaviour counts as an explanation of their behaviour. How well the story is anchored in the evidence can also be observed in the MAS, e.g., how many items of evidence confirm or falsify the story. Moreover, an MAS framework would allow a direct bottom-up formation and anchoring of a story if all parameters of the MAS are fixed. It could also be used to confirm or falsify how certain events in the real world might have unfolded, but more importantly it could explain how they might have unfolded or not if something else were the case.

Also, the hybrid approach to storytelling could be used to find inconsistencies inside the MAS framework w.r.t., the theories that have been used to implement the MAS models. Perhaps there is considerable evidence outside the MAS that indicates that the agents' behaviour should have been different. An MAS framework would allow us to identify what is the underlying cause of the undesired behaviour, at which level of abstraction it happens, and whether it can be addressed inside the model that represents that particular level of abstraction. Subsequently, the respective model can be directly revised according to a new evidence-based theory and re-implemented in the MAS framework.

In conclusion, the hybrid approach can be used to represent stories that are generated based on the agents' behaviour. In the case of an MAS framework for deception analysis, these stories can be explanations for the behaviour of agents that can be offered to analysts. Analysts can then use the explanations to directly perform inference to the best explanation for different complex scenarios, or, if they believe there is an alternative best explanation that has been inferred outside the MAS, they can use this to refine the models inside the MAS framework and achieve the desired behaviour from the artificial agents.

8.3.2 MAS Stories of Deception...or not

Below, I describe how a storytelling approach could be used to enhance the explainability powers of an MAS framework for deception analysis w.r.t., the desiderata for this framework.

Whereas **representational power** addresses different levels of abstraction through the modelling of deception, it does not address the different levels of explanation for analysts. It is regarding this aspect where using a storytelling approach would be most useful to include in the MAS framework.

Once we have explainable MAS models of deception, we can use them to directly observe the links between cause and effect, and event causation prevention. In other words, we could **automate counterfactual reasoning**. Both historians and intelligence analysts, apart from looking at case studies and observing these types of links, they also engage in writing articles, reports, or even entire books to explain what happened, how it happened, and why it happened the way it happened. In other words, they aim to perform an inference to the best explanation. One thing that usually emerges from these reports, is an explanation or a theory that can be interpreted as a story or a narrative.

Following this line of thought, perhaps it would be preferable to explainees (both analysts and the ones who they write the reports for) to be told a story that describes a complex chain of events, one that connects the cause to the effect, narrated in any order that would maximise one's understanding of the events. This is where **social interactivity** could be addressed. According to Miller [177], explanations consist of two processes: (i) a cognitive process, which represents an abductive inference that determines a causal attribution, and (ii) a social process, which represents the knowledge that is exchanged between an explainer and an explainee. The approach to model ToM from Chapter 4 could be very useful here to model the social process between artificial storytelling agents and their interlocutors (explainees). Storytelling agents could model the minds of the explainees in order to deliver narratives that are efficient and that achieve the goal of explaining a phenomenon and, thus, of reaching shared beliefs with their explainees in different social contexts. This approach would improve the models inside the MAS framework w.r.t., models' explainability properties, as it would lead to the design of **self-explainable** artificial agents that are able to explain their practical reasoning and decision making directly to the analyst. Hence, the analysts' work would be reduced, as they would not need to go through the artificial agent's code and interpret it.

Storytelling could also account for **scalability**, especially in Open MAS. A story can represent how a new agent character enters the story, or how the chain of events moves from one location (domain) to another, or perhaps the narrative style of the story allows for referring back to evidence or arguments that have a certain temporal property (using information available at a different time), e.g., something happened in the past that affected the current chain of events, or perhaps if this chain of events continues in this direction, something relevant to the story will happen in the future. Another way to refer to scalability is using the point of view (PoV) of the story, where one can explain a phenomenon from the PoV of a single agent, from the PoV of multiple agents that can either have been actively involved in a context or have only been passive observers, or perhaps from the PoV of a Big Brother observer, one that has access to multiple viewpoints of agents, as well as extensive knowledge of the chains of events and the domains in which these events have happened.

8.4 Conclusion

In this chapter I have evaluated how this thesis has addressed deception as part of a complex system of agents that interact in different contexts using two modelling approaches in MAS. The resulting MAS models enable us to explain and understand the potential dishonest behaviour of artificial agents and what causes it.

Deceptive behaviour, whether it happens in a virtual, physical, or hybrid world, needs to be understood as part of a bigger context or system, and not just by reducing it to the analysis of the observable behaviour. An MAS approach to deception does not just allow the analysis of behavioural data, but also allows one to represent different components of deception and to play with the mechanisms of the system that produces the agent behaviour in order to better understand what causes it or what prevents it from being caused.

In conclusion, this thesis' research direction can be further improved with the creation of an MAS framework that is capable of addressing agent deception in a manner that emulates and enhances the critical thinking of intelligence analysts, both in terms of levels of abstraction through MAS models, as well as in terms of levels of explanation through applying a storytelling approach.

Chapter 9

Future Directions

With this chapter I conclude this thesis; I summarise the contribution of this thesis and indicate possible future research directions.

9.1 Conclusion

This thesis began with two strongly intertwined questions: 1) *Can we use artificial agents to improve our understanding of deception?* and 2) *How may artificial agents deceive?*

In asking the first question, I have made the assumption that we can use agent-based modelling to gain insights into how deception plays out, mainly, to gain insights into event causation and event causation prevention, where by event I mean the occurrence of deception. This led me to ask the second question.

In asking the second question, I have made the assumption that *deception is an intentional process of a deceptive agent seeking to cause another agent to believe something is true that the deceptive agent believes is false, with the aim of achieving an ulterior goal or desire.* I have developed this view of deception from Levine and McCornack's work on deceptive communication in humans as I hope to have clarified to the reader throughout the thesis [153, 172, 154]. The work described in this thesis has been conceptually based mostly on their work, as well as Buller and Burgoon's [38], but to a lesser extent as IDT is over-reliant on deceptive cues, something which the work presented in this thesis aims to avoid. The crucial concept adopted from IDT is deception as social interaction.

In this thesis I have explored deception as interaction between artificial agents. I have been mostly interested in how to model the exchange of knowledge between agents in a way that is meaningful to deception research.

The first part of my thesis described, in Chapter 2, the relevant literature regarding both human and machine deception covering both internal cognitive mechanisms, as well as socio-cognitive mechanisms and large-scale evolutionary systems. Regarding the literature on human and machine deception, I have given a critical account on why cue-reliant approaches and data-oriented approaches are severely limited to meaningfully model deception. I have also explained why these limitations have led me to pick TDT, IDT, and IMT2 as conceptual pillars of my work. Regarding the mechanisms used to represent complex reasoning and interactions in MAS and the evolutionary mechanisms for ABM, they have provided the necessary techniques to meaningfully model deceptive agents and the social interactions in which they take part.

The second part of my thesis covered the definition of deception in computational terms and the modelling and engineering of socio-cognitive agents. I have defined what computational deception is and described a taxonomy for it in Chapter 3. In Chapter 4, I have shown how to model ToM using BDI and speech acts in MAS. I have shown that agents with ToM can reach states of shared beliefs with other agents more efficiently than agents that do not have this capability and that they can do so even under the conditions of uncertainty of communication. I have also shown how shared beliefs can be used for efficient task delegation in different contexts, by taking into account the different skills, knowledge and preferences of the agents that are communicating with each other. Moreover, this approach on Artificial ToM can be used by agents to derive new knowledge by simulating different interaction outcomes in various contexts. The agents can simulate what-if scenarios to see what the other agents would say or not say and what they would come to believe or what they would not come to believe if they were to interact in various social contexts.

The third part of my thesis covered the modelling and engineering of dishonest and deceptive agents. In Chapter 5, I have formalised and represented three types of dishonest behaviour that a socially-enabled artificial agent can execute, namely lying, bullshitting, and deceiving. I have explained, through a running example, how these behaviours differ from each other in terms of what the agent communicates, what it knows, and what it knows about the interlocutor, e.g., its ToM of the interlocutor. In Chapter 6, I have modelled deception between two agents in a question-answering game, adopting a richer representation for the agent architectures. This model follows the principles of TDT, IMT2 and IDT. In terms of IDT, the model represents an interpersonal interaction between two agents considering the social factors such as truth-bias and communicative skill. In terms of IMT2 integration, the model takes into account the fact that agents employ the same reasoning mechanisms for both lying and truth-telling, thus the cognitive load of deceptive agents does not differ between the two linguistic behaviours. In terms of TDT, the model takes into account contextual information that is available to the agent, such as what information is available in that context, what can be said and not said, and how performing a speech act compared to another might make the interlocutor infer a false belief.

The fourth part of my thesis explored how deception and machine deception influences society and what are the potential solutions for addressing deception in complex social systems. In Chapter 7, I have looked at deception from the evolutionary perspective of Machine Behaviour. More specifically, I described how different regulatory systems can be influenced by the deceptive behaviour of agents, human or artificial. Using this approach I have shown how different regulatory systems can influence the large-scale behaviour of self-interested agents over time.

In Chapter 8, I have evaluated the work I have done in this thesis. Also, based on the properties exhibited and the components of deception represented by the models, I have proposed the idea of working towards an MAS framework for Intelligence Analysis that satisfies a set of desired principles for the study of deception in complex systems. Having in mind the contributions of my thesis, I have offered a description of what can potentially be done towards achieving this MAS framework.

9.2 Future Work

This research branches out into several lines of potential future work, some of which I have already mentioned in the respective chapters. Therefore, I describe them according to their potential short-term, medium-term, and long-term goals.

In the **short-term**, there are several additions that can be made to the models from Chapters 4, 5, 6, and 7. For the model in Chapter 4, one could introduce an environment layer, such that agents can reason about other's beliefs regarding changes in environment. These changes in the environment can represent subtle communicative acts that can be easily misinterpreted by socially-aware agents. Regarding the modelling of dishonest behaviour in Chapter 5, one can define and formally represent other types of linguistic behaviour, such as half-truths, sarcasm and metaphorical or fictional uses of language, in order to contrast it with the properties of deception. Regarding the model in Chapter 6, one could potentially explore how intended and unintended deception takes place if there are more than two agents interacting in the same social context. This could be done through an implementation in Jason or in any other BDI-based AOPL of an MAS with an arbitrary number of agents, as I have already described in Chapter 4 the implementation of the reasoning mechanisms for the social agent architectures and in Chapter 6 the deceptive reasoning mechanism. Finally, regarding the several PGG models in Chapter 7, one could introduce a graph component in order to represent how the behaviours of agents in the six PGGs are influenced by the structure of their social networks over time instead of their society's population composition.

In the **medium-term**, the work I have presented in my thesis can be used for improving Intelligence Analysis. This could be done through an MAS framework for deception analysis, as I have outlined in Chapter 8. This could help one represent the distinction between evidence items and pure argument components of a story, which is important when we deal with the concept of *anchoring* [21] and how agents might use different anchoring strategies in different social legal, or security contexts. It is absolutely crucial for the MAS models of framework to be developed with a grounding in solid theories of deception and that pseudo-scientific approaches, such as the ones that stemmed from Paul Ekman's research [75, 76], to be avoided in its development stage. Even if this is a medium-term goal of the research presented in this thesis, it cannot be done by a single individual, because (i) specialist knowledge is required for the multitude of components involved in its design, (ii) because its modular tools have to be continuously tested and improved based on user needs and software engineering needs, and thus it involves a multidisciplinary research and development group approach.

In the **long-term**, the work in this thesis could be used towards building socially-intelligent artificial agents that are able to engage in meaningful conversation with their interlocutors. In this thesis I have laid some of the foundations for completing the work proposed by Charles L. Hamblin in [112], namely on *"How to Build a Machine Worth Talking To"* [i] Hamblin's idea was influenced by Alan Turing's perspective on linguistics, namely that of the problem of mechanising dialogues [272]. Hamblin's book proposes a theory of categorising and using an agent's or a machine's

[i] Philip Staines edited Hamblin's manuscript following his death. What resulted three decades after Hamblin's death is the book *Linguistics and the Parts of the Mind: Or how to Build a Machine Worth Talking to* [112].

commitment stores to deal with three main components of a dialogue, namely facts, acts, and sentiments. Most importantly, however, in Chapter 8 of his book, Hamblin stresses the ability of machines to model other minds:

> "[...] we have noticed that, in order to carry on an intelligent conversation, a speaker needs to know alot about his hearer's backgrounds. He needs to have an appreciation of how his hearers think and feel, what he can take for granted that they will know and how he can expect them to react.[...] I want to maintain, in short, that our linguistic behaviour indicates a quite particular facility at modelling other people's minds, and imagine the reactions of real or hypothetical people under wide ranges of circumstances [...]" [112, p. 114-115]

Hamblin deemed this absolutely necessary for machines to carry out "intelligent conversations", but this is missing from the theory he out laid in his book.

In this thesis I have proposed an approach for agents to model other agents' minds, namely that of Artificial Theory-of-Mind in Chapter 4 of this thesis, and I have shown how this approach can be used to model the complex reasoning behind deceptive communication. This approach can be considered the missing piece that Hamblin mentioned in his theory of commitments, that which enables machines to reason about what Hamblin calls "*hypothetical commitments*". It is, at least to me, fascinating how the problem of machine deception is so strongly tied to the problem of intelligence as defined both by Hamblin and Turing. While Turing argues that in order for a machine to be intelligent, it must cause the interrogator/listener to believe something that is false (that is to trick it into thinking the machine is human), Hamblin argues that for a machine to be intelligent, then the machine must be capable of conversing intelligently, but do do so it must be able to model the mind of the interrogator. To continue their line of thought, in this thesis, I argue that for machines to be able to deceive the interlocutor, and hence prove that they are intelligent, then they must be capable of modelling other minds. Another line of future work would be to explore the relationship between deception using Theory-of-Mind and Mikhail Bakhtin's literary theory on polyphony and unfinalisability. This would indeed be a research endeavour that would be more appealing to philosophers of language and literary theory, but it might be very relevant for designing and testing ethical AI systems. According to Bakhtin, the world is open and free and no amount of dialogue and explanation, however detailed and complex, can never really account for the whole truth. Thus, dialogue is unfinalisable [183]. Dialogue, Bakhtin argues, is also polyphonic, as it can be used similarly to how Dostoyevsky's characters use internal dialogue to make sense of the world and of themselves in the world, e.g., to find meaning in complex systems. Bakhtin claims that the characters themselves become autonomous in terms of their world view and the way in which they argue for their world view in dialogues (internal or external):

> "In the consciousness of the critics, the direct and fully weighted signifying power of the characters' words destroys the monologic plane of the novel and calls forth an unmediated response-as if the character were not an object of authorial discourse, but rather a fully valid, autonomous carrier of his own individual word." [13, p. 5]

This polyphonic view on literature seems quite aligned with the paradigm of MAS, where agents of a model act autonomously according to their world view. Hence, an AI system might describe the world in a dialogue (internally or by exchanging arguments with other agents), and it could make use of the polyphonic style where it would model the multiple voices of the agents it engages with and which are relevant to the dialogue (all voices are valid according to Bakhtin) [12]. To check if it is being deceived, or if it is possible to deceive, the AI system might model a polyphonic story taking into account these valid voices of others. The model of a polyphonic narrative could, therefore also be used in the development of storytelling for argumentation, in which agents that stand for the sources of arguments are not reduced to objects of the model, but are autonomous subjects;

> Furthermore, the very orientation of the narrative-and this is equally true of narration by the author, by a narrator, or by one of the characters-must necessarily be quite different than in novels of the monologic type. The position from which a story is told, a portrayal built, or information provided must be oriented in a new way to this new world -a world of autonomous subjects, not objects. [13, p. 7]

This capability seems to relate to the work I have done in this thesis, and it also seems to link very well with Hamblin's concept of a machine *worth talking to.*

In conclusion, much remains to be explored in the areas of deception and machine deception. Or, as Simon Parsons would say "*...many more boxes to be opened...*" [ii].

[ii] Simon once said (pun intended) that my "problem" as a PhD student is that I "love opening boxes", in terms of research. However, what Simon did not know was that I also used to have quite an obsession with collecting Amazon boxes.

Appendix A

Appendix for Chapter 7

In this appendix I describe the details of the PGG components from Chapter 7.

A.1 Long Run Average frequencies for PGG2 vs PGG6

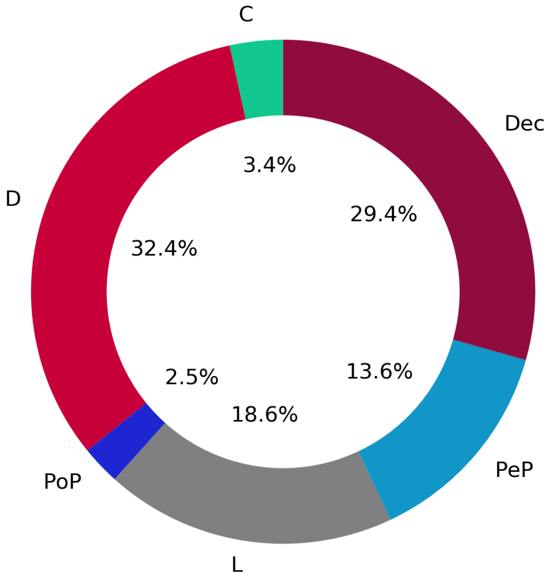


Figure A.1: Long-run average frequencies in **PGG2** for $s \xrightarrow{\infty}$, where D and Dec frequencies are significantly different $p < 0.001$.

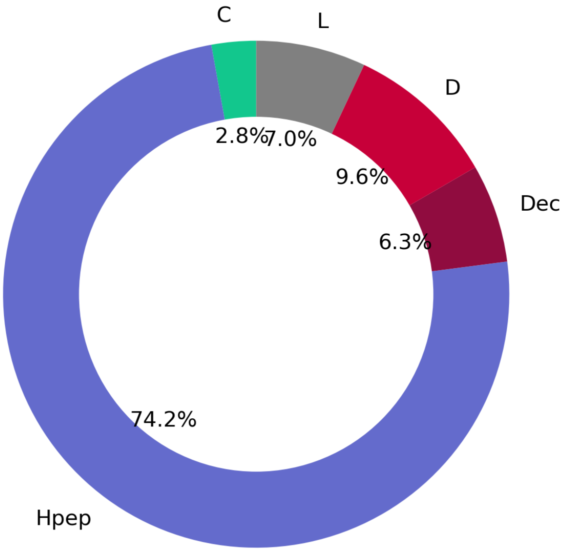


Figure A.2: Long-run average frequencies in **PGG6** for $s \xrightarrow{\infty}$, where D and Dec frequencies are significantly different $p < 0.001$.

A.2 Effects of Social Learning in PGG2 and PGG6

Figure A.3: Effect of social learning in PGG2.

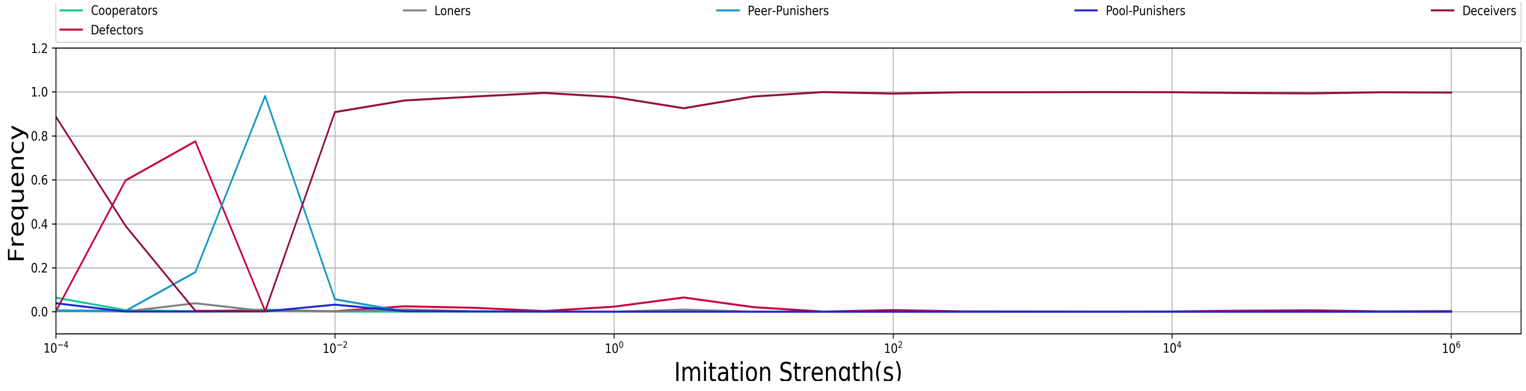
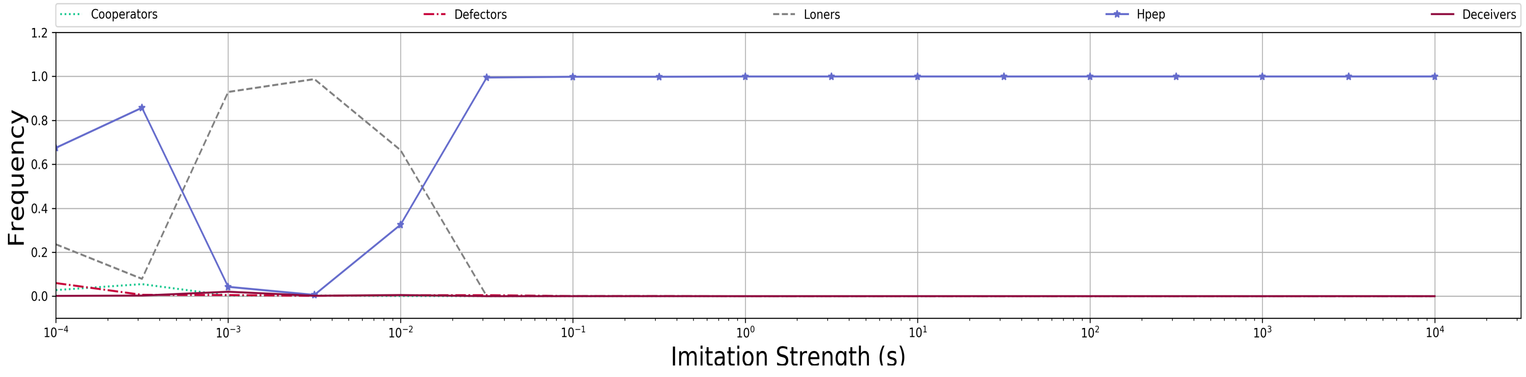


Figure A.4: Effect of social learning in PGG6.



A.3 PGG Components

In a PGG we have the following strategies and set of universal parameters.

1. A set of strategies $S : \{C, D, L, PeP, PoP, Dec, Int, Hpep, Hpop\}$;
2. N = the number of agents in a population to play a PGG;
3. n_{Si} = the number of agents in a population with a given strategy Si;
4. M = the number of agents from a population N that is selected to play a PGG;
5. r = multiplication factor that is always $1 < r < M - 1$;
6. s = imitation strength;
7. c = investment (contributed amount) of a cooperative agents in a PGG.
8. c_{Si} = cost of a given strategy Si;
9. μ = mutation rate;
10. β = cost of paying Punishers to exist;
11. B = pool punishment for Defection;
12. b = punishment for Defection;
13. c_b = cost of punishing a Defector;
14. G = cost of pool punishment;
15. Γ = punishment for Deception;
16. σ = Loner payoff.

17. $P_\sigma = \frac{\binom{n_L}{M-1}}{\binom{N-1}{M-1}}$ represents the probability that all other $M-1$ sampled individuals are Loners.

The payout of a PGG where n_{FR} represents the number of *free-riders* is adopted from [254] and from [1]. This payout corresponds to 7.1 in Chapter 7.

$$Payout^* = c \times r \times \frac{N - n_{FR} - n_L - 1}{N - n_L - 1} \tag{A.1}$$

Because Deceivers are Free-Riders that pretend to cooperate in the PGG, we need to discount them form the payout. Hence our payout is:

$$Payout = c \times r \times \frac{N - n_D - n_L - 1}{N - n_L - 1} \tag{A.2}$$

This payout corresponds to 7.2 in Chapter 7.

A.4 Computing Payoffs

A.4.1 Cooperation, Defection & Punishment Average Payoffs

In a voluntary PGG with first order peer and pool punishment we have five (5) types of agents: Cooperators, Defectors, Peer-Punishers, Pool-Punishers, and Loners.

Cooperator Average Payoff

$$\Pi_C = P_\sigma \times \sigma + (1 - P_\sigma) \times (Payout - c) - \beta \frac{M-1}{N-1} \tag{A.3}$$

Defector Average Payoffs

$$\Pi_D = P_\sigma \times \sigma + (1 - P_\sigma) \times Payout - cost_D \frac{M-1}{N-1} \tag{A.4}$$

Where $cost_D$ depends on the PGG that is played:

1. Pool-Punishment and Peer-Punishment: $n_{PoP} \times B + n_{PeP} \times b$;

2. Pool Punishment: $n_{PoP} \times B$ or $n_{H_{PoP}} \times B$;

3. Peer-Punishment: $n_{PeP} \times b$ or $n_{H_{PeP}} \times b$.

Pool-Punishment Average Payoffs

$$\Pi_{PoP} = P_\sigma \times \sigma + (1 - P_\sigma) \times ((Payout - c) - G) + reward_{PoP}\frac{M-1}{N-1} \tag{A.5}$$

Where $reward_{PoP}$ depends on the PGG that is played:

1. Together with Peer Punishers: $\beta\frac{n_C}{n_{PoP}+n_{PeP}}$;

2. Together with Interrogators: $\beta\frac{n_C}{n_{PoP}+n_{Int}}$.

Peer-Punishment Average Payoffs

$$\Pi_{PeP} = P_\sigma \times \sigma + (1 - P_\sigma) \times (Payout - c) + reward_{PeP}\frac{M-1}{N-1} - (c_b \times n_D)\frac{M-1}{N-1} \tag{A.6}$$

Where $reward_{PeP}$ depends on the PGG that is played:

1. Together with Pool Punishers: $\beta\frac{n_C}{n_{PoP}+n_{PeP}}$;

2. Together with Interrogators: $\beta\frac{n_C}{n_{PeP}+n_{Int}}$.

A.4.2 Trust in PGGs

We introduce trust between agents as a special parameter. We consider trust to be proportional to the number of Cooperators in games, but due to the complexity of the game we are modelling we need to make the distinction between genuine cooperators, represented by Cooperators and Interrogators/Punishers, and total cooperators which includes Deceivers, denoted by $N - n_D$. Deceivers are pretending to cooperate, thus they influence the overall trust between members of a population. We have derived this definition of trust based on Truth-Default Theory in deception literature, which states that human agents are biased to trust others by default

[153, 152]. We use $t = \frac{N-n_D}{N}$ to represent the trust between agents in a population of size N. Trust has the following properties in our model:

1. Trust increases both the likelihood of cooperation and deception.

2. Trust reduces the likelihood of defection.

3. Low trust means more defectors in a selected population M. This means lower payoff for Defectors and Deceivers.

A.4.3 Deceiver Payoff

Deceivers receive the payoff of a Cooperator without making the initial contribution. They are distinguished from Defectors because they are not subject to punishment as they conceal their defection. However this concealment is costly; it increases with the number of other agents that must be convinced, but decreases with overall trust among the population and the deceivers' innate communicative skill.

1. $commSkill$ = communicative skill of the Deceiver.

   (a) Reduces the cost of deception.

   (b) The higher the communicative skill, the more likely it is for a deceiver to succeed in deception.

2. $\gamma = 1 - commSkill$ the deceivers' risk of getting caught

3. $cogLoad = (n_C + n_{Int} + n_{Dec} + n_P) \times (1 - t) \times (1 - commSkill)$ is the cognitive load of a Deceiver. Where:

   (a) $n_C + n_{Int} + n_{Dec} + n_P$ represents the number of agents that need convincing. Here we also add the number of deceivers, because a deceiver considers them cooperators. n_P represents the number of Punishers (Peer or Pool).

(b) $(1 - t) \times (1 - commSkill)$ represents the cost to convince an agent.

- Proportional to the number of cooperators in a PGG;

4. $leakage = n_{Int} \times \gamma \times \Gamma$ represents the leakage of the Deceiver.

   (a) Increases the cost of deception.

   (b) Leakage means that the Deceiver leaves a track of evidence that might lead an Interrogator to find out about deception.

Definition 29 ***Cost of Deceiver*** *Let the cost of deception c_{Dec} be a function of $cogLoad$ and $leakage$, where $c_{Dec} = cogLoad + leakage$.*

Deception Average Payoff

$$\Pi_{Dec} = P_\sigma \times \sigma + (1 - P_\sigma) \times Payout - c_{Dec} \frac{M-1}{N-1} \tag{A.7}$$

A.4.4 Interrogator Payoff

Interrogators receive the same payout as the Peer-Punishers minus the cost of peer-punishing. They are different from Peer-Punishers as they do not punish Defectors. However,Interrogators need to hunt down Deceivers and punish them, therefore they need to pay a cost for interrogation. This cost increases with the number of agents in a population they need to interrogate as well as with the number of Deceivers they are likely to reveal and punish.

1. c_Γ = cost of punishing a Deceiver. Is multiplied by:

   (a) the probability of a deceiver's risk of getting caught γ.

   (b) the number of Deceivers n_{Dec}

2. c_{interr} = cost of interrogating an agent. It is multiplied by:

(a) the numbers of agents that need to be interrogated. These are both Cooperators and Deceivers $n_C + n_{Dec}$.

3. $\gamma =$ the likelihood of revealing a Deceiver. This is the same as the risk of a Deceiver being caught in the same population.

Definition 30 ***Cost of Interrogator*** *Let the cost of being an Interrogator c_{Int} be a function of c_Γ and c_{interr}, where $c_{Int} = \gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$.*

Interrogation Average Payoff

$$\Pi_{Int} = P_\sigma \times \sigma + (1 - P_\sigma) \times (Payout - c) - c_{Int}\frac{M-1}{N-1} + reward_{Int}\frac{M-1}{N-1} \quad \text{(A.8)}$$

Where $reward_{Int}$ depends on the type of PGG:

1. Together with Pool-Punishers: $\beta \times \frac{n_C}{n_{PoP}+n_{Int}}$;

2. Together with Peer-Punishers: $\beta \times \frac{n_C}{n_{PeP}+n_{Int}}$.

A.4.5 Hybrid Interrogators Payoffs

Cooperation in the previous PGGs suffers due to the weakness of Punishers against Deceivers and to the weakness of Interrogators against Defectors. In order to counter both defection and deception, we replace Peer-Punishers and Pool-Punishers with Hybrid Interrogators. On top of interrogating the population to find and punish Deceivers, the Hybrids also play the role of Punishers (either Peer or Pool). The role of Hybrids is to maintain cooperation by dealing with both defection and deception at the same time. This stops Deceivers and Defectors from gaining ground in a cyclical way.

Definition 31 ***Cost of Pool Hybrid Interrogators*** *Let the cost of being an Interrogator c_{Hpop} be a function of c_Γ and c_{interr}, where $c_{Hpop} = G + \gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$.*

Definition 32 ***Cost of Peer Hybrid Interrogators*** *Let the cost of being an Interrogator c_{Hpep} be a function of c_b, c_Γ and c_{interr}, where $c_{Hpep} = c_b \times n_D + \gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$.*

Pool-Hybrid Interrogation Average Payoff

$$\Pi_{H_{PoP}} = P_\sigma \times \sigma + (1 - P_\sigma) \times ((Payout - c) - G) - c_{H_{PoP}} \frac{M-1}{N-1} + (\beta \times \frac{n_C}{n_{H_{PoP}}}) \frac{M-1}{N-1} \tag{A.9}$$

Peer-Hybrid Interrogation Average Payoff

$$\Pi_{H_{PeP}} = P_\sigma \times \sigma + (1 - P_\sigma) \times (Payout - c) - c_{H_{PeP}} \frac{M-1}{N-1} + (\beta \times \frac{n_C}{n_{H_{PeP}}}) \frac{M-1}{N-1} \tag{A.10}$$

A.5 Design of PGGs

A.5.1 Punishment with Tax

A PGG that includes Pool and Peer Punisher agents without second order punishment. Instead of second-order punishment, we introduced a tax β for Punishers to exist in the population that is paid by the Cooperators. See Table A.1.

A.5.2 Punishment with Tax & Deception

A PGG where Deceiver agents are added into the mix. Deceivers pretend to be Cooperators, while scraping the same payout as Defectors. See Table A.2.

A.5.3 Pool-Punishment with Tax, Deception & Interrogation

A PGG that includes Deceiver agents along with Pool-Punishers and Interrogators. The punishment tax β takes into account both Pool-Punishers and Interrogators. Interrogators are introduced to capture and punish Deceivers. See Table A.3.

A.5.4 Peer-Punishment with Tax, Deception & Interrogation

A PGG that includes Deceiver agents along with Peer-Punishers and Interrogators. The punishment tax β takes into account both Peer-Punishers and Interrogators. See Table A.4.

A.5.5 Deception & Pool-Hybrid Interrogation

A PGG that includes Deceivers and Pool-Hybrid Interrogators. The punishment tax β takes into account both Pool-Hybrid Interrogators. See Table A.5.

A.5.6 Deception & Peer-Hybrid Interrogation

A PGG that includes Deceivers and Peer-Hybrid Interrogators. The punishment tax β takes into account both Peer-Hybrid Interrogators. See Table A.6.

A.6 Tables

Payoffs with Punishment		
Strategy	**Cost**	**Payoff**
Cooperators	β	$Payout - c - \beta$
Defectors	$n_{PoP} \times B + n_{PeP} \times b$	$Payout - n_{PoP} \times B - n_{PeP} \times b$
Peer-Punishers	$c_b \times n_D$	$Payout - c - c_b \times n_D + \beta \frac{n_C}{n_{PoP}+n_{PeP}}$
Pool-Punishers	G	$Payout - c - G + \beta \frac{n_C}{n_{PoP}+n_{Int}}$
Loners	N/A	σ

Table A.1: Payoffs for agents in a PGG without Deception. The payoffs resemble the ones in Sigmund 2010, the only difference being the additional tax β that is paid by the Cooperators.

Payoffs with Deception		
Strategy	**Cost**	**Payoff**
Cooperators	β	$Payout - c - \beta$
Defectors	$n_{PoP} \times B + n_{PeP} \times b$	$Payout - n_{PoP} \times B - n_{PeP} \times b$
Peer-Punishers	$c_b \times n_D$	$Payout - c - c_b \times n_D + \beta \frac{n_C}{n_{PoP}+n_{PeP}}$
Pool-Punishers	G	$Payout - c - G + \beta \frac{n_C}{n_{PoP}+n_{PeP}}$
Deceivers	$cogLoad$	$Payout - c_{Dec}$
Loners	N/A	σ

Table A.2: Payoffs for agents in a PGG with Deceivers present. In this PGG, there is no leakage from the Deceivers, because there is no type of agent that interrogates them. '*There is no sound made by a falling tree if there's no one to hear it.*'.

Payoffs with Interrogation and Pool Punishment		
Strategy	**Cost**	**Payoff**
Cooperators	β	$Payout - c - \beta$
Defectors	$n_{PoP} \times B$	$Payout - n_{PoP} \times B$
Pool-Punishers	G	$Payout - c - G + \beta \frac{n_C}{n_{PoP}+n_{Int}}$
Interrogators	$\gamma \times c_{\Gamma} \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$	$Payout - c - c_{Int} + \beta \frac{n_C}{n_{PoP}+n_{Int}}$
Deceivers	$cogLoad + leakage$	$Payout - c_{Dec}$
Loners	N/A	σ

Table A.3: Payoffs for agents in a PGG with only Pool-Punishers where both Deceivers and Interrogators are present. In this PGG, leakage is added to the the cost of deceiving.

Payoffs with Interrogation and Peer-Punishment		
Strategy	**Cost**	**Payoff**
Cooperators	β	$Payout - c - \beta$
Defectors	$n_{PeP} \times b$	$Payout - n_{PeP} \times b$
Peer-Punishers	$c_b \times n_D$	$Payout - c - c_b \times n_D + \beta \frac{n_C}{n_{PeP}+n_{Int}}$
Interrogators	$\gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$	$Payout - c - c_{Int} + \beta \frac{n_C}{n_{PoP}+n_{Int}}$
Deceivers	$cogLoad + leakage$	$Payout - c_{Dec}$
Loners	N/A	σ

Table A.4: Payoffs for agents in a PGG with only Peer-Punishers where both Deceivers and Interrogators are present. In this PGG, leakage is added to the the cost of deceiving.

Payoffs with Pool Hybrids		
Strategy	**Cost**	**Payoff**
Cooperators	β	$Payout - c - \beta$
Defectors	$n_{Hpop} \times B$	$Payout - n_{Hpop}$
Pool Hybrids	$G + \gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$	$Payout - c - c_{Hpop} + \beta \frac{n_C}{n_{Hpop}}$
Deceivers	$cogLoad + leakage$	$Payout - c_{Dec}$
Loners	N/A	σ

Table A.5: Payoffs for agents in a PGG where both Deceivers and Hybrid Interrogators are present. In this PGG, the Hybrid Interrogators play both the role of Pool-Punishers and the role of Interrogators, hunting down Defectors and Deceivers.

Payoffs with Peer Hybrids		
Strategy	**Cost**	**Payoff**
Cooperators	β	$Payout - c - \beta$
Defectors	$n_{Hpep} \times b$	$Payout - n_{Hpep}$
Peer Hybrids	$c_b \times n_D + \gamma \times c_\Gamma \times n_{Dec} + c_{interr} \times (n_C + n_{Dec})$	$Payout - c - c_{Hpep} + \beta \frac{n_C}{n_{Hpep}}$
Deceivers	$cogLoad + leakage$	$Payout - c_{Dec}$
Loners	N/A	σ

Table A.6: Payoffs for agents in a PGG where both Deceivers and Hybrid Interrogators are present. In this PGG, the Hybrid Interrogators play both the role of Peer-Punishers and the role of Interrogators, hunting down Defectors and Deceivers.

Bibliography

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

[2] Pushkal Agarwal, Sagar Joglekar, Panagiotis Papadopoulos, Nishanth Sastry, and Nicolas Kourtellis. Stop tracking me bro! Differential tracking of user demographics on hyper-partisan websites. In *Proceedings of The Web Conference 2020*, page 1479–1490. ACM, 2020.

[3] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[4] Jonathan Albright. Welcome to the era of fake news. *Media and Communication*, 5(2):87–89, 2017.

[5] Alexa. Alexa internet. keyword research, competitor analysis, and website ranking, 2020.

[6] J McKenzie Alexander. Evolutionary game theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019.

[7] Martin Andrews and Sam Witteveen. Unsupervised natural question answering with a small model. *arXiv preprint arXiv:1911.08340*, 2019.

[8] Marc Artiga and Cédric Paternotte. Deception: a functional account. *Philosophical Studies*, 175(3):579–600, 2018.

[9] Robert Axelrod. *The complexity of cooperation: Agent-based models of competition and collaboration*, volume 3. Princeton University Press, 1997.

[10] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[11] Chris L Baker and Joshua B Tenenbaum. Modeling human plan recognition using Bayesian theory of mind. *Plan, activity, and intent recognition: Theory and practice*, pages 177–204, 2014.

[12] Mikhail Bakhtin. The dialogical principle. *Theory and History of Literature*, 13:73–73, 1984.

[13] Mikhail Bakhtin. *Problems of Dostoevsky's Poetics.* University of Minnesota Press, 1984.

[14] Albert Bandura and Richard H Walters. *Social learning theory*, volume 1. Prentice-hall Englewood Cliffs, NJ, 1977.

[15] Luca Barlassina and Robert M. Gordon. Folk psychology as mental simulation. In E. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Center for the Study of Language and Information, Stanford, 2004.

[16] Rafael A Barrio, Tzipe Govezensky, Robin Dunbar, Gerardo Iniguez, and Kimmo Kaski. Dynamics of deceptive interactions in social networks. *Journal of The Royal Society Interface*, 12(112):20150798, 2015.

[17] Charlie Beckett. Wikitribune: can crowd-sourced journalism solve the crisis of trust in news? *POLIS: journalism and society at the LSE*, 2017.

[18] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa. Jade–a fipa-compliant agent framework. In *Proceedings of PAAM*, volume 99, page 33. London, 1999.

[19] Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. Emotions in argumentation: an empirical evaluation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. AAAI Press, 2015.

[20] Yoella Bereby-Meyer, Sayuri Hayakawa, Shaul Shalvi, Joanna D Corey, Albert Costa, and Boaz Keysar. Honesty speaks a second language. *Topics in cognitive science*, 2018.

[21] Floris J Bex. *Evidence for a good story: A hybrid theory of arguments, stories and criminal evidence*. PhD thesis, University of Groningen, 2009.

[22] Floris J Bex. *Arguments, stories and criminal evidence: A formal hybrid theory*, volume 92. Springer Science & Business Media, 2011.

[23] Floris J Bex and Trevor Bench-Capon. Arguing with stories. In *Narration as Argument*, pages 31–45. Springer, 2017.

[24] Floris J Bex and Bart Verheij. Story schemes for argumentation about the facts of a crime. In *Proceedings of the 2010 AAAI Fall Symposium Series*, 2010.

[25] Floris J Bex and Douglas Walton. Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68, 2016.

[26] Shweta Bhatt, Sagar Joglekar, Shehar Bano, and Nishanth Sastry. Illuminating an ecosystem of partisan websites. In *Companion Proceedings of the*

The Web Conference 2018, pages 545–554. International World Wide Web Conferences Steering Committee, 2018.

[27] Bloomberg. How faking videos became easy and why that's so scary. https://www.bloomberg.com/news/articles/2018-09-10/how-faking-videos-became-easy-and-why-that-s-so-scary-quicktake, 2018.

[28] Olivier Boissier, Rafael H Bordini, Jomi F Hübner, Alessandro Ricci, and Andrea Santi. Multi-agent oriented programming with JaCaMo. *Science of Computer Programming*, 78(6):747–761, 2013.

[29] Charles F Bond Jr and Bella M DePaulo. Individual differences in judging deception: Accuracy and bias. *Psychological bulletin*, 134(4):477, 2008.

[30] Charles F Bond Jr and William E Fahey. False suspicion and the misperception of deceit. *British Journal of Social Psychology*, 26(1):41–46, 1987.

[31] Shane Bonetti. Experimental economics and deception. *Journal of Economic Psychology*, 19(3):377–395, 1998.

[32] Rafael H Bordini, Mehdi Dastani, Jürgen Dix, and A El Fallah Seghrouchni. *Multi-Agent Programming*. Springer, 2009.

[33] Rafael H. Bordini, Jomi Fred Hübner, and Michael Wooldridge. *Programming Multi-Agent Systems in AgentSpeak using Jason (Wiley Series in Agent Technology)*. John Wiley & Sons, 2007.

[34] Laura Bradford, Mateo Aboy, and Kathleen Liddell. COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. *Journal of Law and the Biosciences*, 7(1), 05 2020. lsaa034.

[35] Michael Bratman et al. *Intention, plans, and practical reason*, volume 10. Harvard University Press Cambridge, MA, 1987.

[36] Torben Braüner, Patrick Blackburn, and Irina Polyanskaya. Being deceived: Information asymmetry in second-order false belief tasks. *Topics in cognitive science*, 2019.

[37] Tom Buchanan and Monica T Whitty. The online dating romance scam: causes and consequences of victimhood. *Psychology, Crime & Law*, 20(3):261–283, 2014.

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

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

[40] Martin Caminada. Truth, lies and bullshit; distinguishing classes of dishonesty. In *Proceedings of the Social Simulation Workshop @IJCAI*. Citeseer, 2009.

[41] Thomas L Carson. The definition of lying. *Noûs*, 40(2):284–306, 2006.

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

[43] Cristiano Castelfranchi and Rino Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*, volume 18. John Wiley & Sons, 2010.

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

[45] Cristiano Castelfranchi and Yao-Hua Tan. The role of trust and deception in virtual societies. *International Journal of Electronic Commerce*, 6(3):55–70, 2002.

[46] Tathagata Chakraborti and Subbarao Kambhampati. (When) can ai bots lie? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 53–59, 2019.

[47] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

[48] Roderick M Chisholm and Thomas D Feehan. The intent to deceive. *The journal of Philosophy*, 74(3):143–159, 1977.

[49] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[50] David Christian and R Michael Young. Strategic deception in agents. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 218–226. IEEE, 2004.

[51] Michael Suk-Young Chwe. *Rational ritual: Culture, coordination, and common knowledge*. Princeton University Press, 2013.

[52] Danielle K Citron and Robert Chesney. Deep fakes: A looming crisis for national security, democracy and privacy? *Lawfare*, 2018.

[53] Andy Clark. Embodied, situated, and distributed cognition. *A companion to cognitive science*, pages 506–517, 2017.

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

[55] Oana Cocarascu and Francesca Toni. Detecting deceptive reviews using argumentation. In *Proceedings of the International Workshop on AI for Privacy and Security*, pages 1–8. ACM, 2016.

[56] Philip Cohen. Foundations of collaborative task-oriented dialogue: What's in a slot? In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 198–209, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[57] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[58] Daniel Crevier. *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, Inc., 1993.

[59] Ellen Daniel. Jimmy Wales: Advertising-only model has been "incredibly destructive" for journalism. *Verdict Magazine*, 2020.

[60] Mehdi Dastani. 2APL: a practical agent programming language. *Autonomous agents and multi-agent systems*, 16(3):214–248, 2008.

[61] Bethan L Davies. Grice's cooperative principle: meaning and rationality. *Journal of pragmatics*, 39(12):2308–2331, 2007.

[62] Fiorella De Rosis, Valeria Carofiglio, Giuseppe Grassano, and Cristiano Castelfranchi. Can computers deliberately deceive? A simulation tool and its application to Turing's imitation game. *Computational Intelligence*, 19(3):235–263, 2003.

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

[64] Harmen De Weerd and Bart Verheij. The advantage of higher-order theory of mind in the game of limited bidding. In *Proceedings of the Workshop on Reasoning about Other Minds*, volume 751, pages 149–164, 2011.

[65] Chris Deaton, Blake Shepard, Charles Klein, Corrinne Mayans, Brett Summers, Antoine Brusseau, Michael Witbrock, and Doug Lenat. The comprehensive terrorism knowledge base in cyc. In *Proceedings of the 2005 International Conference on Intelligence Analysis*. Citeseer, 2005.

[66] Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.

[67] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.

[68] Virginia Dignum. Responsible autonomy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4698–4704. AAAI Press, 2017.

[69] Prashant Doshi, Xia Qu, Adam Goodie, and Diana Young. Modeling recursive reasoning by humans using empirically informed interactive POMDPs. In *Proceedings of the 9th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1223–1230, 2010.

[70] Mark Dras, Debbie Richards, Meredith Taylor, and Mary Gardiner. Generating and detecting deceptive language in virtual agents. In *Proceedings of the International Workshop on Interacting with ECAs as Virtual Characters*, page 38, 2010.

[71] Michelle Drouin, Daniel Miller, Shaun MJ Wehle, and Elisa Hernandez. Why do people lie online?"Because everyone lies on the internet". *Computers in Human Behavior*, 64:134–142, 2016.

[72] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[73] Marta Dynel. To say the least: Where deceptively withholding information ends and lying begins. *Topics in cognitive science*, 2018.

[74] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition).* WW Norton & Company, 2009.

[75] Paul Ekman and Wallace V Friesen. Nonverbal Leakage and Clues to Deception. *Psychiatry*, 32(1):88–106, Feb 1969.

[76] Paul Ekman and Wallace V Friesen. Detecting deception from the body or face. *Journal of personality and Social Psychology*, 29(3):288, 1974.

[77] Paul Ekman and Maureen O'Sullivan. Hazards in detecting deceit. *Psychological methods in criminal investigation and evidence*, pages 297–332, 1989.

[78] Chris Eliasmith. Computation and dynamical models of mind. *Minds and Machines*, 7(4):531–541, 1997.

[79] Rino Falcone and Cristiano Castelfranchi. Social trust: A cognitive approach. In *Trust and deception in virtual societies*, pages 55–90. Springer, 2001.

[80] Rino Falcone, Munindar Singh, and Yao-Hua Tan. *Trust in cyber-societies: integrating the human and artificial perspectives*, volume 2246. Springer Science & Business Media, 2001.

[81] Don Fallis. Shedding light on keeping people in the dark. *Topics in cognitive science*, 2018.

[82] Kimberly Ferguson-Walter, Sunny Fugate, Justin Mauger, and Maxine Major. Game theory for adaptive defensive cyber deception. In *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security*, pages 1–8, 2019.

[83] John Ferris. The intelligence-deception complex: An anatomy. *Intelligence and National Security*, 4(4):719–734, 1989.

[84] Debora Field and Allan Ramsay. Sarcasm, deception, and stating the obvious: Planning dialogue without speech acts. *Artificial Intelligence Review*, 22(2):149–171, 2004.

[85] Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire. KQML as an agent communication language. In *Proceedings of the Third International Conference on Information and Knowledge Management*, pages 456–463. ACM, 1994.

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

[87] Luciano Floridi. *Philosophy and computing: An introduction.* Psychology Press, 1999.

[88] Luciano Floridi. On the intrinsic value of information objects and the infosphere. *Ethics and information technology*, 4(4):287–304, 2002.

[89] Luciano Floridi. The method of levels of abstraction. *Minds and Machines*, 18(3):303–329, 2008.

[90] Steven P Fonseca, Martin L Griss, and Reed Letsinger. Agent behavior architectures a mas framework comparison. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 86–87, 2002.

[91] Michael Franke, Giulio Dulcinati, and Nausicaa Pouscoulous. Strategies of deception: Under-informativity, uninformativity, and lies—misleading with different kinds of implicature. *Topics in cognitive science*, 2019.

[92] Harry G Frankfurt. *On bullshit.* Princeton University Press, 2009.

[93] Alex Garland. *Ex machina.* Faber & Faber, 2015.

[94] Henry SJ Garnet. *A Treatise of Equivocation, ca. 1598 (ed. by D Jardine).* Longman, Brown, Green and Longmans London, 1851.

[95] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Proceedings of the International Workshop on Agent Theories, Architectures, and Languages*, pages 1–10. Springer, 1998.

[96] Linda M Geven, Gershon Ben-Shakhar, Merel Kindt, and Bruno Verschuere. Memory-based deception detection: Extending the cognitive signature of lying from instructed to self-initiated cheating. *Topics in cognitive science*, 2018.

[97] Antje Gimmler. Deliberative democracy, the public sphere and the internet. *Philosophy & Social Criticism*, 27(4):21–39, 2001.

[98] Uri Gneezy. Deception: The role of consequences. *The American Economic Review*, 95(1):384–394, 2005.

[99] Jennifer Golbeck. *Computing with social trust.* Springer Science & Business Media, 2008.

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

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

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

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

[104] Pär Anders Granhag, Aldert Vrij, and Bruno Verschuere. *Detecting deception: Current challenges and cognitive approaches.* John Wiley & Sons, 2015.

[105] Jesse Gray and Cynthia Breazeal. Manipulating mental states through physical action. *International Journal of Social Robotics*, 6(3):315–327, 2014.

[106] Gian Maria Greco and Luciano Floridi. The tragedy of the digital commons. *Ethics and Information Technology*, 6(2):73–81, 2004.

[107] Elizabeth Green. Why do americans stink at math. *The New York Times Magazine*, 23, 2014.

[108] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[109] Jürgen Habermas. *The theory of communicative action: Lifeworld and systems, a critique of functionalist reason*, volume 2. John Wiley & Sons, 2015.

[110] Christos Hadjinikolis, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney. Opponent modelling in persuasion dialogues. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 164–170. AAAI Press, 2013.

[111] Charles L Hamblin. Questions in Montague grammar. *Foundations of language*, 10(1):41–53, 1973.

[112] Charles L Hamblin. *Linguistics and the Parts of the Mind: Or how to Build a Machine Worth Talking to.* Cambridge Scholars Publishing, 2018.

[113] Jeffrey T Hancock, Mor Naaman, and Karen Levy. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 2020.

[114] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. Modeling agents with a theory of mind. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 02*, pages 217–224. IEEE, 2009.

[115] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[116] Maria Hartwig, Pär A Granhag, Leif Stromwall, Ann G Wolf, Aldert Vrij, and Emma Roos af Hjelmsäter. Detecting deception in suspects: Verbal cues as a function of interview strategy. *Psychology, Crime & Law*, 17(7):643–656, 2011.

[117] Maria Hartwig, Pär Anders Granhag, Leif A Strömwall, and Ola Kronkvist. Strategic use of evidence during police interviews: When training to detect deception works. *Law and Human Behavior*, 30(5):603–619, 2006.

[118] Martin Heidegger. *Being and time: A translation of Sein und Zeit.* SUNY press, 1996.

[119] Brenna Helm, Thomas Vander Ven, and Howard T Welser. The electric hookup: Individual and social risks related to hookup app use among emerging adults. In *Recent Advances in Digital Media Impacts on Identity, Sexuality, and Relationships*, pages 62–81. IGI Global, 2020.

[120] John Herrman. Inside Facebook's (Totally Insane, Unintentionally Gigantic, Hyperpartisan) Political-Media Machine. *The New York Times Magazine*, 2016.

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

[122] Richards J Heuer. Cognitive factors in deception and counterdeception. *Strategic military deception*, pages 45–94, 1980.

[123] Richards J Heuer. *Psychology of Intelligence Analysis.* Center for the Study of Intelligence, 1999.

[124] Martin Hilbert. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin*, 138(2):211, 2012.

[125] Koen V Hindriks. Programming rational agents in GOAL. In *Multi-agent programming*, pages 119–157. Springer, 2009.

[126] Koen V Hindriks, Frank S De Boer, Wiebe Van der Hoek, and John-Jules Ch Meyer. Agent programming in 3APL. *Autonomous Agents and Multi-Agent Systems*, 2(4):357–401, 1999.

[127] Jomi F Hubner, Jaime S Sichman, and Olivier Boissier. Developing organised multiagent systems using the MOISE+ model: programming issues at the system and agent levels. *International Journal of Agent-Oriented Software Engineering*, 1(3-4):370–395, 2007.

[128] Aaron Hunter, Francois Schwarzentruber, and Eric Tsang. Belief manipulation through propositional announcements. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1109–1115. AAAI Press, 2017.

[129] Alistair Isaac and Will Bridewell. Mindreading deception in dialog. *Cognitive Systems Research*, 28:12–19, Jun 2014.

[130] Alistair Isaac and Will Bridewell. *White lies on silver tongues: Why robots need to deceive (and how)*, volume 2, pages 157–72. Oxford University Press, 2017.

[131] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, pages 1–5, 2019.

[132] Andrew J I Jones. On The Logic of Self-deception. *South American Journal of Logic*, 1:387–400, 2015.

[133] Parisa Kaghazgaran, Majid Alfifi, and James Caverlee. TOmCAT: Target-Oriented Crowd Review ATtacks and Countermeasures. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 302–312, 2019.

[134] Parisa Kaghazgaran, James Caverlee, and Majid Alfifi. Behavioral analysis of review fraud: Linking malicious crowdsourcing to amazon and beyond. In

Proceedings of the 11th International AAAI Conference on Web and Social Media, 2017.

[135] Parisa Kaghazgaran, James Caverlee, and Anna Squicciarini. Combating crowdsourced review manipulators: A neighborhood-based approach. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pages 306–314, 2018.

[136] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. Coercion and deception in persuasive technologies. In *Proceedings of the International Trust Workshop @AAMAS/IJCAI/ECAI/ICML*, pages 38–49. CEUR-WS, 2018.

[137] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. Implementing argumentation-enabled empathic agents. In *Proceedings of the European Conference on Multi-Agent Systems*, pages 140–155. Springer, 2018.

[138] Marek Kowalsk. Faceswap. Accessed on 2020-02-08, https://github.com/MarekKowalski/FaceSwap, 2018.

[139] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[140] Srijan Kumar, Meng Jiang, Taeho Jung, Roger Jie Luo, and Jure Leskovec. MIS2: Misinformation and misbehavior mining on the web. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 799–800. ACM, 2018.

[141] Yannis Labrou and Tim Finin. A semantics approach for KQML—a general purpose communication language for software agents. In *Proceedings of the 3rd International Conference on Information and Knowledge Management*, pages 447–455. ACM, 1994.

[142] Wesley Lacson and Beata Jones. The 21st century darknet market: Lessons from the fall of silk road. *International Journal of Cyber Criminology*, 10(1), 2016.

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

[144] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *Proceedings of the 29th Conference on Innovative Applications of Artificial Intelligence*, 2017.

[145] Ralph Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3):49–51, 2011.

[146] Carolyn Lauckner, Natalia Truszczynski, Danielle Lambert, Varsha Kottamasu, Saher Meherally, Anne Marie Schipani-McLaughlin, Erica Taylor, and Nathan Hansen. "catfishing," cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males. *Journal of Gay & Lesbian Mental Health*, 23(3):289–306, 2019.

[147] Jennifer Lavoie and Victoria Talwar. Care to share? children's cognitive skills and concealing responses to a parent. *Topics in cognitive science*, 2018.

[148] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[149] Ivan Leudar and Alan Costall. On the persistence of the 'problem of other minds' in psychology: Chomsky, grice and theory of mind. *Theory & Psychology*, 14(5):601–621, 2004.

[150] Hector J Levesque, Philip R Cohen, and José HT Nunes. On acting together. In *Proceedings of the 8th AAAI Conference on Artificial Intelligence*, volume 90, pages 94–99, 1990.

[151] Hector J Levesque, Raymond Reiter, Yves Lespérance, Fangzhen Lin, and Richard B Scherl. Golog: A logic programming language for dynamic domains. *The Journal of Logic Programming*, 31(1-3):59–83, 1997.

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

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

[154] Timothy R Levine and Steven A McCornack. Theorizing about deception. *Journal of Language and Social Psychology*, 33(4):431–440, 2014.

[155] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[156] David C Logan. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of experimental botany*, 60(3):712–714, 2009.

[157] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409*, 2019.

[158] Walid Magdy, Yehia Elkhatib, Gareth Tyson, Sagar Joglekar, and Nishanth Sastry. Fake it till you make it: Fishing for catfishes. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 497–504. ACM, 2017.

[159] James E Mahon. History of deception: 1950 to the present. *Encyclopedia of Deception*, pages 618–619, 2014.

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

[161] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[162] Christina Masden and W Keith Edwards. Understanding the role of community in online dating. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 535–544. ACM, 2015.

[163] Peta Masters and Sebastian Sardina. Deceptive path-planning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4368–4375. AAAI Press, 2017.

[164] Michael Mateas and Phoebe Sengers. Narrative intelligence. Technical report, 1999.

[165] James Mayfield, Yannis Labrou, and Tim Finin. Evaluation of KQML as an agent communication language. In *Proceedings of the International Workshop on Agent Theories, Architectures, and Languages*, pages 347–360. Springer, 1995.

[166] Peter McBurney. What are models for? In *Proceedings of the European Workshop on Multi-Agent Systems*, pages 175–188. Springer, 2011.

[167] Peter McBurney, William Nash, and Andrew Jones. Lies and Deception. Technical report, King's College London, Department of Informatics, Jan 2014.

[168] Peter McBurney and Simon Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of logic, language and information*, 11(3):315–334, 2002.

[169] Peter McBurney and Simon Parsons. Dialogue games for agent argumentation. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 261–280. Springer US, 2009.

[170] Peter McBurney, Rogier M Van Eijk, Simon Parsons, and Leila Amgoud. A dialogue game protocol for agent purchase negotiations. *Autonomous Agents and Multi-Agent Systems*, 7(3):235–273, 2003.

[171] Peter John McBurney. *Rational interaction*. PhD thesis, University of Liverpool, 2002.

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

[173] Richard W. McVinney. Deep fakes & deep fears. https://rwmcvinney.wordpress.com/author/rwmcvinney/, 2019.

[174] Johnathan Mell, Gale M Lucas, and Jonathan Gratch. Welcome to the real world: How agent strategy increases human willingness to deceive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1250–1257. IFAAMAS, 2018.

[175] Victor S Melo, Alison R Panisson, and Rafael H Bordini. Argumentation-based reasoning using preferences over sources of information. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1337–1338. IFAAMAS, 2016.

[176] John-Jules Ch. Meyer. Logics for intelligent agents and multi-agent systems. In Jörg H. Siekmann, editor, *Computational Logic*, volume 9 of *Handbook of the History of Logic*, pages 629 – 658. North-Holland, 2014.

[177] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[178] Marvin Minsky and Doug Riecken. A conversation with Marvin Minsky about agents. *Communications of the ACM*, 37(7):22–29, 1994.

[179] Kevin D Mitnick and William L Simon. *The Art of Deception: Controlling the Human Element of Security.* John Wiley & Sons, 2011.

[180] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 279–288. ACM, 2019.

[181] Sanjay Modgil, Francesca Toni, Floris J Bex, Ivan Bratko, Carlos I Chesnevar, Wolfgang Dvořák, Marcelo A Falappa, Xiuyi Fan, Sarah Alice Gaggl, Alejandro J García, et al. The added value of argumentation. In *Agreement Technologies*, pages 357–403. Springer, 2013.

[182] Kimberlee Morrison. Fake social media accounts incite police investigations. Available at `http://www.adweek.com/digital/fake-social-media-accounts-incite-police-investigations/`, 2014.

[183] Gary Saul Morson and Caryl Emerson. *Mikhail Bakhtin: Creation of a prosaics.* Stanford University Press, 1990.

[184] Francesca Mosca, Ştefan Sarkadi, Jose M Such, and Peter McBurney. Agent EXPRI: Licence to Explain. In *International Workshop on Explainable, Trans-*

parent Autonomous Agents and Multi-Agent Systems, pages 21–38. Springer, 2020.

[185] Francesca Mosca and Jose M Such. ELVIRA: an Explainable Agent for Value and Utility-driven Multiuser Privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems.* IFAAMAS, 2021.

[186] Marco Mulder, Jan Treur, and Michael Fisher. Agent modelling in metatem and desire. In *Proceedings of the International Workshop on Agent Theories, Architectures, and Languages*, pages 193–207. Springer, 1997.

[187] A Nettel. The enthymeme between persuasion and argumentation. In *Proceedings of the Conference on Argumentation of the International Society for the Study of Argumentation*, pages 1359–1365, 2011.

[188] Harvey P Newquist. *The brain makers : genius, ego, and greed in the quest for machines that think.* Sams Publ., 1994.

[189] Thanh Thi Nguyen, Cuong M Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 2019.

[190] Pablo Noriega. *Agent mediated auctions: the fishmarket metaphor.* Institut d'Investigació en Intelligència Artificial, 1999.

[191] Martin A Nowak. *Evolutionary dynamics: exploring the equations of life.* Harvard University Press, 2006.

[192] Martin A Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428(6983):646, 2004.

[193] Jay F Nunamaker, Douglas C Derrick, Aaron C Elkins, Judee K Burgoon, and Mark W Patton. Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, 28(1):17–48, 2011.

[194] High-Level Expert Group on Business-to Government Data Sharing. Towards a European strategy on business-to-government data sharing for the public interest. EU Technical Report. Technical report, European Commission, 2020.

[195] Nir Oren, Tim Norman, Alun Preece, and Stuart Chalmers. Policing virtual organizations. In *Proceedings of the 2nd European Conference on Multi-Agent Systems*, pages 499–508, 2004.

[196] Fabio Paglieri, Cristiano Castelfranchi, Célia da Costa Pereira, Rino Falcone, Andrea Tettamanzi, and Serena Villata. Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. *Computational and Mathematical Organization Theory*, 20(2):176–194, 2014.

[197] Alison R Panisson, Victor S Melo, and Rafael H Bordini. Using preferences over sources of information in argumentation-based reasoning. In *Proceedings of the 5th Brazilian Conference on Intelligent Systems*, pages 31–36. IEEE, 2016.

[198] Alison R Panisson, Felipe Meneguzzi, Moser Silva Fagundes, Renata Vieira, and Rafael H Bordini. Formal semantics of speech acts for argumentative dialogues. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1437–1438. IFAAMAS, 2014.

[199] Alison R Panisson, Felipe Meneguzzi, Renata Vieira, and Rafael H Bordini. Towards practical argumentation in multi-agent systems. In *Proceedings of the Brazilian Conference on Intelligent Systems*, pages 98–103. IEEE, 2015.

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

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

[202] Simon Parsons, Elizabeth Sklar, and Peter McBurney. Using argumentation to reason with and about trust. In *Proceedings of the International Workshop on Argumentation in Multi-Agent Systems*, pages 194–212. Springer, 2011.

[203] Simon Parsons, Yuqing Tang, Elizabeth Sklar, Peter McBurney, and Kai Cai. Argumentation-based reasoning in agents with varying degrees of trust. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 879–886. IFAAMAS, 2011.

[204] David Pereira, Eugénio Oliveira, and Nelma Moreira. Modelling emotional bdi agents. In *Proceedings of the Workshop on Formal Approaches to Multi-Agent Systems*, 2006.

[205] G D Plotkin. A structural approach to operational semantics. Technical Report DAIMI FN-19, Computer Science Department, Aarhus University, 1981.

[206] Alexander Pokahr, Lars Braubach, and Winfried Lamersdorf. Jadex: A bdi reasoning engine. In *Multi-agent programming*, pages 149–174. Springer, 2005.

[207] Simon Pope and Audun Jøsang. Analysis of competing hypotheses using subjective logic. Technical report, Queensland University Brisbane (Australia), 2005.

[208] Simon Pope, Audun Jøsang, and David McAnally. Formal methods of countering deception and misperception in intelligence analysis. *Proceedings of the 11th ICCRTS Coalition Command and Control in the Networked Era*, 2006.

[209] Karl Popper. *The logic of scientific discovery.* Routledge, 2005.

[210] Stephen Porter, Leanne ten Brinke, Alysha Baker, and Brendan Wallace. Would I lie to you? "leakage" in deceptive facial expressions relates to psychopathy and emotional intelligence. *Personality and Individual Differences*, 51(2):133–137, 2011.

[211] Stefan Poslad, Phil Buckle, and Rob Hadingham. The FIPA-OS agent platform: Open source for open standards. In *proceedings of the 5th international conference and exhibition on the practical application of intelligent agents and multi-agents*, volume 355, page 0, 2000.

[212] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[213] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477, 2019.

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

[215] Anand S Rao and Michael P Georgeff. Modeling rational agents within a bdi-architecture. *KR*, 91:473–484, 1991.

[216] Anand S Rao, Michael P Georgeff, et al. Bdi agents: from theory to practice. In *Proceedings of the 1st International Conference on Multi-Agent Systems*, volume 95, pages 312–319, 1995.

[217] Diogo Rato, Brian Ravenet, Rui Prada, and Ana Paiva. Strategically misleading the user: Building a deceptive virtual suspect. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems*, pages 1711–1713, 2017.

[218] John Rawls. *A theory of justice*. Harvard university press, 2009.

[219] Aunshul Rege. What's love got to do with it? exploring online dating scams and identity fraud. *International Journal of Cyber Criminology*, 3(2), 2009.

[220] Lauren Reichart Smith, Kenny D Smith, and Matthew Blazka. Follow me, what's the harm: Considerations of catfishing and utilizing fake online personas on social media. *J. Legal Aspects Sport*, 27:32, 2017.

[221] David Resnick. The ethics of science. *London: Rout*, 1998.

[222] Alessandro Ricci, Mirko Viroli, and Andrea Omicini. Cartago: A framework for prototyping artifact-based environments in mas. In *International Workshop on Environments for Multi-Agent Systems*, pages 67–86. Springer, 2006.

[223] Kevin Roose. Here Come the Fake Videos, Too. `https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html`, 2018.

[224] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11. IEEE, 2019.

[225] Janet Rothwell, Zuhair Bandar, James O'Shea, and David McLean. Silent talker: a new computer-based system for the analysis of facial cues to deception. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(6):757–777, 2006.

[226] Janet Rothwell, Zuhair Bandar, James O'Shea, and David McLean. Charting the behavioural state of a person using a backpropagation neural network. *Neural Computing and Applications*, 16(4-5):327–339, 2007.

[227] Donald Rumsfeld. US Department of Defense news briefing, February 12, 2002. US DoD Press Conference Transcript, url=https://www.nato.int/docu/speech/2002/s020606g.htm, 2002.

[228] Sean Russell, Howell Jordan, Gregory MP O'Hare, and Rem W Collier. Agent factory: a framework for prototyping logic-based aop languages. In *Proceedings of the German Conference on Multi-Agent System Technologies*, pages 125–136. Springer, 2011.

[229] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3:1, 2019.

[230] Chiaki Sakama. Dishonest reasoning by abduction. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. AAAI Press, 2011.

[231] Chiaki Sakama. Dishonest arguments in debate games. In *Proceedings of the International Conference on Computational Models of Argument*, volume 75, pages 177–184. IOS Press, 2012.

[232] Chiaki Sakama. A formal account of deception. In *Proceedings of the 2015 AAAI Fall Symposium Series*, 2015.

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

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

[235] Chiaki Sakama, Martin Caminada, and Andreas Herzig. A formal account of dishonesty. *Logic Journal of the IGPL*, 23(2):259–294, 2015.

[236] Eugene Santos and Deqing Li. On deception detection in multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(2):224–235, 2009.

[237] Stefan Sarkadi. Deception. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5781–5782. AAAI Press, 2018.

[238] Stefan Sarkadi. Argumentation-based dialogue games for modelling deception. In Federico Castagna, Francesca Mosca, Jack Mumford, Stefan Sarkadi, and Andreas Xydis, editors, *Online Handbook of Argumentation for AI: Volume 1*, pages 38–42. arXiv, 6 2020. eprint at https://arxiv.org/abs/2006.12020.

[239] Stefan Sarkadi. Deceptive autonomous agents. In *Proceedings of the Defence and Security Doctoral Symposium at Shrivenham*. Cranfield University, 2020.

[240] Stefan Sarkadi, Peter McBurney, and Simon Parsons. Deceptive storytelling in artificial dialogue games. In *Proceedings of the 2019 AAAI Spring Symposium Series*, 2019.

[241] Stefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, and Simon Parsons. Towards an approach for modelling uncertain theory of mind

in multi-agent systems. In *Proceedings of the 6th International Conference on Agreement Technologies*, pages 3–17, Bergen, Norway, 2018. Springer.

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

[243] Mike S Schäfer. Digital public sphere. *The international encyclopedia of political communication*, pages 1–7, 2015.

[244] Roger C Schank. *Tell me a story: Narrative and intelligence.* Northwestern University Press, 1995.

[245] Michael Schillo, Petra Funk, and Michael Rovatsos. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence*, 14(8):825–848, 2000.

[246] Aaron Schlenker, Omkar Thakoor, Haifeng Xu, Fei Fang, Milind Tambe, Long Tran-Thanh, Phebe Vayanos, and Yevgeniy Vorobeychik. Deceiving cyber adversaries: A game theoretic approach. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 892–900. IFAAMAS, 2018.

[247] Paul Schweizer. Triviality arguments reconsidered. *Minds and Machines*, 29(2):287–308, 2019.

[248] John Scott and Gordon Marshall. explanandum and explanans, 2009.

[249] John Rogers Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

[250] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

[251] Jaeeun Shim and Ronald C Arkin. Biologically-inspired deceptive behavior for a robot. In *Proceedings of the International Conference on Simulation of Adaptive Behavior*, pages 401–411. Springer, 2012.

[252] Jaeeun Shim and Ronald C Arkin. A taxonomy of robot deception and its benefits in hri. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2328–2335. IEEE, 2013.

[253] Yoav Shoham. Agent-oriented programming. *Artificial intelligence*, 60(1):51–92, 1993.

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

[255] Craig Silverman, Jane Lytvynenko, Lam Thuy Vo, and Jeremy Singer-Vine. Inside the partisan fight for your news feed. Available at `https://www.buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-news-feed`, 2017.

[256] Gerardo I Simari. From data to knowledge engineering for cybersecurity. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6403–6407. AAAI Press, 2019.

[257] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In *Proceedings of the OTM Confederated International Con-*

ferences "On the Move to Meaningful Internet Systems", pages 1223–1237. Springer, 2002.

[258] Elizabeth Sklar, Simon Parsons, and Mathew Davies. When is it okay to lie? A simple model of contradiction in agent-based dialogues. In *Proceedings of the International Workshop on Argumentation in Multi-Agent Systems*, pages 251–261. Springer, 2004.

[259] Beate Sodian, Catherine Taylor, Paul L Harris, and Josef Perner. Early deception and the child's theory of mind: False trails and genuine markers. *Child development*, 62(3):468–483, 1991.

[260] Roy A Sorensen. *A Cabinet of Philosophical Curiosities: A Collection of Puzzles, Oddities, Riddles and Dilemmas.* Oxford University Press, 2016.

[261] Helena Sousa, Manuel Pinto, Elsa Costa, et al. Digital public sphere: weaknesses and challenges. *Comunicação e Sociedade*, 23:9–12, 2013.

[262] Russell Spivak. Deepfakes": The newest way to commit one of the oldest crimes. *The Georgetown Law Technology Review*, 3(2):339–400, 2019.

[263] Eugen Staab and Martin Caminada. On the profitability of incompetence. In *Proceedings of the International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 76–92. Springer, 2010.

[264] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 2017.

[265] Frank J Stech and C Elässer. Deception detection by analysis of competing hy-

pothesis,". In *Proceedings of the 2005 International Conference on Intelligence Analysis*, volume 8, pages 12–15. Citeseer, 2005.

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

[267] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[268] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2016.

[269] Sarah Thorne. Hey Siri, tell me a story: Digital storytelling and AI authorship. *Convergence*, page 1354856520913866, 2020.

[270] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020.

[271] Alice Toniolo, Timothy J Norman, Anthony Etuk, Federico Cerutti, Robin Wentao Ouyang, Mani Srivastava, Nir Oren, Timothy Dropps, John A Allen, and Paul Sullivan. Supporting reasoning with different types of evidence in intelligence analysis. In *Proceedings of the 14th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 781–789. IFAAMAS, 2015.

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

[273] Sara L Uckelman. Deceit and indefeasible knowledge: The case of dubitatio. *Journal of Applied Non-Classical Logics*, 21(3-4):503–519, 2011.

[274] Rainer Unland. Software agent systems. In *Industrial Agents*, pages 3–22. Elsevier, 2015.

[275] Marco Valtorta, Jiangbo Dang, Hrishikesh Goradia, Jingshan Huang, and Michael Huhns. Extending heuer's analysis of competing hypotheses method to support complex decision analysis. In *Proceedings of the 2005 International Conference on Intelligence Analysis (IA-05)*. Citeseer, 2005.

[276] Estee Van der Walt, Jan HP Eloff, and Jacomine Grobler. Cyber-security: Identity deception detection on social media platforms. *Computers & Security*, 78:76–89, 2018.

[277] Hans Van Ditmarsch. Dynamics of lying. *Synthese*, 191(5):745–777, 2014.

[278] Hans van Ditmarsch, Petra Hendriks, and Rineke Verbrugge. Editors' review and introduction: Lying in logic, language, and cognition. *Topics in Cognitive Science*, 2020.

[279] Rogier M van Eijk, Frank S de Boer, Wiebe Van Der Hoek, and John-Jules C Meyer. Open multi-agent systems: Agent communication and integration. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 218–232. Springer, 1999.

[280] Tim Van Gelder. What might cognition be, if not computation? *The Journal of Philosophy*, 92(7):345–381, 1995.

[281] Tim Van Gelder. Can we do better than ACH? *AIPIO News*, 55, 2008.

[282] Dieter Vanderelst and Alan Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48:56–66, 2018.

[283] Lav R Varshney, Nitish Shirish Keskar, and Richard Socher. Limits of detecting text generated by large-scale language models. *arXiv preprint arXiv:2002.03438*, 2020.

[284] Renata Vieira, Álvaro F Moreira, Michael Wooldridge, and Rafael H Bordini. On the formal semantics of speech-act based communication in an agent-oriented programming language. *Journal of Artificial Intelligence Research*, 29:221–267, 2007.

[285] Gian Volpicelli. Wikipedia's Jimmy Wales wanted to save journalism. He didn't. *WIRED Magazine*, 2019.

[286] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[287] Aldert Vrij and Pär Anders Granhag. Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, 1(2):110–117, 2012.

[288] Aldert Vrij, Maria Hartwig, and Pär Anders Granhag. Reading lies: nonverbal communication and deception. *Annual review of psychology*, 70:295–317, 2019.

[289] Alan R Wagner and Ronald C Arkin. Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics*, 3(1):5–26, 2011.

[290] D. Walton and E. Krabbe. *Commitment in Dialogue: Basic concept of interpersonal reasoning.* SUNY Press, Albany NY, 1995.

[291] Douglas Walton. *The place of emotion in argument.* Penn State Press, 2010.

[292] Ning Wang, David V Pynadath, and Susan G Hill. The impact of POMDP-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 997–1005. IFAAMAS, 2016.

[293] Jacqueline Kory Westlund and Cynthia Breazeal. Deception, secrets, children, and robots: What's acceptable. In *Proceedings of the Workshop on The Emerging Policy and Ethics of Human-Robot Interaction @HRI*, 2015.

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

[295] Robert A Wilson and Lucia Foglia. Embodied cognition. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[296] Michael Winikoff. Jack$^{\text{TM}}$ intelligent agents: an industrial strength platform. In *Multi-Agent Programming*, pages 175–193. Springer, 2005.

[297] Michael Wooldridge. Verifiable semantics for agent communication languages. In *Proceedings of the International Conference on Multi Agent Systems (Cat. No. 98EX160)*, pages 349–356. IEEE, 1998.

[298] Ben Wright, Mark Roberts, David W Aha, and Ben Brumback. When agents talk back: Rebellious explanations. In *Proceedings of the Workshop on Explainable Planning @ICAPS*, 2019.

[299] Liang Wu and Huan Liu. Detecting crowdturfing in social media. *Encyclopedia of Social Network Analysis and Mining*, pages 1–9, 2017.

[300] Grace Hui Yang and Yue Yu. Use of interpersonal deception theory in counter social engineering. In *Proceedings of the International Workshop on Rumours and Deception in Social Media*. CEUR-WS, 2018.

[301] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265. IEEE, 2019.

[302] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1143–1158. ACM, 2017.

[303] Eliezer Yudkowsky. The AI-box experiment. *Singularity Institute*, 2002.

[304] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, pages 405–417. ACM, 2017.

[305] Frank Zenker. From Stories-via Arguments, Scenarios, and Cases-to Probabilities: Commentary on Floris J. Bex's "The Hybrid Theory of Stories and Arguments Applied to the Simonshaven Case" and Bart Verheij's "Analyzing the Simonshaven Case With and Without Probabilities". *Topics in cognitive science*, 2019.