university of groningen / faculty of mathematics and natural sciences / artificial intelligence

The Influence of Lying in a Negotiation Setting: Colored Trails

A Master's Thesis

Presented by

Sverre Brok

August 30, 2023

Internal Supervisor(s):
Prof. dr. L.C. (Rineke) Verbrugge
Dr. H.A. (Harmen) de Weerd

Artificial Intelligence (Multi-Agent Systems)
University of Groningen, The Netherlands

Abstract

With the growing capabilities of artificial agents, it is essential to shed light on the extent intelligent systems are willing to lie in negotiations. The current study examines lying in the game of Colored Trails, which is a negotiation setting in which two agents want to reach their own goal location that the other agent doesn't know and take turns proposing a new distribution of resources.

Previous research suggests that a theory of mind capability, that is, the capability of an agent to attribute mental content such as beliefs to another agent, is needed for an agent to lie. The Colored Trails setting has previously been used to investigate the benefits of theory of mind in negotiations. We extend this research by introducing agents that are capable of theory of mind with the ability to lie. Since the classical framework of Colored Trails did not enable agents to lie, we incorporated the possibility for both agents to send a message that tells the receiver that a particular goal location is the sender's goal location.

This thesis presents two main contributions. First, a graphical user interface has been made where the behavior of the agents in Colored Trails can be analyzed. Second, different experiments have been performed to examine to what extent agents that can lie are able to achieve better outcomes than their honest counterparts.

The results of the experiments show that, in general, higher orders of theory of mind provide more benefit to an agent in Colored Trails than the ability to lie. Moreover, comparing our work with previous research revealed that, in the uncertain environment of our Colored Trails setting, even the ability to send goal location messages influenced the results to a lesser extent than the theory of mind capability of an agent.

Overall, an honest agent performed as well as an agent with the ability to lie. The results contribute to a foundation for making artificial intelligence more trustworthy. Future research could extend this work by investigating more specific conditions in Colored Trails where lying might be beneficial or expanding this work to other frameworks.

Acknowledgements

Although saying it might be a bit cliché, this work would certainly not have been possible without the support of others. First and foremost, I would like to express my sincere gratitude to Rineke Verbrugge and Harmen de Weerd for all the valuable meetings we have had. They both provided constant encouragement, feedback, and invaluable insights throughout the research process. Whenever I felt stuck, they both stayed positive and helped me persevere and gain new ideas. A special thanks to Rineke Verbrugge for consistently offering constructive feedback and to Harmen de Weerd for providing me with his code and guidance.

The computational resources for this study were made possible by the Center for Information Technology (2023). I am grateful for being granted access to the high-performance computing facilities, which were crucial in conducting the extensive experiments central to this thesis.

Finally, I would like to thank all others who, in various ways, contributed to the final product of this thesis. Your support, whether through discussions, offering words of encouragement, having lunch breaks, or distracting me from the hard work, has been deeply appreciated.

Contents

1 Introduction 1

2 Related Work 7

2.1 On the interface of communication and artificial intelligence 8

2.2 Negotiating agents 10

2.3 Lying agents 12

2.4 Theory of mind 13

2.5 Colored Trails 16

3 Methodology 19

3.1 Game setting: Colored Trails 19

3.2 Theory of mind in Colored Trails 20

3.2.1 Initial offer 23

3.2.2 Model of zero-order theory of mind 24

3.2.3 Model of first-order theory of mind 25

3.2.4 Model of higher-order theory of mind 27

3.2.5 Learning within games 29

3.2.6 Learning across games 32

3.3 Lying in Colored Trails 33

3.3.1 Communication 37

3.3.1.1 Receiving a goal location message 38

3.3.1.2 Sending a goal location message 42

3.3.2 When to send a goal location message? 43

3.3.3 Lying and misleading agents 45

3.3.3.1 Exceptions and conventions on lying and misleading . . 46

3.4 Implementation 48

3.5 Experiments 51
3.5.1 Experiments settings and parameter table 52
3.5.2 Experiment 1: Setting the probability of a ToM_0 agent sending messages 54
3.5.3 Experiment 2: Does lying and misleading outperform honesty? . 55
3.5.4 Experiment 3: Is there a benefit to sending goal location messages in Colored Trails? 55
3.5.5 Experiment 4: Varying the learning speed 56
3.5.6 Experiment 5: Games where a Pareto improvement exists 56

4 Results 57
4.1 Experiment 1: Setting the probability of a ToM_0 agent sending messages . 58
4.1.1 Separating ToM_0 initiators and ToM_0 responders 63
4.2 Experiment 2: Does lying and misleading outperform honesty? 65
4.2.1 Investigating other metrics 69
4.2.2 Comparing the score gain for each combination of agent types . . 73
4.3 Experiment 3: Is there a benefit to sending goal location messages in Colored Trails? 81
4.4 Experiment 4: Varying the learning speed 84
4.5 Experiment 5: Games where a Pareto improvement exists 87

5 Discussion & Conclusion 93
5.1 Conclusion 97
5.2 Future work 97

References 101

A The Concept of Pareto Efficiency 113

B Additional Plots 115
B.1 Experiment 1: Setting the probability of a ToM_0 agent sending messages . 115
B.2 Experiment 2: Does lying and misleading outperform honesty? 118
B.3 Experiment 4: Varying the learning speed 119

C Installation 121
C.1 Windows 122
C.1.1 Eclipse 122

C.1.1.1 Installation Eclipse 122
C.1.1.2 Run the graphical user interface 122
C.1.2 IntelliJ IDEA 126
C.1.2.1 Installation IntelliJ IDEA 126
C.1.2.2 Run the graphical user interface 126
C.2 Linux 127

1 Introduction

Programs using Artificial Intelligence (AI) that communicate with people such as Alexa, ChatGPT, and Bard are becoming more commonplace. An article by IBM (n.d.) starts with: “AI is no longer the future – it’s now here in our living rooms and cars and, often, our pockets.” Unfortunately, with these AI tools, the spread of misinformation is growing (Hurst, 2023). Besides the spread of misinformation, agents are already claimed to be able to lie to humans (Kneer, 2021; Rogers, Webber, & Howard, 2023). A recent article investigates how trustworthiness can be regained by humans in human-robot interactions after a robot tells a lie (Rogers et al., 2023).

An example where agents can use a lying strategy is negotiation. It is not new that artificial agents are used in negotiations (Kraus, 1997). What is new, is that these AI systems become more sophisticated and that artificial agents get integrated into more fields, such as smart buildings (Li, Logenthiran, Phan, & Woo, 2017). Within these smart buildings, agents negotiate over the energy supply and demand to provide optimal energy usage and minimal electricity costs. AI agents may use lying in such negotiations to obtain a better position. If we aim to (re)gain trust in AI, we need transparency of the AI system and examine when and in what ways an AI system is “willing” to lie. But first, we will delve into the definition of lying.

Definition of Lying. The concept of lying is familiar to most human beings. Lying is a fact of daily life (DePaulo & Kashy, 1998; DePaulo et al., 2003). Intuitively and generally speaking, people care about social norms and honesty; however, in daily diaries, adults report telling on average one or two lies a day (DePaulo & Kashy, 1998; DePaulo et al., 2003). People lie about their feelings, preferences, attitudes, opinions, achievements, failures, and so forth. Lies can but do not have to be harmful. Examples of lies that may not be considered harmful are so-called *white* lies. A parent might tell white lies to their child to increase the child's confidence. Just as common might be lies to children about whether Santa is real. While many everyday lies may not be especially harmful to other people, other lies may harm close friendships.

Deception may not be limited to humans only. Some forms of lying and deception have evolved in communication by other animals as well. An example of an animal that uses deceit in communication is the firefly of the Photuris species (El-Hani, Queiroz, & Stjernfelt, 2010; Peterson, 2011). The light of the Photuris species is hard to distinguish from the light of the Photinus species, even for other fireflies. (The light of the two species may be hard to distinguish, but the names too.) The Photinus firefly uses light to communicate with other Photinus fireflies and to attract mates. Male fireflies of the Photinus species send out signals and wait for a Photinus female to respond. When they find a Photinus female firefly responding to their signal, they approach the source to mate. The Photuris firefly adapted to this behavior and emits deceitful light signals that mimic a female Photinus firefly. The Photuris firefly thus deceives and attracts male Photinus fireflies. Whenever the Photinus male comes too close, the Photuris firefly, which is larger than the Photinus firefly, eats the male Photinus firefly.

This example of the Photuris firefly deceiving the Photinus firefly involves evolution. Another example of deception that has evolved by some animals, such as the chameleon, is the use of camouflage to blindside their prey or hide from predators (Green, Duarte, Kellett, Alagaratnam, & Stevens, 2019). Many argue that there is no intentional aspect of inducing a false belief in a target in these examples (Hyman, 1989), and it may thus not be considered a form of lying. Chevalier-Skolnikoff (1986), however, has argued that there are primates - adult chimpanzees - that can develop cognitive capacities sufficient for intentional deception, at least in some forms. Besides primates, Corvids show signs of "tactical" deception by withholding intentions and providing false information (Bugnyar & Heinrich, 2006; Bugnyar & Kotrschal, 2002). Corvids may pretend to cache food in a location while actually hiding it elsewhere to deceive other Corvids that observed the caching behavior and later try to steal the food.

Lying is a phenomenon that comes in many forms, so its definition is not so obvious and it is hard to define. The dictionary definition (Merriam-Webster, n.d.) of lying is:

> *"to make an untrue statement with intent to deceive"*.

There are numerous problems with this definition. According to Mahon (2016), (a similar version of) this definition is both too narrow and too broad. It is too narrow, as the definition requires falsity of the statement. According to this definition, someone that makes a true statement but believes that said statement is false is not lying although the statement was intended to deceive. The definition is too broad, as it allows for intended deceit about something other than what is being stated without intended deceit about what is actually stated. For example, the intent of the speaker might be that the addressee does not have to believe the said statement, but rather the speaker intends to deceive the addressee about something else.

Definitions of lying have been discussed at length (see, e.g., Fallis, 2009; Mahon, 2016; Van Ditmarsch, Hendriks, & Verbrugge, 2020). The definition given by Van Ditmarsch et al. (2020) is:

Definition 1 (Lying)**.** You lie if you say something that you believe to be false with the intention that the addressee believes that you and the addressee commonly believe that it is true (Van Ditmarsch et al., 2020).

Using this definition, the liar intends that the *addressee* must not only believe that the statement said by the *liar* is true but also that the *addressee* believes that the *liar* believes that the statement is true. Additionally, the *addressee* must believe that the *liar* believes that the *addressee* believes the statement is true, and so on. The result of this "and so on" reasoning is called common belief (Van Ditmarsch, Van Eijck, & Verbrugge, 2009) of which the definition of lying given above is a result (Van Ditmarsch et al., 2020).

Van Ditmarsch et al. (2020) make a few distinctions between lying and closely related phenomena. Lying is a verbal act, and it should be distinguished from deceiving without lying, bullshit (nonsense), bluffing (uncertainty), white lies (intention is good), and omissions. Moreover, there is a difference between lies, on the one hand, and metaphoric and ironic statements, on the other hand. The speaker in the former intends to deceive, while the intention of the latter is for the listener to recognize the falsehood. For specific definitions and examples of these differences, we refer to Van Ditmarsch et al. (2020). Additionally, we refer to Frankfurt (2005) for a comprehensive explanation of the difference between bullshit and lying. Briefly, Frankfurt (2005) differentiates a bullshitter from a liar by their lack of concern of the truth.

While definitions of lying and deception differ between scholars, we distinguish between deception and lying as follows. According to Mahon (2016), deception is defined as follows:

Definition 2 (To deceive)**.** To intentionally cause to have a false belief that is known or believed to be false (Mahon, 2016).

In contrast to lying, Mahon (2016) calls deception an achievement or success verb. While a false belief of the receiver must be achieved to call something deception, lying is concerned with the "intent" of the speaker.

The present studies. In this thesis, we investigate the influence of lying in negotiations. This is done by modeling agents in a multi-agent system capable of lying. Agents that are capable of lying will have a *theory of mind* capability. Theory of mind is what Premack and Woodruff (1978) termed the capability of someone being able to attribute mental states to someone else, such as beliefs, desires, knowledge, goals, and so forth. Using Definition 1 of lying, theory of mind is required for lying to capture the intent of an agent to deceive the other agent with its lie. We will look into lying as a possible negotiation strategy within a (mixed-motive) framework where both cooperation and competition are important to reach a beneficial outcome for the participating agents.

The modeling will be done in the influential setting of the *Colored Trails* game, first introduced by Grosz, Kraus, and colleagues (2004; 2010). In this negotiation setting, agents alternate in making offers of the distribution of the available colored chips with which they can obtain points. Agents aim to obtain the highest number of points possible. Because we aim to include lying in this setting, which is not possible by only making offers, we introduce the concept of sending goal location messages in Colored Trails. As such, agents will be able to communicate their goal position to the trading partner besides making offers.

Now that the main concepts of this thesis are introduced, we provide the following core research question:

> *What is the influence of lying by artificial agents in the multi-agent negotiation setting of Colored Trails?*

In particular, we seek an answer to what extent agents capable of lying outperform similar agents that are not capable of lying.

The remainder of this thesis is organized as follows. In Chapter 2, we present related work to lying in AI. Since the definition of lying we adopt requires a theory of mind capability, we also discuss some research on theory of mind. Additionally, we discuss related research that used the negotiation setting of the Colored Trails game. In Chapter 3, we explain the specific Colored Trails game that we have adopted, how agents can use theory of mind, and how agents can lie in our model. The graphical user interface we developed in Java is discussed in this chapter, as well as the experiments we performed. The results of these experiments can be found in Chapter 4. Finally, we conclude and discuss in Chapter 5.

2
Related Work

The topic of lying was already of interest in ancient times. Many philosophical works on lying quote the church father St. Augustine, who was active around the 4th century and analyzed lying in his work called *De Mendacio* (Augustine, 1956). Another example of early interest in lying is the so-called Liar Paradox (see, e.g., Beall, Glanzberg, & Ripley, 2020) that is credited to the Cretan prophet Epimenides of Knossos and dates back to around 600 BC. The paradox follows from the statement of Epimenides, *a Cretan*, who says: “All Cretans are liars”. This statement was long considered to have no reasonable truth value, since both the statement being true and it being false would lead to a contradiction. However, Van Ditmarsch et al. (2020), for instance, justify that this seeming paradox is not necessarily a paradox since the negation of “All Cretans are liars” is not “All Cretans tell the truth” but “Some Cretan tells the truth”, and this Cretan does not have to be Epimenides. Therefore, Epimenides’ statement can simply be false. A more modern example (and a *real* paradox) is “This sentence is false”. If this sentence is false, it must be true, but if this sentence is true, it must be false. The research question of how to resolve these paradoxes has still been asked in recent discussions (Beall et al., 2020).

Current research on lying includes the article by Van Ditmarsch et al. (2020) where some recent trends in research on lying are described from a multidisciplinary perspective.

They argue that a comprehensive account of lying requires a multidisciplinary approach, since the act of lying involves many aspects. For example, among other skills needed to tell a lie such as linguistic knowledge, a speaker wishing to tell a lie must compute the change in beliefs brought about by the lie (studied in dynamic epistemic logic). Van Ditmarsch et al. (2020) outline seven articles that focus on various aspects of the human trait of lying, which are included, together with the article of Van Ditmarsch et al. (2020), in a special issue on lying of *Topics in Cognitive Science* (*topiCS*) (Gray, 2020). Another recent example is the broadly multidisciplinary *Oxford Handbook on Lying* (Meibauer, 2019). Compared to Meibauer (2019), Van Ditmarsch et al. (2020) more strongly focus on the logical aspect of lying as well as the philosophical, linguistic, and psychological aspects of lying.

In the remainder of this chapter, we discuss some literature on communication in AI in Section 2.1 and, more specifically, on negotiation in AI in Section 2.2. In Section 2.3, we will see that AI systems already attempt to lie and deceive. In particular, we will look at why it is important to investigate agents that are capable of lying. While the "intent" of an AI system to deceive might still be debatable (Livet & Varenne, 2020; Roff, 2020),[1] AI agents become more sophisticated, so if they don't lie and deceive already, they might be able to in the near future. Next, since we consider lying to require the use of theory of mind, we discuss some related work on theory of mind in Section 2.4. We conclude this chapter with Section 2.5 about related research within our negotiation framework, the setting of Colored Trails.

2.1 On the interface of communication and artificial intelligence

While there are many definitions of communication, in this thesis we focus on communication expressed as messages in linguistic form. We thereby exclude behavioral communication such as communication between fireflies, where fireflies transmit light signals to effect a certain response, as mentioned in the introduction of this thesis. Historically speaking, communication has been conceptualized mostly as a human process (Dance, 1970). However, nowadays communication is an important aspect of computer science and AI as well, as it enables an AI system to interact with its environment and

[1] There is still an ongoing (philosophical) debate about whether AI systems can think (Livet & Varenne, 2020). This debate also captures the discussion about whether these systems can hold beliefs and intentions, and therefore whether they can have an *intent* to deceive.

make decisions based on the information it receives. Communication in AI refers to the exchange of information between AI systems and humans, or between different AI systems. Examples of communication between AI systems and humans are dialogue systems such as chat-bots or Siri. Communication between machines, i.e., machine-to-machine communication, involves the exchange of data and information between different AI systems. Machine-to-machine communication is, for example, a critical component of the Internet of Things, where billions of physical devices around the world are connected, enabling them to communicate real-time data without involving a human being (Ranger, 2020). Other examples of machine-to-machine communication are autonomous vehicles that communicate with one another to avoid collisions, industrial sensors that communicate with one another to optimize processes, and a smart grid system where devices communicate with one another to optimize energy distribution.

Communication plays an important role in multi-agent systems (Wooldridge, 2009). Communication is needed to share information and knowledge, and agents can perform communicative actions in an attempt to influence or alter the mental state of other agents (Wooldridge, 2009). There exist various agent communication languages. Two commonly used agent communication languages are KQML and FIPA-ACL (for a review on these two agent communication languages, see, e.g., Soon, On, Anthony, & Hamdan, 2019). Agent-based models have also been used to explore the origins and evolution of communication and language (Scott-Phillips, Kirby, & Ritchie, 2009; Steels, 2003, 2011). In addition, De Weerd, Verbrugge, and Verheij (2015) use agent-based models to determine to what extent higher orders of theory of mind help agents to establish effective communication.

Grice's maxims of communication (the maxim of quantity, quality, relation, and manner) describe how people intuitively communicate (Grice, 1975). These four maxims also provide a framework for effective communication. Grice was also aware that these maxims are often not respected. For example, according to the maxim of quality, one should not tell a lie.

For humans, there is more to communication than simply communicating information to another human with a plain verbal statement. Besides a clear verbal statement, non-verbal communication such as intonation and body language play a role in how a statement is received (Mehrabian, 1971). Effective communication in negotiations, for example, requires more than solely effectively communicating your interests. It also requires active listening, an openness to different perspectives, and building trust. Raiffa, Richardson, and Metcalfe (2002) describe a thorough analysis of such collaborative decision making. While non-verbal communication may be less fundamental in machine-to-machine communication, it is important to keep in mind other aspects of communication in dialogues between AI systems and humans.

2.2 Negotiating agents

Negotiation is the process of joint decision making, and it is widely studied in AI (Baarslag et al., 2013; Chen & Weiss, 2012; Jennings et al., 2001; Kraus, 1997; Parsons, Sierra, & Jennings, 1998; Sierra, Jennings, Noriega, & Parsons, 1997; Weiss, 1999). Parsons et al. (1998) propose a framework for negotiation for autonomous agents.[2] Weiss (1999) provides a book on distributed artificial intelligence, which is an important subject in industrial and commercial applications and involves agents that cooperate and negotiate. Jennings et al. (2001) developed a generic framework for classifying and viewing automated negotiations. Chen and Weiss (2012) introduce an effective approach called OMAC for automated negotiation in complex environments. There also exist challenges to advance the state-of-the-art in the area of negotiating agents. An example of such a challenge is the (Second) International Automated Negotiating Agents Competition (ANAC 2011). Analysis and insights gained from this challenge are presented by Baarslag et al. (2013). Upcoming application domains for negotiating agents are, for instance, autonomous driving or the smart electrical grid, where agents have to negotiate about the division of the available electricity (see, e.g., Alam, Gerding, Rogers, & Ramchurn, 2015). While many successes have been made in the field of autonomous negotiating agents, fully-deployed and truly autonomous negotiators may not yet be present (Baarslag, Kaisers, Gerding, Jonker, & Gratch, 2017).

The article of Baarslag et al. (2017) discusses various challenges and opportunities for (almost) entirely autonomous negotiators. According to Baarslag et al. (2017), we can distinguish three strands of research in automated negotiation. The first strand is *Negotiation support systems*, where people are assisted and trained in negotiation by these systems (see, e.g., Johnson, Gratch, & DeVault, 2017). The second strand is *Game-theoretical approaches and trading bots*. This part of the research focuses more on equilibrium strategies according to game theory, and on algorithmic trading and bidding by agents, for instance, in the financial sector (see, e.g., Wellman, Greenwald, & Stone, 2007). The third strand is *Negotiation analytical approaches* that considers agents to create a belief about the opponent types or strategies. Because of these three strands of research,

[2] While there are many definitions of the word *agent*, we adopt the definition as given by Wooldridge (1999): "*An agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives.*"

we now have agents that can exist independently in the real world, choose independently among a set of strategies, and engage in supportive interdependence; however, not all at the same time. To enhance research on (fully-)autonomous negotiation, Baarslag et al. (2017) argue that three major themes have to be aligned and advanced together: accurate representation, long-term perspective, and user trust.

Another active area of research within multi-agent negotiation is argumentation-based communication, which is defined as the exchange of arguments between autonomous agents in negotiations to reach a shared understanding or make a decision (Rahwan et al., 2003; Sierra et al., 1997). Argumentation-based negotiation is commonly used in multi-agent systems where agents may have different preferences, beliefs, or goals, and need to negotiate and cooperate to make a decision or resolve a conflict. Argumentation-based communication requires agents to reason about uncertain or incomplete information and to handle conflicts or inconsistencies in a structured manner. Kraus, Sycara, and Evenchik (1998) propose a formal logical framework of mental attitudes of agents based on argument types identified from human negotiation patterns. Kraus (1997) describes some of her group's projects where they have successfully taken an interdisciplinary approach to build coordinated and cooperative intelligent agents by combining AI techniques with techniques from fields such as game theory, operations research, and philosophy. Rahwan et al. (2003) identify the main research motivations and ambitions behind work in the field of argumentation-based communication and provide the main challenges and open questions in the field.

Recently, AI methods have also been applied in complex games that involve cooperation and negotiation such as Diplomacy (Bakhtin et al., 2022; Kramár et al., 2022). An important and challenging part of the board game Diplomacy is the negotiation phase, where humans coordinate their actions to both cooperate and compete with each other. In previous research on Diplomacy with multi-agent systems, the "negotiation" phase was eliminated (see, e.g., Paquette et al., 2019), and thus an explicit communication channel between agents was not considered. Kramár et al. (2022) constructed agents for Diplomacy that can communicate the plans they have for future steps in the game. While Kramár et al. (2022) do not model lying or deceiving agents explicitly, the agents were able to deviate from their contracts made in the game (and might thus be considered dishonest). Bakhtin et al. (2022) present an AI agent, named Cicero, that achieved human-level performance in the game of Diplomacy in an online league of human players, where Cicero ranked in the top 10% of participants who played more than one game. Using both a dialogue and strategic reasoning module, Cicero was able to form "intents" and communicate its corresponding (dis-)honest message. In the next section, we will discuss research that focuses more on lying agents.

2.3 Lying agents

The idea of machines being able to lie and deceive may originate from Turing's imitation game in 1950 (Turing, 1950), where a machine is tested to send messages indistinguishable from a human participant. Castelfranchi (2000) predicted in 2000 that "there will be problems of deception – and consequently of trust – not between humans (via machines) but between humans and artificial entities and among the artificial agents themselves. Agents are and will be designed, selected, or trained to deceive, and people will be deceived by and will deceive their own agents."

For the moment, let us not consider the subtle difference (or ambiguity) between definitions of lying and deception that exist in the literature. Recent developments in the growing area of research at the interface between deception and AI are presented by Masters, Smith, Sonenberg, and Kirley (2021). The complexities that arise with machine deception are addressed by Sarkadi (2018). Although deception by machines is still primarily used as a tool by humans (Masters et al., 2021), it might not be a long shot before autonomous agents can deploy deceptive and lying behavior without human intervention, along with the inherent risks (Sarkadi, 2018). Hence, one of the main reasons to study and model lying and deceptive behavior is that it could be adopted by autonomous and malicious (software) agents (see, e.g., Sarkadi, Panisson, Bordini, McBurney, Parsons, & Chapman, 2019). When we understand all the possible ways an AI agent can lie or deceive, only then might we be able to mitigate unwanted deception by AI agents.

Panisson, Sarkadi, McBurney, Parsons, and Bordini (2018) are some of the first to attempt to model agent attitudes such as lies, bullshit, and deception in the practical context of an Agent-Oriented Programming Language. The dishonest agents provided by Panisson et al. (2018) work under the assumption of complete certainty. A perfect model of the other party might, however, in most practical cases not be available. Hence, Sarkadi, Panisson, Bordini, McBurney, Parsons, and Chapman (2019) follow up on this research and design, implement, and evaluate a model for deception in which agents engage in mental simulation (theory of mind) to determine the optimal deceptive action. Moreover, in their multi-agent system, they integrate components of two major theories of deception, namely *Information Manipulation Theory 2* and *Interpersonal Deception Theory*. In the model of deception by Sarkadi, Panisson, Bordini, McBurney, Parsons, and Chapman (2019), deceptive behavior does not have to be a lie.

When we limit ourselves to lying behavior only, there is little research on modeling lying agents. For one, this might be due to differences in definitions that researchers adopt for deceptive and lying behavior (Masters et al., 2021). Panisson et al. (2018) model lies explicitly besides deception, but use a rather simple definition of lying. They model lying as only making a false statement (without intent to deceive). Others have also tried to model lies (Caminada, 2009; Sklar, Parsons, & Davies, 2005).

Lying is also attempted to be included in logic (Ågotnes, van Ditmarsch, & Wang, 2018; Van Ditmarsch, 2014; Van Ditmarsch, Van Eijck, Sietsma, & Wang, 2012). Ågotnes et al. (2018) mainly investigate true lies (announcement of something false that makes it true) in the setting of Gerbrandy's logic (Gerbrandy & Groeneveld, 1997) of believed (public) announcement logic, wherein agents may have or obtain incorrect beliefs. Besides the analysis of true lies, they present results relating to lying in general. Ågotnes et al. (2018), however, restrict themselves to a simple definition of lying where they call an announcement a lie if the announced formula is false. Van Ditmarsch et al. (2012) do take into account the intent to deceive (more specifically, the intent to be believed) in the definition of lying, and model lying as a communicative act intended to change the beliefs of the agents in a multi-agent system. Van Ditmarsch (2014) discusses some further literature on the intentional aspect of lying but also discusses some alternatives on the view of taking intention into account. Moreover, Van Ditmarsch (2014) proposes various logics of lying for different speaker perspectives (the agent who is lying) and addressee perspectives.

Verbrugge and Mol (2008) experimented with humans playing a competitive game where a limited form of communication could be used to mislead the other player. They investigated the tactics and related this to the order of theory of mind used by a player. In the next section, we will look more closely at work related to theory of mind.

2.4 Theory of mind

Theory of mind (Premack & Woodruff, 1978) is the capability of an individual to attribute mental content to others such as beliefs, intentions, knowledge, and goals. Without this theory of mind, so-called zero-order theory of mind, individuals are limited to reasoning about world facts only. An individual capable of zero-order theory of mind only understands sentences such as "Alice is reading a book". First-order theory of mind allows individuals not only to reason about world facts but also to reason about the unobservable mental content of others. An individual capable of first-order theory of mind should understand, for example, a sentence that includes the reason why Alice is reading: "Alice is reading a book since she believes reading strengthens her brain". Finally, an individual's ability to recursively model the mental states of other individuals is called higher-order theory of mind (Verbrugge, 2009). In this thesis, we follow Verbrugge (2009) and say that higher-order theory of mind is second-order theory of mind or higher. Using higher orders of theory of mind, individuals understand sentences such as "Bob does not know that Alice believes that reading strengthens her brain". In this example, an individual capable of second-order theory of mind can attribute mental content to what Bob does not know about the beliefs of Alice.

The human ability to make use of higher orders of theory of mind is well-established experimentally (see, e.g., Miller, 2009; Perner & Wimmer, 1985). Possible explanations for the emergence of social cognition, which includes theory of mind and lying, are the Machiavellian intelligence hypothesis (Whiten & Byrne, 1988), the Vygotskian intelligence hypothesis (Vygotsky & Cole, 1978), and the mixed-motive interaction hypothesis (Verbrugge, 2009).

According to the Machiavellian intelligence hypothesis, there is a competitive advantage to the cognitively demanding ability of (higher-order) theory of mind. Empirical research using agent-based models shows that there is indeed an advantage of using higher-order theory of mind in competitive settings (see, e.g., Devaine, Hollard, & Daunizeau, 2014; De Weerd, Verbrugge, & Verheij, 2013b).

In contrast, the Vygotskian intelligence hypothesis explains the emergence of higher-order theory of mind in humans in cooperative settings instead of competitive settings. De Weerd, Verbrugge, and Verheij (2015) show that agents with a first-order theory of mind capability achieve a cooperative solution more efficiently compared to a situation where both agents use only a zero-order theory of mind capability. Higher-order theory of mind only helps agents to achieve a cooperative solution more quickly but does not benefit these agents when such a solution has already been found.

The third hypothesis for the emergence of social cognition in humans is the mixed-motive interaction hypothesis. In mixed-motive settings, both cooperation and competition play a role, such as negotiations. While agents may benefit from a theory of mind capability in either purely cooperative or purely competitive settings, it seems more likely that the capability of theory of mind has emerged from mixed-motive settings (De Weerd, 2015). Studies suggest that theory of mind allowed us to survive and deal with more complex and unpredictable environments (De Weerd, Verbrugge, & Verheij, 2017, 2022). Higher-order theory of mind has also been associated with better negotiation skills (De Weerd et al., 2017).

Other research on theory of mind includes but is not limited to the work of Panisson, Sarkadi, and colleagues. In particular, Panisson, Sarkadi, McBurney, Parsons, and Bordini (2019) propose formal semantics for agents to update their theory of mind through communication with other agents. Sarkadi, Panisson, Bordini, McBurney, and Parsons (2019) add uncertainty to the modeling of other agents' minds during communication.

It is clear that theory of mind is needed, for example, to reason about what others believe, or to predict how our actions influence other people. In particular, studies suggest that the ability to tell a lie requires theory of mind (Lavoie & Talwar, 2020; Talwar, Gordon, & Lee, 2007; Talwar & Lee, 2008; Wimmer & Perner, 1983). Talwar and Lee (2008) found that there is a close connection between children's development of lie-telling and their development of theory of mind. When children grow older, the way a lie is told goes through various stages, but in essence, their ability to tell and maintain a lie improves. Lavoie and Talwar (2020) suggest that as children's theory of mind abilities (and working memory) improve, their abilities to conceal information from others develop. Theory of mind allows them to reason about the mental content of the addressee, and thus to tell a consistent story around the lie. Talwar et al. (2007) show that understanding the concept of lying and being able to maintain a lie over time requires second-order theory of mind (e.g., your ability to infer what the addressee believes about your (the lie-teller's) thoughts).

For a lie to be maintained successfully in subsequent statements, the lie-teller must assess the mental state of the addressee to avoid leaking semantic information that contradicts earlier formed beliefs by the addressee. For example, the lie-teller aims to ensure that subsequent statements are consistent with the addressee's beliefs. In the literature, the ability to maintain consistency between verbal statements during deception is referred to as semantic leakage control (Talwar & Lee, 2002). By explicitly modeling the mental state of the addressee, the lie-teller can make statements that are consistent with the mental state that the lie-teller assigns to the addressee, which includes the beliefs the addressee has about the lie-teller's beliefs or plans. For the lie-teller to assign beliefs to the addressee about the beliefs or plans of the lie-teller, the lie-teller needs to be capable of second-order theory of mind.

In the study by Talwar and Lee (2002), children between 3 and 7 years of age were left alone in a room with a music-playing toy placed behind their backs. Before the experimenter left the room and the child was left alone, the child was told not to look at the toy. After the experimenter entered the room again, the experimenter asked the child whether the child had looked at the toy or not. Talwar and Lee (2002) found that the majority of children between three and five years old blurted out the name of the toy while they denied having turned around and peeked at the toy. Children who were six to seven years old were better at semantic leakage control and feigned ignorance of the toy's identity. This clearly shows that younger children are more susceptible to semantic leakage. This result is in line with the theory of mind capability of children.

Just like humans, it is reasonable to assume that software agents require theory of mind to "actively" deceive or to detect deception (Isaac & Bridewell, 2017). Hence, in this thesis, agents that can lie also have the ability to use theory of mind.

2.5 Colored Trails

The Colored Trials setting, introduced by Grosz, Kraus, and colleagues (2004; 2010), is a framework that is commonly used to investigate decision making in mixed-motive situations, that is, situations where the agents have conflicting motives to cooperate or to compete with each other.[3] Colored Trails is a multi-agent system setting where agents negotiate the exchange of resources to achieve their individual goals. A typical setting of the Colored Trails game consists of a five-by-five board of colored tiles, where two or more agents attempt to reach a certain (different) goal location. Using colored chips, the agents may move onto a tile with the same color. An agent obtains points with each step in the shortest path to its goal position, by reaching its goal position, and by having left-over chips.

The Colored Trails setting represents a multi-issue bargaining situation, where each issue is represented by a color, while different paths toward the goal location represent different acceptable solutions. Agents competing in the Colored Trails game have overlapping issues, so both competitive and cooperative elements are involved. Negotiation in such a mixed-motive setting can be seen as the task of sharing a metaphorical pie (Raiffa et al., 2002). Agents participating in Colored Trails aim to both cooperate to achieve a high number of points for both negotiators (enlarge the pie) and compete to reach their individual goal location and have a large number of chips in possession (obtain as large a piece of the pie as possible).

Besides agents, the Colored Trails game can be played by humans or a heterogeneous group of humans and agents (Kraus et al., 2004). In particular, the Colored Trails frame-

[3] Also see https://coloredtrails.atlassian.net/wiki/spaces/coloredtrailshome/.

work is a useful research test-bed for investigating the decision making of agents in a negotiation setting (Gal et al., 2010). Previous research used this framework to investigate the benefits of using theory of mind in negotiations (De Weerd, Verbrugge, & Verheij, 2013a; De Weerd et al., 2017, 2022; Ficici & Pfeffer, 2008).

De Weerd et al. (2013a) include incomplete information in Colored Trails, where agents do not know each other's goal location. In this uncertain setting, they show how theory of mind can present individuals with an advantage over others who lack a theory of mind ability. De Weerd et al. (2013a) argue that this may be a reason why our ability to reason about the mental content of others may have evolved. De Weerd et al. (2017) follow up on this and identify settings in which there is an evolutionary incentive to reason using higher orders of theory of mind, which could explain the emergence of the human-like theory of mind abilities. They show that the use of higher-order theory of mind can be beneficial in settings where agents can observe more of the behavior of their trading partner (multiple rounds of offers), and thus the ability to make use of higher-order theory of mind can be associated with better negotiation skills.

De Weerd et al. (2022) investigate how the predictability of the environment affects the effectiveness of (higher-order) theory of mind. They studied agents with different theory of mind capabilities in the Colored Trails setting as a single-shot bargaining situation where three agents meet one another in a negotiation. Using three types of environments differing in predictability, they show that, within a given setting, theory of mind reasoning is more beneficial when the environment is less predictable.

Finally, Ficici and Pfeffer (2008) investigate the use of theory of mind by humans and construct different computer agents that fit human reasoning. They show that humans use theory of mind by modeling other players to think strategically. Moreover, with an experiment using software agents as trading partners for human participants, De Weerd et al. (2017) show that theory of mind agents can even encourage the use of higher-order theory of mind in human participants (also see De Weerd, Broers, & Verbrugge, 2015), and may thus be used to train people in the application of their theory of mind and negotiation skills.

3
Methodology

This chapter deals with describing our implementation of an extension of the Colored Trails game and the set-up of our experiments. The Colored Trails game has been used in research settings to investigate various aspects of communication and negotiation between agents (see Section 2.5). In our setting, we will use the Colored Trails game to implement lying and investigate the influence of lying.

In Section 3.1, our Colored Trails game is described in detail. Then, in Section 3.2, we discuss the implementation of theory of mind in the Colored Trails setting, adapted from De Weerd et al. (2017). We discuss how our agents can lie in Section 3.3. Finally, we describe our graphical user interface and experiments in Section 3.4 and Section 3.5, respectively.

3.1 Game setting: Colored Trails

We will adopt the typical setting of the Colored Trails game consisting of a five-by-five board of colored tiles, where two agents attempt to reach a certain goal location using colored chips (De Weerd et al., 2013a, 2017). Both agents start in the center square and get assigned a certain number of colored chips similar to the colors of the tiles of the board. A colored chip can be used by an agent to move onto an adjacent tile with the same color to get closer to its goal location. The ultimate goal of an agent is to get the highest number of points possible. An agent obtains points by moving toward its goal location, reaching its goal location, and by having leftover chips after moving toward its goal location. An agent obtains 100 points for every tile it can move onto in the shortest path from its starting location to its goal location. When an agent reaches its goal location, it obtains 500 points. An agent receives 50 points for every colored chip it has left.

At initialization of the game, the tiles of the game board are randomly colored from five different colors. Moreover, each agent obtains four randomly colored chips, and each agent gets assigned a goal location randomly chosen from twelve possible goal locations. The possible goal locations are three or four steps away from the center square. The starting position is the same for both agents and is the center square of the game board. The goal locations of the agents can be different but also the same.

An example of the described Colored Trails game board is given in Figure 3.1. Agent i can take two steps toward its goal location using a yellow and a fuchsia-colored chip. After these two steps, agent i is left with two chips. This gives agent i a total of 300 points with the initial distribution of colored chips. In contrast, agent r can take one step toward its goal location using a yellow chip and is left with three chips. This gives agent r a total of 250 points with the initial distribution of colored chips.

While an agent knows its own goal location, the goal location of the trading partner is not known by the agent. Hence, an agent has incomplete information about the game. An agent is thus uncertain about the preferences of its trading partner. By negotiating, agents can offer a new distribution of chips. Agents alternate in making offers, and either agent can make a new offer, accept the previous offer, or withdraw from negotiation. Both agents cooperate and compete to get a higher number of points. However, to stimulate the negotiation process, one point is subtracted from both agents for every offer that is made.

3.2 Theory of mind in Colored Trails

In this section, we describe the model used for agents to achieve a theory of mind capability following De Weerd et al. (2017). This model allows for agents to generalize over different games of Colored Trails and allows for sequential games, that is, a Colored Trails game where the agents make offers one after the other. We follow De Weerd et al. (2017) and say that

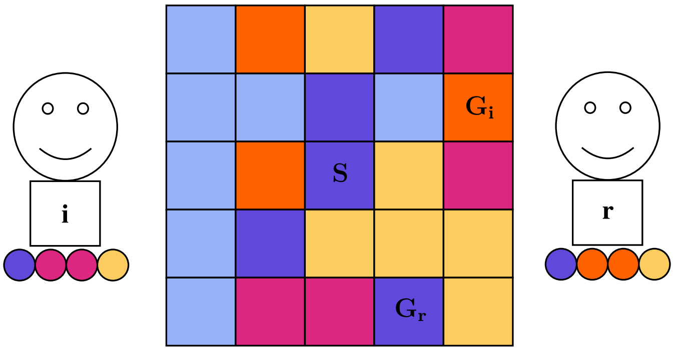


Figure 3.1: Example setup for a five-by-five Colored Trails game with two agents, agents i and r. The board consists of the following five colors (top row, from left to right): light blue, dark orange, dark yellow, purple, and fuchsia (dark pink). The starting position of both agents is denoted with S, and the goal positions of agents i and r are denoted with G_i and G_r, respectively. Both agents i and r have one chip short to reach their goal, that is, agent i needs a dark orange chip, while agent r needs a fuchsia-colored chip. If the two were to exchange a fuchsia chip for a dark orange chip, both agents would reach their goal location, which increases the metaphorical pie. This can be achieved by negotiating over the distribution of colored chips.

> *"an agent achieves theory of mind by taking the perspective of its trading partner, and determining what its own decision would be if the agent had been in the position faced by its trading partner."*

For convenience, we use the shorthand *ToM*$_k$, $k \geq 0$, for an agent that can use theory of mind up to and including the kth-order, but not beyond.

While reading this section, we strongly advise also looking at the paper by De Weerd et al. (2017) as it contains examples and a more elaborate intuition on how higher-order theory of mind may benefit agents. Nevertheless, we tried to be as clear and specific as possible in describing the (implemented) model of De Weerd et al. (2017). For clarity, we indicate occurrences where we deviate from the notation of De Weerd et al. (2017). Moreover, following De Weerd et al. (2017), we omit variables from functions if they can be derived from the context.

Following the representation of De Weerd et al. (2017), a Colored Trails game is a tuple $\mathcal{CT} = \langle \mathcal{N}, \mathcal{D}, \mathcal{L}, \pi_i, \pi_r, D_0 \rangle$, where:

- $\mathcal{N} = \{i, r\}$ is the set of agents;
- $\mathcal{D}$ is the set of possible distributions of chips;
- $\mathcal{L}$ is the set of possible goal locations;

- $\pi_i, \pi_r : \mathcal{L} \times \mathcal{D} \to \mathbb{R}$ are the utility functions for agents i and r, respectively, such that, for example, $\pi_i(l, D)$ is the utility agent i obtains when $l \in \mathcal{L}$ is its goal location and $D \in \mathcal{D}$ is the chip distribution; and
- D_0 is the initial distribution of chips.

In addition, we define $\mathcal{C}$, with $|\mathcal{C}| = 5$, to be the set of possible colors the chips and the tiles of the board can take on.

Note that this representation does not model the board explicitly. Namely, as the utility functions contain information about the board configuration, the board is implicitly modeled through the utility functions of the agents. In particular, utility is a function of a possible goal position and a distribution of chips and represents the score the agent obtains. The score is calculated as mentioned in Section 3.1.

Note that the notation in De Weerd et al. (2017) uses a time index in the utility function indicating that one point is subtracted for every offer that is made. For notational convenience, we can exclude the time index of the utility function and subtract points for making offers only at the end of the negotiation. After the negotiation has ended, the total score of the agent is thus given by the utility score minus the points subtracted from the negotiation (the total number of offers). An important note, however, is that agents do take into account the point subtraction for each offer that they make; this is only not incorporated in the utility function.

To calculate the utility, an agent needs to have a starting position and a goal position. Both the starting position and the goal position are inherently defined in the utility function. By using the utility function, we abstract away from each agent finding the optimal route from its starting position toward its goal position. Hence, we assume that an agent does not make any mistakes in finding the optimal route and that an agent does not consider the possibility of the other agent making such mistakes. Note that, while an agent knows its goal position, it does not know the goal position of the other agent.

By negotiating and making offers, agents can improve their final score. Let $O_t \in \mathcal{D}$ denote an offer made by an agent at time step t, $t \in \{0, 1, 2, \dots\}$. The chips of the trading partner are known to an agent, so both agents know the set of possible offers $\mathcal{D}$. During negotiation, the initiator and responder take turns in making offers. Consequently, we obtain a finite sequence of offers that we denote by $\{O_0, O_1, \dots\}$, followed by withdraw or accept.

In the following subsections, we describe in detail how agents capable of theory of mind play the Colored Trails game. Without loss of generality, we consider the point of view of agent i. By changing the subscripts from i to r (or vice versa), one obtains analogous formulas for the point of view of agent r. Before diving into the mathematical model of the agents, let us describe the general procedure that the agents follow.

The agent that makes the first offer $O_0 \in \mathcal{D}$ is called the initiator, and its trading partner is called the responder. The first offer O_0 is thus always made by the initiator, but the initiator can also choose to withdraw from negotiation instantly (see Section 3.2.1). Both agents form beliefs about the likelihood of an offer going to be accepted. Offers that have been rejected by the trading partner as well as offers that are received give information about the likelihood of an offer going to be accepted by the trading partner. These beliefs are updated during a game by discriminating the offers based on the colors of the chips (see Section 3.2.5). At the beginning of a new round of negotiation, the beliefs are set to the learned behavior of the trading partner (see Section 3.2.6). Besides the beliefs about whether specific offers are going to be accepted by the trading partner, agents that can use theory of mind construct beliefs about the goal position of the trading partner. These goal location beliefs are updated whenever an offer is received. After updating all its beliefs, the agent decides whether to accept the previous offer O_{t-1}, make a counteroffer O_t, or withdraw from negotiation, based on which of the three options yields the highest expected utility. For agent i with goal location $l_i \in \mathcal{L}$, the utility corresponding to accepting the previous offer is $\pi_i(l_i, O_{t-1})$, and the utility corresponding to withdrawing from negotiation is $\pi_i(l_i, D_0)$. The utility corresponding to making a new offer is non-deterministic and is based on the beliefs of the agent.

3.2.1 Initial offer

The initiator makes the first offer. Previous research has shown that making the first offer in negotiations, in general, is influential (Raiffa et al., 2002; Rosette, Kopelman, & Abbott, 2014; Van Poucke & Buelens, 2002). The first offer in Colored Trails is special because the initiator can decide first on negotiating or keeping the initial chip distribution, so it serves as an anchor for the entire negotiation process. In the following sections, we will describe the exact mathematical model of an agent deciding on its action as it depends on the theory of mind capability of the agent, but for now, we point out the difference between making the initial offer to making the other offers.

Each agent constructs beliefs about whether an offer will be accepted. A *ToM*$_k$ agent, $k \geq 0$, calculates the expected value $EV_i^{(k)}$ of making an offer $O \in \mathcal{D}$ (see the next subsections). The agent compares the expected value of making the best possible offer, that is the offer yielding the highest expectation, with the utility of withdrawing. Different from the other steps in the negotiation, the initiator cannot choose to accept the previous offer, since there simply has not been made a previous offer in the current negotiation round. When the expected value of making the best offer is greater than the utility resulting from withdrawing from negotiation, the agent makes the offer and starts the negotiation. Otherwise, the agent withdraws from the negotiation and both agents keep their initial set of colored chips.

3.2.2 Model of zero-order theory of mind

Let us consider agent i as being an agent limited to using zero-order theory of mind, that is, a *ToM*$_0$ agent. A *ToM*$_0$ agent is not able to attribute mental content to the trading partner or reason about the trading partner's goal position. However, a *ToM*$_0$ agent can model the behavior of its trading partner by constructing zero-order beliefs $b^{(0)} : \mathcal{D} \rightarrow [0,1]$ about the likelihood $b^{(0)}(O)$ that the trading partner accepts certain offers $O \in \mathcal{D}$. These zero-order beliefs will be used to get an estimated value of continuing negotiations and are based on observations only.

Using these zero-order beliefs, a *ToM*$_0$ agent will still be able to make reasonable offers by forming an expectation about how its score would change if it were to make a particular offer. More specifically, the expected value a *ToM*$_0$ agent assigns to making counteroffer $O \in \mathcal{D}$ given its goal location $l_i \in \mathcal{L}$ and its zero-order beliefs $b^{(0)}$ is given by

$$EV_i^{(0)}(O, l_i, b^{(0)}) = b^{(0)}(O) \cdot \pi_i(l_i, O) + \big(1 - b^{(0)}(O)\big) \cdot \pi_i(l_i, D_0) - 1. \tag{3.1}$$

This equation results from a summation from the zero-order belief that offer $O \in \mathcal{D}$ is going to be accepted multiplied by the utility corresponding to this offer, $\pi_i(l_i, O)$, and the zero-order belief that offer $O \in \mathcal{D}$ is going to be rejected multiplied by the utility of the initial distribution, $\pi_i(l_i, D_0)$. We subtract one point as making an offer comes with a cost of one point.[1] There is of course the possibility that the trading partner makes a counteroffer, but since a *ToM*$_0$ agent does not attribute mental content to the trading partner, it simply considers two cases: an offer can either be accepted or rejected.

A *ToM*$_0$ agent calculates the expected value of each of the possible offers it can make. Then, the *ToM*$_0$ agent randomly selects an offer that maximizes the expected value, that is,

$$O_t^* = \arg\max_{O \in \mathcal{D}} EV_i^{(0)}(O, l_i, b^{(0)}). \tag{3.2}$$

[1] Because we leave out the time index in the utility function, we subtract here one point for making an offer compared to the notation by De Weerd et al. (2017).

A *ToM*$_0$ agent will decide rationally between making an offer, accepting the previous offer, or withdrawing from negotiation by choosing the option that will yield the agent the highest (expected) score. Mathematically, a *ToM*$_0$ agent will select the option that is in accordance with its response function:

$$ToM_{0i}(O_{t-1}, l_i, b^{(0)}) = \begin{cases} O_t^* & \text{if} \quad EV_i^{(0)}(O_t^*, l_i, b^{(0)}) > \pi_i(l_i, D_0) \text{ and} \\ & \qquad EV_i^{(0)}(O_t^*, l_i, b^{(0)}) > \pi_i(l_i, O_{t-1}) \\ \text{accept} & \text{if} \quad \pi_i(l_i, O_{t-1}) > \pi_i(l_i, D_0) \text{ and} \\ & \qquad \pi_i(l_i, O_{t-1}) \geq EV_i^{(0)}(O_t^*, l_i, b^{(0)}) \\ \text{withdraw} & \text{otherwise.} \end{cases} \tag{3.3}$$

Herein, observe that a *ToM*$_0$ agent chooses to offer O_t^* when the expected score corresponding to this offer is higher than the utility corresponding to accepting the previous offer made by the trading partner and the utility corresponding to withdrawing from negotiation. If this is not the case, but the utility corresponding to accepting the previous offer made by the trading partner is higher than the utility corresponding to withdrawing from negotiation, then the *ToM*$_0$ agent accepts offer O_{t-1}. If both cases are not satisfied, the agent chooses to withdraw from the negotiation. Notice that an agent can't accept any offer made in previous steps of the negotiation except for the very last offer made by the trading partner. So, once an offer is rejected, it cannot be accepted in later steps in the negotiation.

3.2.3 Model of first-order theory of mind

Let us now consider agent i as a *ToM*$_1$ agent. Being a *ToM*$_1$ agent allows the agent to attribute mental content such as beliefs and goals to the trading partner. A *ToM*$_1$ agent can reason about whether the trading partner will accept certain offers using this attributed mental content, that is, a *ToM*$_1$ agent can make predictions about the future behavior of its trading partner using attributed mental states. To achieve this, a *ToM*$_1$ agent can place itself in the position of its trading partner and model what its action would have been if the *ToM*$_1$ agent were in the position of the trading partner.

A *ToM*$_1$ agent can construct first-order beliefs $b^{(1)} : \mathcal{D} \to [0, 1]$ that represent what the zero-order beliefs of the *ToM*$_1$ agent would have been if it had been in the position of his trading partner (De Weerd et al., 2017). More specifically, using the first-order beliefs, a *ToM*$_1$ agent considers that its trading partner believes that the probability of the *ToM*$_1$ agent accepting a given offer $O \in \mathcal{D}$ is $b^{(1)}(O)$. While a *ToM*$_0$ agent only looks at whether or not an offer will be accepted by the trading partner, a *ToM*$_1$ agent can also make predictions about what counteroffer the trading partner could make and take this into account when deciding on its action. Because of the sequential nature of the negotiation process, a *ToM*$_1$ agent can look one step further ahead into the negotiation.

When an agent receives an offer, its beliefs may change. A *ToM*$_1$ agent uses this belief adjustment to model what the trading partner's action will be after making an offer $O \in \mathcal{D}$. In particular, a *ToM*$_1$ agent determines how making offer $O \in \mathcal{D}$ would change its zero-order beliefs if it had been in the position of its trading partner, and makes further calculations using the adjusted first-order beliefs $U(b^{(1)}, O)$ (see also De Weerd et al., 2017). A *ToM*$_1$ agent uses these adjusted first-order beliefs to predict its trading partner's behavior by using the *ToM*$_0$ response function as given by Equation (3.3).

Now, the expected value the *ToM*$_1$ agent assigns to making counteroffer $O \in \mathcal{D}$ given its goal location $l_i \in \mathcal{L}$, its first-order beliefs $b^{(1)}$, and the partner's goal position $l \in \mathcal{L}$ is given by

$$EV_i^{(1)}(l, O) = \begin{cases} \pi_i(l_i, D_0) - 1 & \text{if } \hat{O}^{(1)} = \text{withdraw}, \\ \pi_i(l_i, O) - 1 & \text{if } \hat{O}^{(1)} = \text{accept}, \\ \max\left\{\pi_i\left(l_i, \hat{O}^{(1)}\right), \pi_i(l_i, D_0)\right\} - 2 & \text{otherwise}, \end{cases} \tag{3.4}$$

where

$$\hat{O}^{(1)} = \textit{ToM}_{0r}(O, l, U(b^{(1)}, O)) \tag{3.5}$$

is the offer the *ToM*$_1$ agent expects its trading partner to make in response to offer $O \in \mathcal{D}$. Here, we also subtract one point for every offer that is made. Hence, we subtract two points when the *ToM*$_1$ agent expects its trading partner to make a counteroffer.

Note that we used goal location $l \in \mathcal{L}$ in the above equations for the goal location of the trading partner. The actual goal location of the trading partner is not known, but a *ToM*$_1$ agent can construct beliefs about the goal location of the trading partner. Each offer the trading partner makes reveals some information about its goal location, as a rational trading partner would not offer a distribution that decreases its score. Goal location beliefs are given by a probability distribution $p^{(1)} : \mathcal{L} \to [0, 1]$, where $p^{(1)}(l)$ denotes the likelihood that a *ToM*$_1$ agent assigns to its trading partner having goal location $l \in \mathcal{L}$.

While a *ToM*$_1$ agent has an additional toolkit by modeling the beliefs and goals of the trading partner, a *ToM*$_1$ agent might learn through repeated interactions that its first-order beliefs fail in accurately modeling the behavior of the trading partner. A *ToM*$_1$ agent, therefore, has a confidence variable $c_1 \in [0, 1]$ that denotes the likelihood the *ToM*$_1$ agent assigns to the predictions of its first-order theory of mind. A *ToM*$_1$ agent weighs its predictions about the (expected) utility of an offer according to this confidence variable.

More specifically, the expected value a *ToM*$_1$ agent assigns to making counteroffer $O \in \mathcal{D}$ is given by

$$EV_i^{(1)}(O) = (1 - c_1) \cdot EV_i^{(0)}\left(O, l_i, b^{(0)}\right) + c_1 \cdot \sum_{l \in \mathcal{L}} p^{(1)}(l) \cdot EV_i^{(1)}(l, O). \tag{3.6}$$

A *ToM*$_1$ agent randomly selects an offer that maximizes the expected value, that is,

$$O_t^* = \arg\max_{O \in \mathcal{D}} EV_i^{(1)}(O). \tag{3.7}$$

A *ToM*$_1$ agent will decide rationally between making an offer, accepting the previous offer, or withdrawing from negotiation by choosing the option that will yield the *ToM*$_1$ agent the highest (expected) score. A *ToM*$_1$ agent will thus select the option that is in accordance with its response function:

$$\mathit{ToM}_{1i}(O_{t-1}) = \begin{cases} O_t^* & \text{if} \quad EV_i^{(1)}(O_t^*) > \pi_i(l_i, D_0) \text{ and} \\ & \qquad EV_i^{(1)}(O_t^*) > \pi_i(l_i, O_{t-1}) \\ \text{accept} & \text{if} \quad \pi_i(l_i, O_{t-1}) > \pi_i(l_i, D_0) \text{ and} \\ & \qquad \pi_i(l_i, O_{t-1}) \geq EV_i^{(1)}(O_t^*) \\ \text{withdraw} & \text{otherwise.} \end{cases} \tag{3.8}$$

3.2.4 Model of higher-order theory of mind

Finally, let us consider agent i as a *ToM*$_k$ agent, $k \geq 2$. A *ToM*$_k$ agent is capable of higher-order theory of mind and considers the possibility that its trading partner takes into account that the *ToM*$_k$ agent has beliefs and goals as well. This allows a *ToM*$_k$ agent to manipulate the beliefs of its trading partner when its trading partner uses a lower-order theory of mind. A *ToM*$_k$ agent might also decide to reveal its goal position to the trading partner by choosing its offer such that the trading partner can exclude many other goal locations. Whether a *ToM*$_k$ agent decides to reveal its goal location or manipulate the beliefs of the trading partner depends on which of the options yields the *ToM*$_k$ agent the highest expected score.

Analogously to a *ToM*$_1$ agent, for every order of theory of mind, an agent has additional beliefs $b^{(k)} : \mathcal{D} \rightarrow [0, 1]$, goal location beliefs $p^{(k)} : \mathcal{L} \rightarrow [0, 1]$, and a confidence $c_k \in [0, 1]$

in its kth-order theory of mind. Therefore, a *ToM*$_k$ agent has $k+1$ hypotheses about the future behavior of its trading partner. A *ToM*$_k$ agent continuously updates these beliefs according to which hypothesis fits the behavior of its trading partner best.

Recall that the beliefs of the trading partner may change when it receives an offer. Just like a *ToM*$_1$ agent, a *ToM*$_k$ agent, $k \geq 2$, uses adjusted beliefs to predict its trading partner's behavior by using the response function of a *ToM*$_{k-1}$ agent. Assuming that the trading partner's goal position is $l \in \mathcal{L}$, the expected value a *ToM*$_k$ agent assigns to making counteroffer $O \in \mathcal{D}$ is given by

$$EV_i^{(k)}(l, O) = \begin{cases} \pi_i(l_i, D_0) - 1 & \text{if } \hat{O}^{(k-1)} = \text{withdraw}, \\ \pi_i(l_i, O) - 1 & \text{if } \hat{O}^{(k-1)} = \text{accept}, \\ \max\left\{\pi_i\left(l_i, \hat{O}^{(k-1)}\right), \pi_i(l_i, D_0)\right\} - 2 & \text{otherwise}, \end{cases} \tag{3.9}$$

where

$$\hat{O}^{(k-1)} = \textit{ToM}_{(k-1)r}(O) \tag{3.10}$$

is the offer the *ToM*$_k$ agent expects its trading partner to make in response to offer O. Using the expected values of the lower orders of theory of mind recursively, the expected value a *ToM*$_k$ agent assigns to making counteroffer $O \in \mathcal{D}$ is given by

$$EV_i^{(k)}(O) = (1 - c_k) \cdot EV_i^{(k-1)}(O) + c_k \cdot \sum_{l \in \mathcal{L}} p^{(k)}(l) \cdot EV_i^{(k)}(l, O). \tag{3.11}$$

A *ToM*$_k$ agent randomly selects an offer that maximizes the expected value, that is,

$$O_t^* = \arg\max_{O \in \mathcal{D}} EV_i^{(k)}(O). \tag{3.12}$$

A *ToM*$_k$ agent will decide rationally between making an offer, accepting the previous offer, or withdrawing from negotiation by choosing the option that yields the highest (expected) score. A *ToM*$_k$ agent will thus select the option that is in accordance with its response function:

$$\mathit{ToM}_{ki}(O_{t-1}) = \begin{cases} O_t^* & \text{if} \quad EV_i^{(k)}(O_t^*) > \pi_i(l_i, D_0) \text{ and} \\ & \qquad EV_i^{(k)}(O_t^*) > \pi_i(l_i, O_{t-1}) \\ \text{accept} & \text{if} \quad \pi_i(l_i, O_{t-1}) > \pi(l_i, D_0) \text{ and} \\ & \qquad \pi_i(l_i, O_{t-1}) \geq EV_i^{(k)}(O_t^*) \\ \text{withdraw} & \text{otherwise.} \end{cases} \tag{3.13}$$

3.2.5 Learning within games

The *ToM*$_0$ agents discussed in Section 3.2.2 form beliefs about the likelihood that a certain offer will be accepted. An agent updates these beliefs when it receives an offer from its trading partner and when the trading partner rejected its offer. Whether or not an offer will be accepted depends on the game board, the distribution of chips, the goal locations of the agents, the history of offers made in this game, and the history of offers made in previous games. (For more about learning across games, see Section 3.2.6.) To generalize the behavior of *ToM*$_0$ agents, we discuss a simple learning heuristic in this subsection as provided by De Weerd et al. (2017).

Many of the belief updates in this subsection make use of an agent-specific learning speed $\lambda \in [0, 1]$. The learning speed is a fixed parameter that represents the degree to which new information influences the beliefs of the agent. A high learning speed means that the agent updates its beliefs fast and based on the latest information, and a low learning speed means that the agent needs a longer time to take the information into account. Agents do not try to model the learning speed of their trading partners. Hence, the beliefs of the agent about the trading partner are generally incorrect, unless the learning speed of the agents is the same.

When a *ToM*$_0$ agent receives offer O_{t-1}, the *ToM*$_0$ agent decreases its belief that its trading partner will accept an offer $O \in \mathcal{D}$ when offering $O \in \mathcal{D}$ assigns more chips of some color $c \in \mathcal{C}$ to the agent itself than offer O_{t-1} does. For a *ToM*$_0$ agent, this results in a belief update

$$U(b^{(0)}, O_{t-1})(O) = (1-\lambda)^m \cdot b^{(0)}(O), \tag{3.14}$$

where m is the number of colors for which offer $O \in \mathcal{D}$ assigns more chips to the agent itself than offer O_{t-1}. For example, if the trading partner offers two blue chips and two red chips to the agent, the agent considers it to be less likely that any offer that assigns more than two blue chips, more than two red chips, or at least one chip of another color to the agent itself will be accepted by the trading partner.

When the trading partner rejects offer O_t (by making a new offer), the agent decreases its belief that its trading partner will accept any offer $O \in \mathcal{D}$ that assigns at least as many chips of a given color $c \in \mathcal{C}$ to the agent itself as offer O_t does. For a *ToM*$_0$ agent, this results in a belief update as given by

$$U^R(b^{(0)}, O_t)(O) = (1 - \lambda)^{m'} \cdot b^{(0)}(O), \tag{3.15}$$

where m' is the number of colors for which offer O assigns at least as many chips of a given color $c \in \mathcal{C}$ to the agent as offer O_t does. For example, if the agent offers two blue chips and two red chips to the trading partner and the trading partner rejects the offer, the agent considers it to be less likely that any offer that assigns at least two blue chips, at least two red chips, or at least zero chips of another color to the agent itself will be accepted by the trading partner. In the case of five available chip colors in the game, this means that all offers are considered to be less likely to be accepted because any offer assigns at least zero chips of the other colored chips. However, the offers that assign at least two blue chips or at least two green chips will be considered even less likely, as can be verified with Equation (3.15).

Agents that are capable of theory of mind also have beliefs about the goal location of the trading partner. These beliefs are updated whenever an agent receives an offer from the trading partner. Agents assume that the trading partner is rational and that the trading partner only makes offers that increase the score of the trading partner. Hence, when a *ToM*$_k$ agent, $k \geq 1$, receives an offer O_{t-1}, the *ToM*$_k$ agent considers it impossible that the goal location of its trading partner is $l \in \mathcal{L}$ for which $\pi_r(l, O_{t-1}) + 1 \leq \pi_r(l, D_0)$, since for these goal locations the trading partner would be better off by withdrawing from negotiation. For the other possible goal locations $l \in \mathcal{L}$, the belief update is proportional to the expected increase in the score of the trading partner if offer O_{t-1} would be accepted. That is, after receiving offer O_{t-1} from its trading partner, a *ToM*$_k$ agent updates its goal location beliefs $p^{(k)}$ of goal location $l \in \mathcal{L}$ as follows:

$$p^{(k)}_{\texttt{new}}(l) := \begin{cases} 0 & \text{if } \pi_r(l, O_{t-1}) + 1 \leq \pi_r(l, D_0) \\ \beta \cdot p^{(k)}_{\texttt{old}}(l) \cdot \dfrac{1 + EV^{(k-1)}_{i \to r}(l, O_{t-1})}{1 + \max_{O \in \mathcal{D}} EV^{(k-1)}_{i \to r}(l, O)} & \text{otherwise,} \end{cases} \tag{3.16}$$

where $p^{(k)}_{\texttt{old}}$ are the old goal location beliefs, $p^{(k)}_{\texttt{new}}$ are the new goal location beliefs, and β is a normalizing constant to ensure $\sum_{l\in\mathcal{L}} p^{(k)}_{\texttt{new}}(l) = 1$. For the sake of clarity, we used the notation $EV^{(k-1)}_{i\to r}$ to indicate that the expected value is calculated using the $(k-1)$th-order beliefs that agent i, a *ToM*$_k$ agent, assigns to agent r, a *ToM*$_{k-1}$ agent.

Furthermore, agents that are capable of theory of mind update their confidence in their order of theory of mind. It might be that an agent's model of the trading partner is not accurate using its order of theory of mind. In that case, the agent can choose to place more confidence in a lower-order of theory of mind that better fits the behavior of its trading partner. Because its trading partner may also change its order of theory of mind, confidences are updated using adaptive expectations. After a *ToM*$_k$ agent, $k \geq 1$, receives offer O_{t-1} from its trading partner, the *ToM*$_k$ agent updates its confidence in using its kth-order theory of mind $c^{(k)}$ by using the update formula

$$c^{(k)}_{\texttt{new}} := (1-\lambda)\cdot c^{(k)}_{\texttt{old}} + \lambda \cdot \sum_{l\in\mathcal{L}} p^{(k)}_{\texttt{old}}(l) \cdot \frac{1 + EV^{(k-1)}_{i\to r}(l, O_{t-1})}{1 + \max_{O\in\mathcal{D}} EV^{(k-1)}_{i\to r}(l, O)}. \tag{3.17}$$

Herein, $c^{(k)}_{\texttt{old}}$ and $c^{(k)}_{\texttt{new}}$ are the old and new confidence in using its kth-order theory of mind, respectively. A *ToM*$_k$ agent assigns higher confidence to its kth-order theory of mind when the agent assigns a high expected value to the offer O_{t-1} made by the agent's trading partner compared to the offer that the agent would have selected itself if it had been a *ToM*$_{k-1}$ agent in the position of its trading partner.

At the start of a new round of negotiation, a *ToM*$_k$ agent, $k \geq 1$, resets its confidence in its kth-order theory of mind to 1. Moreover, note that a *ToM*$_k$ agent, $k \geq 1$, can model its trading partner as a *ToM*$_{k-1}$ agent. This means that for $k \geq 2$, a *ToM*$_k$ agent can model its trading partner being able to put higher confidence in its lower-order of theory of mind capability. If that were to happen, a *ToM*$_2$ agent could model its trading partner to be a *ToM*$_1$ agent but the *ToM*$_1$ trading partner using a zero-order theory of mind strategy. Since a *ToM*$_2$ agent can also model its trading partner to be a *ToM*$_0$ agent by putting more confidence in its own first-order beliefs, a *ToM*$_k$ agent models its trading partner with the trading partner's confidence fixed to 1. This means that the partner model of a *ToM*$_k$ agent is modeled as a *ToM*$_{k-1}$ agent with its confidence in its $(k-1)$th-order theory of mind fixed to 1. If the *ToM*$_k$ agent uses its $(k-1)$th-order theory of mind, the *ToM*$_k$ agent models its trading partner as a *ToM*$_{k-2}$ agent but the *ToM*$_{k-2}$ agent's confidence is fixed, etc.

A visual representation of how agents capable of theory of mind model their trading partner and use their theory of mind of lower orders is given in Figure 3.2. A *ToM*$_k$ agent,

$k \geq 1$, has a model of itself with a lower-order of theory of mind and a model of its trading partner with a lower-order of theory of mind because an agent capable of kth-order theory of mind can model its trading partner as a *ToM*$_{k-1}$ agent. Since the model of the trading partner is confidence-locked, that is, all its confidence is assigned to its highest order of theory of mind capability, the model of the trading partner does not have a model of itself where it can use a lower-order of theory of mind.

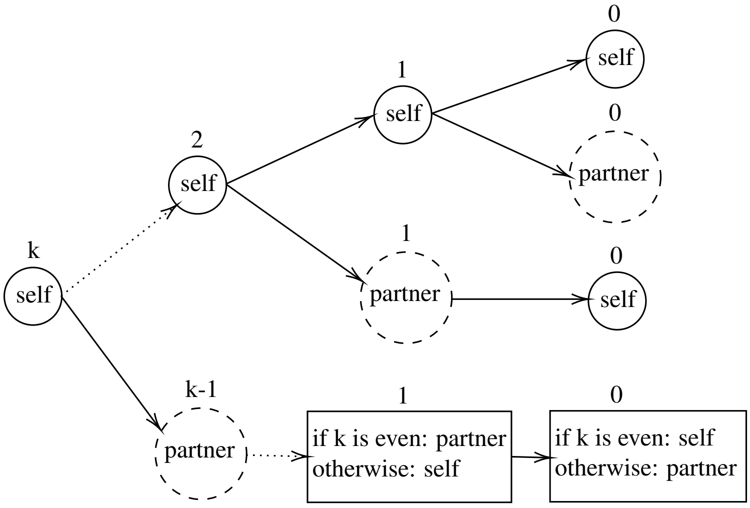


Figure 3.2: Model structure of a *ToM*$_k$ agent. A *ToM*$_k$ agent, $k \geq 1$, models its trading partner as a *ToM*$_{k-1}$ agent. A partner model is always confidence-locked, which means that a modeled partner does not have a (direct) model of itself with a lower-order theory of mind capability. An agent that is not confidence-locked also has a model of itself where it can use its lower-order theory of mind capability. Note that the kth-order beliefs of an agent with at least a kth-order theory of mind capability of whether an offer is going to be accepted are the modeled $(k-1)$th-order beliefs of the trading partner (where $k \geq 1$).

3.2.6 Learning across games

An offer such as three blue chips and one red chip might be much better in one Colored Trails game setting than in another setting. The number of possible game boards in our version of Colored Trails is 5^{25} as there are 25 colored tiles and each tile can have 5 possible colors.[2] Moreover, we have two players each with initially four chips that can be colored in five different ways, so there are $\left(\binom{(5+4-1)}{4}\right)^2 = 4900$ possible distributions of an initial set of chips, where we take the power of 2 to account for the initiator and the

[2] Since agents have the same starting position, the starting position will never be moved on (again) when finding the shortest path to the goal position. Hence, agents will never have to hand in a chip for moving to the starting position, and so the color of the starting position is irrelevant to the game. As a consequence, it could be excluded from the total possible game boards, but for completeness, we left it here.

responder.[3] In addition, we have twelve possible goal locations for each player, resulting in 12^2 possible distributions of goal locations. Taken together, we have over $2 \cdot 10^{23}$ different initial game settings, so it is infeasible for agents to form beliefs for each game setting within our Colored Trails game.

For *ToM*$_0$ agents to generalize across the over $2 \cdot 10^{23}$ different initial game settings, De Weerd et al. (2017) propose a simple learning heuristic that allows *ToM*$_0$ agents to make mutually beneficial offers after a short learning period. In this subsection, we discuss how *ToM*$_0$ agents learn across different Colored Trails games.

To distinguish between offers in different games, the color of the chips is not taken into account. Instead, offers are classified by the number of chips that are transferred from the agent to its trading partner, and the number of chips that are transferred from the trading partner to the agent. Since agents initially possess four colored chips, an agent can offer zero to four chips to the trading partner and vice versa. Hence, there are a total of 25 classes of offers that agents distinguish. For example, agents distinguish between an offer that trades one (blue) chip for one (red) chip and an offer that trades two (blue) chips for two (red) chips, but agents do not distinguish between an offer that trades one blue chip with one red chip and an offer that trades one green chip for one red chip.

During negotiation, for each of the 25 classes of offers, agents keep track of the total number of offers (received and made) and the total number of offers that are accepted by the trading partner. Here, agents consider offers made by the trading partner also as offers that are accepted by the trading partner (as a rational trading partner only makes offers that the agent itself would accept). At the instantiation of a new negotiation round, an agent resets the belief of an offer going to be accepted equal to the observed frequency of similar types of offers that have been accepted. To provide an agent with beliefs before any offer has been observed and to circumvent complete disbelief of an offer going to be accepted, the agent assumes to have had five positive encounters with every offer type.

3.3 Lying in Colored Trails

In Section 3.2, we discussed how agents can have a theory of mind capability, which is adapted from De Weerd et al. (2017). In this section, we will deviate from De Weerd et al. (2017) and other existing literature by introducing agents that are capable of lying in

[3] To calculate the number of combinations with repetition, the formula $\binom{n+k-1}{k}$ can be used, where n is the number of different elements and k is the sample size.

Colored Trails. In this thesis, lying is defined as an agent making a false statement with the intent to deceive. More specifically, we define lying in Colored Trails as follows:

Definition 3 (Lying in Colored Trails)**.** An agent is said to be lying if and only if the agent makes a statement p that the agent believes to be false with the intent that its trading partner believes that p is true and that its trading partner believes that the lying agent also believes p is true.

Regarding this definition, one might argue that the agents discussed in Section 3.2 do not have intentions,[4] since these agents make decisions based on an expected value and choose the action that leads to the highest expected value. In this definition and the subsequent part of our thesis, we adopt the intentional stance (Dennett, 1989) toward our agents. This means that we use the strategy of ascribing intentions to our agents in predicting and explaining their behavior.

Definition 3 of lying in Colored Trails is adapted from the definition of lying as given by Van Ditmarsch et al. (2020), which is also stated in the introduction of this thesis (recall Definition 1). Following Definition 1, common belief is required for lying; however, if we were to require common belief in our setting, a lying agent would need an infinite theory of mind capability. Hence, we restrict the definition of lying in Colored Trails such that a lying agent needs to be capable of second-order theory of mind to capture the intent to change the beliefs that the receiver has about the lying agent itself.

Agents that are not capable of second-order theory of mind, i.e., ToM_0 and ToM_1 agents, are not considered agents capable of lying. When a ToM_0 agent makes a statement that the agent believes to be false, it does so without the intent to change the beliefs of the trading partner, since it does not model its trading partner having beliefs. Hence, we say that a ToM_0 agent simply utters a false statement. However, when a ToM_1 agent makes a false statement, it might intend to change the beliefs of the trading partner without intending to change the beliefs the trading partner attributes to the ToM_1 agent itself, since a ToM_1 agent does not model its trading partner attributing beliefs to the ToM_1 agent itself. A ToM_1 agent is, therefore, not able to lie according to Definition 3; however, we say that a ToM_1 agent can mislead:

[4] Or one might argue whether a computer agent can have intentions at all. Like mentioned in the introduction of Chapter 2, there is still an ongoing (philosophical) debate about whether AI systems can hold beliefs and intentions (Livet & Varenne, 2020).

Definition 4 (Misleading in Colored Trails)**.** An agent is said to be misleading if and only if the agent makes a statement p that the agent believes to be false with the intent to change the beliefs of the trading partner.

Note that misleading is a more general term than lying, that is, an agent capable of lying is also capable of misleading, while an agent capable of misleading is not necessarily capable of lying. The difference lies mainly in that an agent misleading with statement p does not have to model the beliefs of the trading partner about statement p, but with stating statement p, the misleading agent intends to change some beliefs (which can be about statement p but do not have to).

Recall that an offer is simply a distribution of colored chips that an agent offers to the trading partner. Agents capable of theory of mind can use offers to change the beliefs of the trading partner. Offers can therefore be used to deceive the trading partner by making an offer with the intent to cause the trading partner to hold a false belief about something that the deceiving agent knows to be false (recall Definition 2 of deceiving). A *ToM*$_2$ agent who makes an offer that results in its trading partner deducing false information about the agent's goal location is an example of deceptive behavior.

Example 1 (Deceptive behavior in the original Colored Trails setting (De Weerd et al., 2017))**.** In Figure 3.3, a *ToM*$_2$ agent deceives the trading partner by making an offer that results in the trading partner deducing false information about the *ToM*$_2$ agent's goal location. The initiator is the *ToM*$_2$ agent and the responder is a *ToM*$_1$ agent. In this example, both agents can only communicate through offers, so we consider the setting as described in Section 3.2.

In this example, the goal location of agent i is goal location 1. Agent i can move two steps toward its goal location after which it has two chips left; so, with its initial set of chips, agent i can reach a total of 300 points.

Agent i decides to offer two purple and two light blue chips to agent r; so, it offers to exchange a purple chip for a dark orange chip. Agent i models with its second-order theory of mind capability that this offer results in agent r only considering goal locations 2 and 8 as possible goal locations for agent i, assuming agent r is a *ToM*$_1$ agent. Agent i places itself in the position of its trading partner (agent r) and calculates the score change for agent i if agent r would accept the offer, for each possible goal location. The higher the increase in score for agent i, the more likely agent r finds that location to be the actual goal location of agent i.

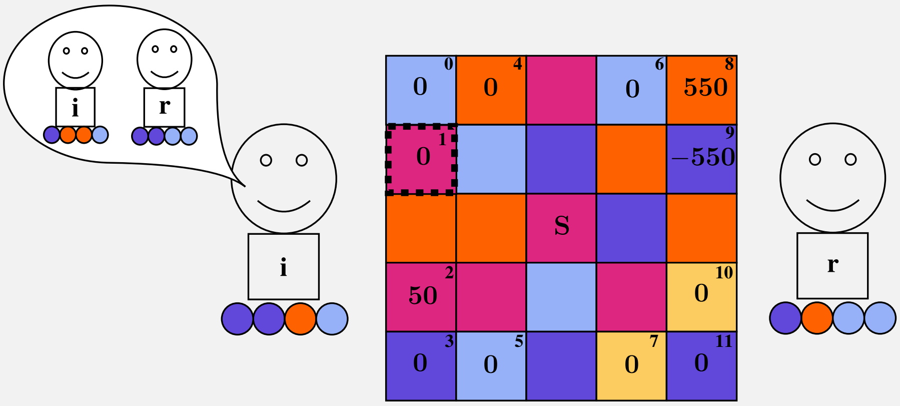


Figure 3.3: Example of deceptive behavior by a *ToM*$_2$ agent (the initiator), agent i, negotiating with a *ToM*$_1$ agent (the responder), agent r. Agent i offers to exchange a purple chip for a dark orange chip. The (larger) centered numbers in the tiles indicate the change in score for agent i if agent r were to accept the offer of agent i and agent i were to have that goal location as its actual goal location. The (smaller) numbers in the top right corner of a tile indicate the goal location numbers. The tile highlighted with the dashed border is the actual goal location of agent i, which is unknown to agent r. The goal location of agent r is not relevant for this example and is therefore not indicated. The starting position of both agents is denoted with S.

In this example, agent i chooses to offer a distribution of chips that does not increase its score; however, agent i calculated that the expected value of this offer was the highest among all distributions of chips it could offer. Agent i might expect that agent r makes a counteroffer in response to this offer that it would accept.

As a result of receiving the offer from agent i, agent r places zero probability mass on goal location 1, the actual goal location of agent i, and only considers goal locations 2 and 8 as possible goal locations for agent i. In this example, agent i, a *ToM*$_2$ agent, deceives agent r as agent i intentionally causes agent r to have false goal location beliefs (recall Definition 2 of deceiving).

While agents capable of second-order theory of mind as described in Section 3.2 are able to deceive, these agents are not able to lie. According to Definition 3, agents cannot possibly lie to the trading partner by simply making offers since an offer is a distribution of chips an agent offers to the trading partner and thereby a commitment: If the trading partner decides to accept the offer, the offer becomes final. Hence, in the original game of Colored Trails, an agent cannot make a statement p that the agent believes to be false. In Sections 3.3.1–3.3.3, we extend the traditional Colored Trails game by introducing agents that will be able to communicate with each other by making statements besides making offers. This enables agents capable of second-order theory of mind to lie to their trading partners.

More specifically, we explain how agents can use the additional communication option in Section 3.3.1. Section 3.3.2 follows with explaining when agents decide on sending a goal location message. In Section 3.3.3, we introduce agents that are capable of lying.

3.3.1 Communication

We first introduce an additional communication option for an agent to be able to make false statements in Colored Trails. One may think of multiple ways an agent can *usefully* communicate some information about the Colored Trails game to the trading partner. Examples of useful communication by an agent in the Colored Trails setting are communicating a goal position (as an agent's goal position is unknown to the trading partner), a preference order of offers, or the number of chips needed to reach its goal location. These three examples are useful communication in Colored Trails in the sense that communicating a goal position, a preference order of offers, or the number of chips needed to reach its goal location gives (additional) information about the game to the trading partner.

In our specific case of Colored Trails with five different colors and each agent having initially four colored chips, there are 9 to 108 offers an agent can make, depending on the initial division of colored chips. In the simple case where all chips have the same color, an agent can make nine different offers. Consequently, an agent can construct $(9 \times 8 =)\,72$ different preference orders of two different offers. However, in cases where the colors are evenly distributed over the chips, the agent can make $(108 \times 107 =)\,11{,}556$ different preference orders of two different offers. When an agent has to decide which of these preference orders to communicate, it must reason about the effect of communicating each of these different preference orders. This means that a *ToM*$_0$ agent has to store and compute the effect of every different preference order of two different offers.

On the other hand, in our Colored Trails setting of a five-by-five board, goal locations are the locations that are at least three or four steps away from the starting location. Communicating the number of chips needed to reach its goal location is limited to three and four chips. Hence, there would be only two possibilities to communicate.

Since communicating a preference order would be highly computationally expensive but communicating the number of chips needed to reach its goal location may be too simple, agents in this thesis can communicate a goal location to the trading partner by a *goal location message*, that is, a message communicating to the receiver that the goal location in the message is the sender's goal location. There are twelve possible goal locations in our Colored Trails game, so reasoning about which goal location to communicate is less computationally expensive than reasoning about which preference order of offers to communicate.

Before sending an offer, an agent has the opportunity to communicate a goal location to the trading partner. Agents can also decide not to send a goal location message before an offer. In that case, an agent only makes an offer and the trading partner takes its turn. Besides making an offer, an agent can decide to withdraw from negotiation or accept the previous offer as mentioned in Section 3.2. In the case that an agent decides to accept the previous offer or withdraw from negotiation, a goal location message cannot be sent since the distribution of chips is already final.

Using the additional communication method of agents being able to send goal location messages, we can construct five new types of agents (in addition to the agents capable of theory of mind in De Weerd et al. (2017)):

- a *ToM*$_0$ agent that can send goal location messages, including goal location messages containing a goal location that is not its actual goal location;
- an honest *ToM*$_1$ agent that can send goal location messages but is not capable of sending goal location messages containing a goal location that is not its actual goal location;
- a misleading *ToM*$_1$ agent that can send goal location messages and can send goal location messages containing a goal location that is not its actual goal location;
- an honest *ToM*$_2$ agent that can send goal location messages but is not capable of sending goal location messages containing a goal location that is not its actual goal location;
- a lying *ToM*$_2$ agent that can send goal location messages and can send goal location messages containing a goal location that is not its actual goal location.

Note that by adopting the intentional stance of Dennett (1989), we may ascribe the intention of agents that are not capable of sending goal location messages containing a goal location that is not its actual goal location, to be honest. We explain the capabilities of the different agents in the subsequent part of Section 3.3.

3.3.1.1 Receiving a goal location message

The result of receiving a goal location message depends on the theory of mind capability of the agent. When a *ToM*$_0$ agent receives a goal location message from the trading partner, the *ToM*$_0$ agent cannot understand the meaning of the message like agents with a first-order or second-order theory of mind capability do, since a *ToM*$_0$ agent cannot model a goal of the trading partner. While a *ToM*$_0$ agent cannot construct beliefs about the trading partner having a goal, it constructs beliefs about the probability of an offer going to be accepted.

When a *ToM*$_0$ agent receives a goal location message, it decreases its belief that its trading partner will accept an offer $O \in \mathcal{D}$ that does not contain a chip of the color corresponding to the goal location indicated by the goal location message. If it is indeed the case that the sender's goal location is the goal location mentioned, the sender needs at least one chip of that color to reach its goal location. When a *ToM*$_0$ agent receives a goal location message containing goal location $l \in \mathcal{L}$, the belief update is given by

$$U^{RG}(b^{(k)}, l)(O) = (1-\lambda)^{\left(1-\mathbb{1}_{\{\texttt{contains\_color}(O,c(l))\}}\right)} \cdot b^{(k)}(O). \tag{3.18}$$

Herein, $\mathbb{1}_{\{x\}}$ is the indicator function that evaluates to 1 if x is true, and 0 otherwise; $c : \mathcal{L} \to \mathcal{C}$ a function that returns the tile color of the goal location; and $\texttt{contains\_color} : \mathcal{D} \times \mathcal{C} \to \{\texttt{true}, \texttt{false}\}$ a function that returns $\texttt{true}$ if and only if offer $O \in \mathcal{D}$ contains a chip with input color $c \in \mathcal{C}$.

While *ToM*$_0$ agents do not model the trading partner having a goal location, *ToM*$_k$ agents with $k \geq 1$ do; so *ToM*$_k$ agents, $k \geq 1$, will be able to infer a goal location of the trading partner from receiving a goal location message and adjust their goal location beliefs. When a *ToM*$_k$ agent, $k \geq 1$, receives a goal location message, it first checks whether it already received a goal location message in the current negotiation round and acts accordingly.

If the *ToM*$_k$ agent did not receive a goal location message from its trading partner before, it agent checks if its goal location belief of the communicated goal location is greater than zero (i.e., a nonzero probability that the trading partner has indeed the goal location as mentioned in the goal location message). If it is the case that the *ToM*$_k$ agent's goal location belief of the communicated goal location is greater than zero, then it believes the trading partner and sets its goal location belief to 1 for the communicated goal location and zero for all other goal locations. If the *ToM*$_k$ agent's goal location belief of the communicated goal location is equal to zero, it does not believe the trading partner, and it will not change its goal location beliefs. Note that this means that an agent capable of theory of mind believes that the goal location in the goal location message is the sender's actual goal location even when the agent considers there to be a really small probability that the mentioned goal location is the sender's actual goal location. However, since they are only beliefs of the agent, the *ToM*$_k$ agent still considers it possible that the goal location mentioned in the goal location message is not the actual goal location of the trading partner.

When a *ToM*$_k$ agent, $k \geq 1$, receives its first goal location message in the current negotiation round and believes the trading partner, the *ToM*$_k$ agent stores its goal location beliefs before updating them. A *ToM*$_k$ agent, $k \geq 1$, stores its goal location beliefs because the *ToM*$_k$ agent considers a possibility that the goal location mentioned in the goal location message is not the actual goal location of the trading partner. A *ToM*$_k$ agent, $k \geq 1$, may disbelieve its trading partner when an additional goal location message is sent containing a different goal location, contradicting the previous goal location message, or when an offer of the trading partner is not rational assuming the goal location of the trading partner is indeed the goal location mentioned by the trading partner. In both cases, after the updates have taken place, all goal location beliefs would be zero, resulting in the *ToM*$_k$ agent believing that no goal location is the actual goal location of the trading partner, which is not possible since each agent has a goal location. Hence, in such cases, a *ToM*$_k$ agent, $k \geq 1$, restores its saved goal location beliefs (its goal location beliefs without the influence of a goal location message) and uses these saved goal location beliefs as its new goal location beliefs.

If a *ToM*$_k$ agent, $k \geq 1$, were to receive a new goal location message, but the *ToM*$_k$ agent already revoked its goal location beliefs, the agent will not believe the trading partner anymore. Consequently, the *ToM*$_k$ agent will not update its goal location beliefs in response to the goal location message and acts as if the new message had not been sent.

As explained in Section 3.2.5, a *ToM*$_k$ agent, $k \geq 1$, updates its goal location beliefs after an offer is made by the trading partner. The *ToM*$_k$ agent will still update its goal location beliefs according to Equation 3.16 due to offers that the trading partner makes even when it received a message in the current negotiation round. Moreover, the saved goal location beliefs are also updated as if the saved goal location beliefs of the *ToM*$_k$ agent were the actual goal location beliefs. Finally, the confidence of the *ToM*$_k$ agent in its kth-order theory of mind, is updated using Equation (3.17) using its current goal location beliefs.

Example 2 (Receiving a goal location message)**.** In Figure 3.4, agent i, a *ToM*$_2$ agent (the initiator), sends a goal location message with its actual goal location, goal location 1, to agent r, a *ToM*$_1$ agent (the responder). Since agent r did not receive a goal location message in this negotiation round yet, agent r checks whether its goal location belief of goal location 1 is nonzero. Since there had not been sent an offer yet, its goal location belief was 0.08 (because there are 12 possible goal locations and each goal location is equally likely). Hence, agent r sets its goal location belief to 1 for goal location 1 and zero for all other goal locations. Agent r stores its goal location beliefs (all with 0.08) because agent r considers the possibility that goal location 1 is not the actual goal location of agent i in which case it will use these saved goal location beliefs again.

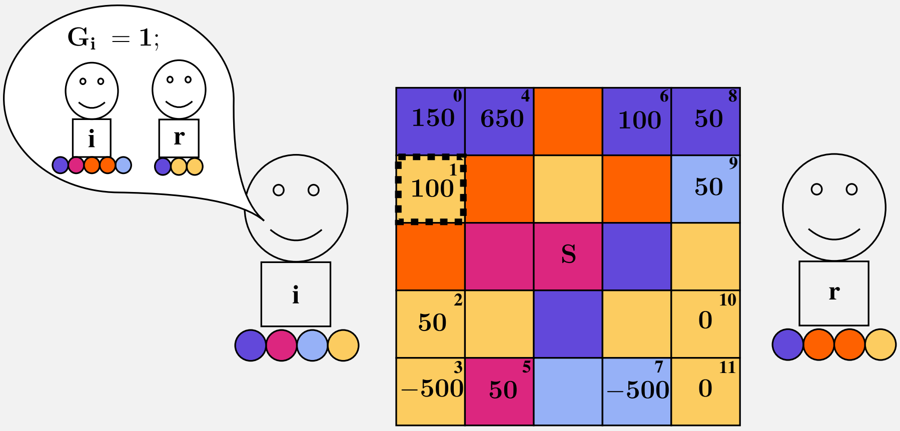


Figure 3.4: Example of an agent receiving a goal location message. The *ToM*$_2$ agent (the initiator), agent i, negotiates with a *ToM*$_1$ agent (the responder), agent r. The initiator sends a goal location message with its actual goal location, goal location 1. Moreover, the initiator offers to exchange one yellow chip for two orange chips. The (larger) centered numbers in the tiles indicate the change in score for the initiator if the responder were to accept the offer of the initiator and the initiator were to have that goal location as its actual goal location. The (smaller) numbers in the top right corner of a tile indicate the goal location numbers. The tile highlighted with the dashed border is the actual goal location of agent i, which is unknown to agent r. The goal location of agent r is not relevant for this example and is therefore not indicated. The starting position of both agents is denoted with S.

Together with the goal location message, an offer is sent by agent i. Agent i offers one yellow chip against two orange chips. After receiving the offer from agent i, agent r updates its current goal location beliefs, which remain the same, since the offer is rational when assuming that goal location 1 is the goal location of agent i, that is, agent r remains to believe with probability 1 that the goal location of the trading partner is goal location 1. In case it turns out that there occurs a contradiction in the beliefs of agent r later in the negotiation, that is, all goal location beliefs become 0, agent r will revoke its goal location beliefs and use its saved goal location beliefs. During negotiation, these saved goal location beliefs are also updated.

3.3.1.2 Sending a goal location message

Whether an agent sends a goal location message depends partly on its theory of mind capability. While zero-order and first-order theory of mind agents will be able to send goal location messages, only an agent that is capable of second-order theory of mind will be able to model the effects of sending messages that reveal a goal location to its trading partner, since agents that are capable of second-order theory of mind can model the trading partner having beliefs about the focal agent's goal location.

A *ToM*$_0$ agent does not model beliefs or a goal location of the trading partner but constructs beliefs only about the probability of an offer going to be accepted. Hence, a *ToM*$_0$ agent does not consider a change in the beliefs of the trading partner resulting from it sending a goal location message. Instead, a *ToM*$_0$ agent sends a goal location message according to some small probability (see Section 3.3.2). When the *ToM*$_0$ agent sends a goal location message, its actual goal location has the highest probability of being chosen. This kind of behavior fits with the interpretation of signals by *ToM*$_0$ agents and is inspired by the caching behavior of ravens (Van der Wall, 1990). When caching their food on a specific site, ravens look out for the presence of potential raiders, and as a consequence, frequently interrupt caching, change cache sites, or recover their food items from a cache site (Bugnyar & Kotrschal, 2002).

While *ToM*$_0$ agents cannot model beliefs and a goal location of their trading partner, agents capable of theory of mind can. A *ToM*$_k$ agent, $k \geq 1$, models its trading partner having beliefs. Since agents change their beliefs as a result of receiving a goal location message, a *ToM*$_k$ agent, $k \geq 1$, will be able to reason what the effect of sending a goal location message is. Whether a *ToM*$_k$ agent, $k \geq 1$, actually sends a goal location message, is discussed in the next section, Section 3.3.2.

Now that we discussed the changes in beliefs that occur when an agent sends a goal location message and receives a goal location message, we can summarize the changes in beliefs due to sending or receiving a goal location message as follows:

- When a *ToM*$_0$ agent receives or sends a goal location message, it changes its zero-order beliefs, that is, the beliefs of the probability of an offer going to be accepted.

- When a *ToM*$_1$ agent receives a goal location message, it may change its goal location beliefs (depending on whether it believes the trading partner). The *ToM*$_1$ agent also changes its first-order beliefs, since they are the zero-order beliefs of the trading partner who sent the goal location message, and a *ToM*$_0$ agent sending a goal location message changes its zero-order beliefs. Moreover, when a *ToM*$_1$ agent sends a message, it changes its first-order beliefs, but it does not change its goal location beliefs.

- When a *ToM*$_2$ agent receives a goal location message, it changes its goal location beliefs. Moreover, the *ToM*$_2$ agent changes its second-order beliefs, as they are the modeled zero-order beliefs of itself what the trading partner models as its first-order beliefs, and thus what the *ToM*$_2$ agent's second-order beliefs are. When a *ToM*$_2$ agent sends a goal location message, it does not change its goal location beliefs, but it does change its second-order beliefs.

Moreover, note that a *ToM*$_k$ agent, $k \geq 1$, has a model of itself where it is a *ToM*$_{k-1}$ agent with $(k-1)$th-order beliefs. Hence, when a *ToM*$_k$ agent, $k \geq 1$, receives or sends a message and its beliefs change, its $(k-1)$th-order beliefs also change.

3.3.2 When to send a goal location message?

Recall from Section 3.2 that an agent can make an offer, withdraw from negotiation, or accept the previous offer of the trading partner. In Section 3.3.1, we introduced how an agent can communicate a goal location message before making an offer. Now, we discuss when a *ToM*$_k$ agent will decide to communicate a goal location message together with an offer.

Before an agent decides on its action, the agent determines the expected value of all possibilities and chooses the combination that results in the highest expected value. An agent first calculates the best offer without sending a goal location message according to the model as described in Section 3.2. After this, the procedure differs depending on the theory of mind capability of the agent.

When a *ToM*$_0$ agent decides to make offer $O \in \mathcal{D}$, i.e., the expected value of making offer $O \in \mathcal{D}$ is maximal, there is a probability $p_0 \in [0, 1]$ that the *ToM*$_0$ agent communicates a goal location message together with making offer $O \in \mathcal{D}$. When a *ToM*$_0$ agent sends a goal location message, the *ToM*$_0$ agent chooses each goal location that is not its actual goal location with probability $p_1 \in [0, (|\mathcal{L}| - 1)^{-1}]$, where $|\mathcal{L}|$ is the number of possible goal locations. The upper limit on p_1 is set such that the sum of the probabilities of sending a false goal location, i.e., $(|\mathcal{L}| - 1) \cdot p_1$, does not exceed 1. Here, we assume that there are at least two goal locations such that there is a goal location that is not the actual goal location of an agent. (Note that, in our Colored Trails setting, we have $|\mathcal{L}| = 12$.) The probability that a *ToM*$_0$ agent sends its actual goal location follows from p_1. More specifically, when a *ToM*$_0$ agent sends a goal location message, it chooses its actual goal location with probability $1 - (|\mathcal{L}| - 1) \cdot p_1$. Probability p_1 will be chosen such that the probability of sending its actual goal location is the highest, that is, p_1 is chosen such that $1 - (|\mathcal{L}| - 1) \cdot p_1 > p_1$, which boils down to $p_1 < (|\mathcal{L}|)^{-1}$.

As discussed in Section 3.2, a *ToM*$_k$ agent, $k \geq 1$, models the expected response of its trading partner when deciding on an offer. When sending a goal location message with an offer, the *ToM*$_k$ agent models what it would do if it were in the position of its trading partner and received a goal location message with location $l \in \mathcal{L}$ together with offer $O \in \mathcal{D}$. The *ToM*$_k$ agent models the change in beliefs and models what the response of the trading partner will be as discussed in Section 3.2 and calculates the expected value of sending the goal location message in combination with the offer. Agents do not model a response of the trading partner where it sends a goal location message.

If a combination of a goal location message together with an offer yields a strictly higher expected value than any offer without a goal location message, the *ToM*$_k$ agent chooses the best combination of a goal location message with an offer. Otherwise, the *ToM*$_k$ agent chooses the best offer without any goal location message similar to De Weerd et al. (2017). Note that an agent only sends an offer (with goal location message) if the expected value of sending that offer (with goal location message) is higher than the value of withdrawing from negotiation and higher than the value of accepting the previous offer of the trading partner, as discussed in Section 3.2.

Example 3 (Sending a goal location message)**.** Consider the same setting as in Figure 3.4 from Example 2 where agent i, a *ToM*$_2$ agent (the initiator), sends a goal location message with its actual goal location, goal location 1, to agent r, a *ToM*$_1$ agent (the responder). Agent i is being honest about its actual goal location. This message is the first goal location message sent as well as the first offer made in this negotiation round.

Agent i chooses to send a goal location message together with an offer since it expects to obtain the highest score from this combination. This means that sending an offer without any goal location message yields a lower expected value than this offer with a goal location message. An agent would not send a goal location message if there was an offer without a goal location message that has the same or a higher expected score.

3.3.3 Lying and misleading agents

Following Definition 3 of lying in Colored Trails, an agent (the sender) can lie by communicating a goal location $l \in \mathcal{L}$ to the trading partner that is different from the sender's actual goal location $l_i \in \mathcal{L}$, that is, $l \neq l_i$, with the intent that the trading partner believes that the sender's goal location is goal location $l \in \mathcal{L}$ and that the trading partner believes that the sender believes the sender's goal location is goal location $l \in \mathcal{L}$. Since a *ToM*$_2$ agent models the trading partner being a *ToM*$_1$ agent, and it thus models that the trading partner attributes beliefs and a goal location to the *ToM*$_2$ agent, a *ToM*$_2$ agent can lie. We have that *ToM*$_0$ and *ToM*$_1$ agents cannot lie, because *ToM*$_0$ agents do not model the trading partner having beliefs at all, and *ToM*$_1$ agents do not model the trading partner having beliefs about the agent itself.

Example 4 (Lying agent)**.** In Figure 3.5, agent i, a *ToM*$_2$ agent (the initiator), sends a goal location message with a false goal location, goal location 9, to agent r, a *ToM*$_1$ agent (the responder). The actual goal location of agent i is goal location 0. With its initial set of chips, agent i can take one step toward its goal location with the fuchsia-colored chip after which agent i has three colored chips left, giving it a total of 250 initial points. In contrast, agent r has goal location 8 and cannot take one step toward its goal location, giving it a total of 200 initial points.

Agent i calculates that sending a goal location message with goal location 9 to agent r together with an offer that assigns the single fuchsia-colored chip to itself results in the highest expected value. This goal location message is a lie as the goal location of agent i is not goal location 9, but goal location 0, and agent i intends (or expects) to change the beliefs of agent r by sending this goal location message. Note that the offer is consistent with the message as the score of agent i would have increased if the offer were accepted and agent i's actual goal location would be goal location 9.

Since agent r needs a fuchsia-colored chip to move one step in the direction of its goal location, most of the offers that agent r benefit assign the fuchsia-colored chip to itself (agent r). While agent i does not know the goal location of agent r, agent i expects an offer in return that assigns one purple, three orange, and one light blue chip to itself, giving agent i a total of 950 points. Whether agent r responds with the expected offer depends on agent r's actual goal location and whether agent i models agent r correctly.

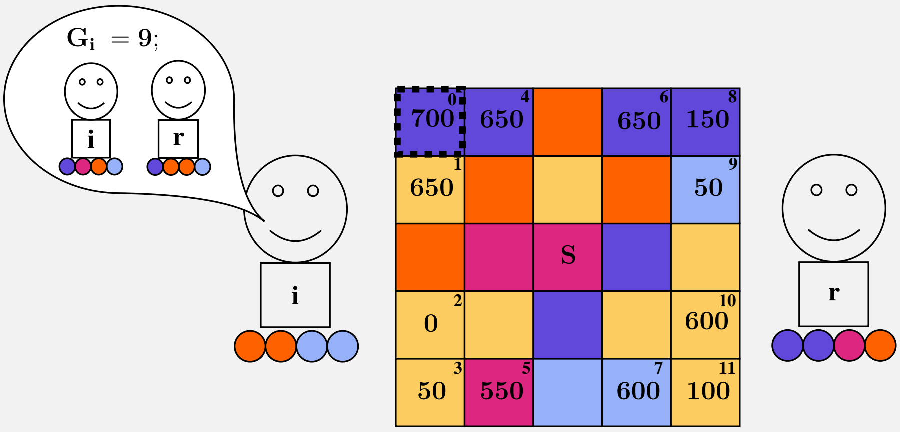


Figure 3.5: Example of an agent lying. The *ToM*$_2$ agent (the initiator), agent i, negotiates with a *ToM*$_1$ agent (the responder), agent r. Agent i sends a goal location message with a false goal location, goal location 9. Moreover, the agent i offers to exchange one orange and one light blue chip for one purple and one fuchsia-colored chip. The (larger) centered numbers in the tiles indicate the change in score for agent i if agent r were to accept the offer of agent i and agent i were to have that goal location as its actual goal location. The (smaller) numbers in the top right corner of a tile indicate the goal location numbers. The tile highlighted with the dashed border is the actual goal location of agent i, which is known to agent r. The goal location of agent r is not relevant for this example and is therefore not indicated. The starting position of both agents is denoted with S.

While *ToM*$_1$ agents are not able to lie, recall that *ToM*$_1$ agents can mislead (recall Definition 4 of misleading in Colored Trails). A *ToM*$_1$ agent models the trading partner having beliefs but not about the *ToM*$_1$ agent itself. However, when a *ToM*$_1$ agent sends a goal location message, it models its trading partner, a *ToM*$_0$ agent, to receive its message and models its trading partner to update its beliefs. As discussed in Section 3.3.1.1, when a *ToM*$_0$ agent receives a goal location message, the *ToM*$_0$ agent considers offers that do not contain the color corresponding to the goal location as less likely to be accepted. Hence, a *ToM*$_1$ agent can still attempt to manipulate the beliefs of a *ToM*$_0$ agent. Hence, *ToM*$_1$ agents can mislead about the color they want by sending a false goal location message. A *ToM*$_0$ agent can make false statements although that is not considered lying or misleading, but rather a behavioral aspect of a *ToM*$_0$ agent.

3.3.3.1 Exceptions and conventions on lying and misleading

An agent uttering a statement p that it believes to be false without intending to deceive the trading partner on statement p is not a lying agent according to our definition, but rather an agent that utters a statement p that it believes to be false. There might occur a special case in the Colored Trails game where a *ToM*$_2$ agent sends a false goal location message, but the intent of the *ToM*$_2$ agent was not to let the trading partner believe the false statement, i.e., the sender expects that the receiver will not believe the message. This might happen, for example, when the sender intends to cause the receiver to revoke its current goal location beliefs. In that case, the receiver will change its goal location beliefs back to how they were before a goal location message had been sent (i.e., using the saved goal location beliefs). In the case where an agent sends a goal location message that is a false statement, but the intent is not to let the trading partner believe that the sender's actual goal location is the goal location mentioned, the false goal location message is not considered a lie according to our definition of lying. Since there might still be an intent to change the beliefs of the trading partner, the agent can be considered to have misled the trading partner.

Example 5 (Revoking goal location beliefs)**.** Consider the setting as in Figure 3.6, where a *ToM*$_2$ agent, agent i (the initiator), negotiates with a *ToM*$_1$ agent, agent r (the responder). In this negotiation round, agent i already sent a goal location message with goal location 10, its actual goal location, which agent r also believes. Then, agent r sent a counteroffer. It is now agent i's turn.

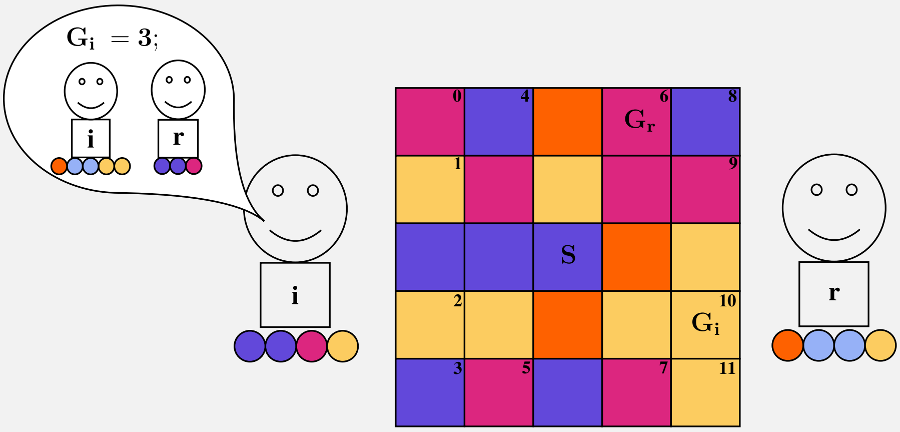


Figure 3.6: Example of an agent that sends a goal location message to revoke the beliefs of the trading partner agent. The *ToM*$_2$ agent (the initiator), agent i, negotiates with a *ToM*$_1$ agent (the responder), agent r. The initiator sends a goal location message with a false goal location, goal location 3, to revoke the beliefs of the responder. Moreover, the initiator offers to exchange two purple chips and one fuchsia-colored chip for an orange chip, two light-blue chips, and a yellow chip. The numbers in the top right corner of a tile indicate the goal location numbers. The starting position of both agents is denoted with S, and the goal positions of agents i and r are denoted with G_i and G_r, respectively.

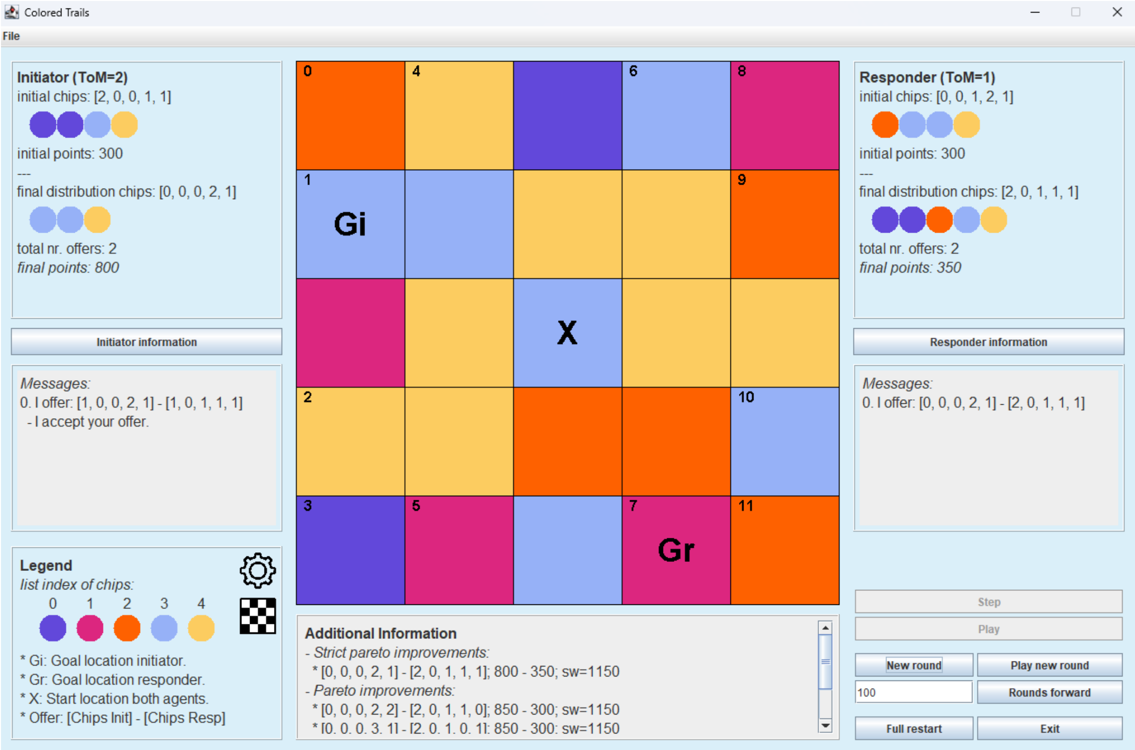
Since it is not rational to make an offer that does not increase the value of that agent, an offer provides information about the goal location of the trading partner. After the offer of agent r, agent i updates its goal location beliefs. It calculates all possible combinations of offers and location messages and determines that sending a goal location that is not the same as the previous goal location message, obtains the highest expected value. In this case, agent i intends to revoke the goal location beliefs of agent r (mislead agent r) such that agent r uses its saved goal location beliefs (without the influence of a goal location message) and such that agent r does not put all probability mass on agent i's previously announced goal location. The reason agent i might make such a decision is based on what agent i expects agent r to respond to the offer and goal location message.

Recall that *ToM*$_k$ agents, $k \geq 1$, have a confidence variable that indicates the confidence in using its kth-order theory of mind. Thus, a *ToM*$_k$ agent, $k \geq 1$, also models itself as an agent with a lower-order of theory of mind and weighs its predictions about the utility of an offer according to this confidence variable. There might be cases where a goal location message is sent by a *ToM*$_2$ agent because its *ToM*$_1$ model expects this goal location message to yield a high utility. Hence, the expected value might be the highest for this combination of goal location message and offer. While we observe the *ToM*$_2$ agent sending the goal location message, it is a weighted decision between all its models. Because of this weighted decision, it is difficult to capture the intent of an agent sending a goal location message and thus determine whether an agent lies according to our definition.

Hence, in the following part of the thesis, we say that a *ToM*$_2$ lies whenever it sends a false goal location, a *ToM*$_1$ misleads whenever it sends a false goal location, and a *ToM*$_0$ agent simply makes a false statement whenever it sends a false goal location.

3.4 Implementation

We implemented the described agents in this chapter in the programming language Java. In addition, we made a graphical user interface (GUI) where two AI agents can play the Colored Trails game, and information about the negotiation is shown to the user. An example of a Colored Trails game in the GUI is given in Figure 3.7. The user takes the role of observer and cannot participate in the negotiation, although the GUI could be extended such that the user can partake in the negotiation. In Chapter C in the appendix, we explain some more details about where to find the code for the GUI and how to run it.



Colored Trails

Initiator (ToM=2)
initial chips: [2, 0, 0, 1, 1]
initial points: 300
---
final distribution chips: [0, 0, 0, 2, 1]
total nr. offers: 2
final points: 800

Messages:
0. I offer: [1, 0, 0, 2, 1] - [1, 0, 1, 1, 1]
- I accept your offer.

Legend
list index of chips:
0 1 2 3 4
* Gi: Goal location initiator.
* Gr: Goal location responder.
* X: Start location both agents.
* Offer: [Chips Init] - [Chips Resp]

Additional Information
- *Strict pareto improvements:*
* [0, 0, 0, 2, 1] - [2, 0, 1, 1, 1]; 800 - 350; sw=1150
- *Pareto improvements:*
* [0, 0, 0, 2, 2] - [2, 0, 1, 1, 0]; 850 - 300; sw=1150
* [0. 0. 0. 3. 1] - [2. 0. 1. 0. 1]: 850 - 300: sw=1150

Responder (ToM=1)
initial chips: [0, 0, 1, 2, 1]
initial points: 300
---
final distribution chips: [2, 0, 1, 1, 1]
total nr. offers: 2
final points: 350

Messages:
0. I offer: [0, 0, 0, 2, 1] - [2, 0, 1, 1, 1]

Figure 3.7: Example of a Colored Trails game using the graphical user interface, as explained in Section 3.4.

The center part of the GUI (see Figure 3.7) is the game board of the Colored Trails game. It contains the colored tiles, the start location (denoted by X), and the goal locations of the initiator (G_i) and the responder (G_r). The numbers in the top left of the twelve colored tiles denote the goal location index. The bottom center part of the GUI contains additional information about Pareto improvements in the game. (See Chapter A in the appendix for a more in-depth explanation of the Pareto principle in the context of this thesis.)

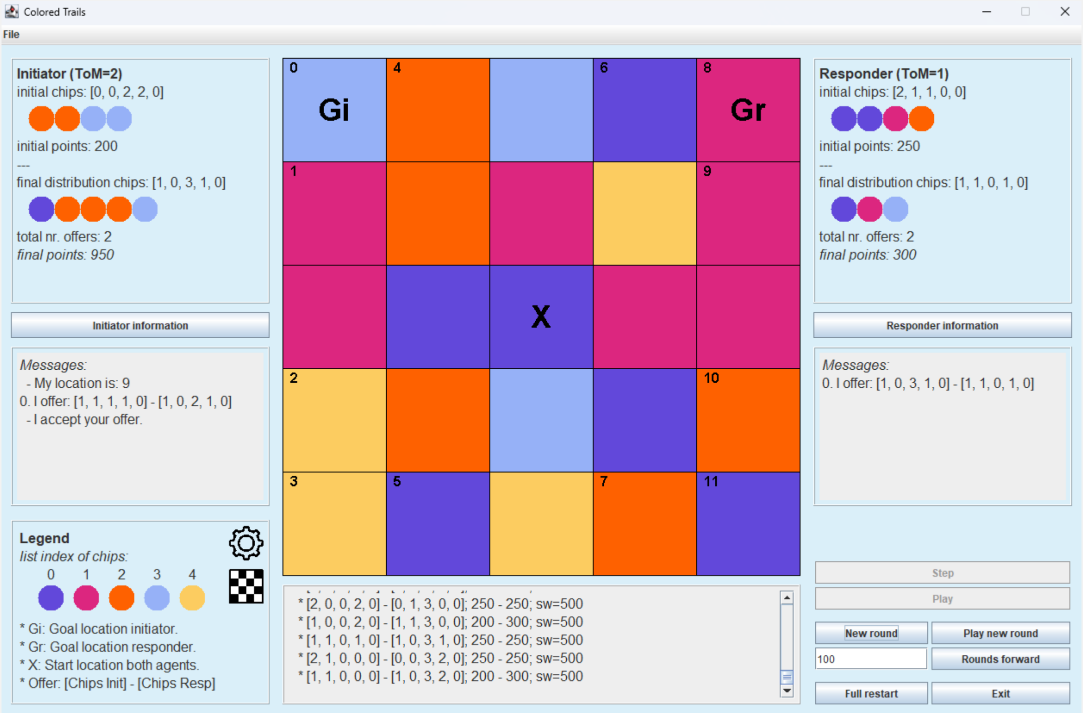
On the left and right of the GUI, we have the initiator and responder, respectively. The theory of mind level ($\texttt{ToM}$) is given next to the agent's name. In the same information box, the initial chips and the initial points are given. Moreover, when the negotiation round has ended, the final distribution of chips, the total number of offers from both agents, and the final points are given (without the points subtracted from making offers).

In the legend in the bottom left part of the GUI, the list indices of colors are given. Thus, an agent offering $[1, 2, 1, 0, 1]$ means that the agent offers a distribution where it offers one purple chip, two fuchsia-colored chips, one orange chip, zero light blue chips, and one yellow chip to itself.

During negotiation, agents can send offers and goal location messages to each other, which will be depicted in the `Messages` box. An offer is depicted by two lists. The first list of chips is always for the initiator, while the second list is for the responder. The user can obtain some extra information about (the beliefs of) the initiator and the responder by clicking on the `Initiator information` and `Responder information` buttons, respectively. Finally, in the bottom right corner of the GUI, there are buttons for the user. The user can view the next step in the negotiation round (`step`), view the agents playing the whole negotiation round (`play`), create a new negotiation round (`New round`), view the agents playing a new negotiation round (`Play new round`), forward a predefined number of negotiation rounds (`Rounds forward`), restart the agents in the negotiation with new beliefs (`Full restart`), and exit the game (`Exit`).

Example 6 (Lying agent in our GUI)**.** Recall Example 4 of a lying agent, where agent i lies about its goal location to obtain a certain response from agent r, its trading partner. This example is also shown in Figure 3.8 in our GUI. The initiator (agent i) calculates that sending a message with goal location 9 to the responder together with an offer that assigns the single fuchsia-colored chip to itself results in the highest expected value. Note that this is a lie because the goal location of the initiator is not goal location 9, but goal location 0. The offer that assigns the fuchsia-colored chip to the initiator is consistent with the goal location message announcing goal location 9 to the responder since this offer increases the points of the initiator if the initiator were to have goal location 9 as its actual goal location.

In this example, the initiator offers a purple, two orange, and a light blue chip to the responder. If the responder were to accept this offer, the responder could take no steps toward its goal location and would be left with four chips, resulting in 200 points, which is lower than its initial points of 250. Hence, the responder chooses not to accept the offer of the initiator. After receiving the goal location message together with the offer from the initiator, the responder updates its beliefs. Since the location message and offer are consistent with its beliefs, the responder believes the initiator and assigns a higher value to the fuchsia-colored chip. Since the responder also needs the fuchsia-colored chip to move one step toward its goal location, the responder does not offer the fuchsia-colored chip to the initiator, but instead, the responder offers more chips from another color to the initiator. The responder offers a purple, a fuchsia-colored, and a light blue chip to itself such that it can take three steps toward its goal location and obtain 300 points if the offer is accepted.



Colored Trails

Initiator (ToM=2)
initial chips: [0, 0, 2, 2, 0]
initial points: 200
---
final distribution chips: [1, 0, 3, 1, 0]
total nr. offers: 2
final points: 950

Messages:
- My location is: 9
0. I offer: [1, 1, 1, 1, 0] - [1, 0, 2, 1, 0]
- I accept your offer.

Legend
list index of chips:
0 1 2 3 4
* Gi: Goal location initiator.
* Gr: Goal location responder.
* X: Start location both agents.
* Offer: [Chips Init] - [Chips Resp]

Responder (ToM=1)
initial chips: [2, 1, 1, 0, 0]
initial points: 250
---
final distribution chips: [1, 1, 0, 1, 0]
total nr. offers: 2
final points: 300

Messages:
0. I offer: [1, 0, 3, 1, 0] - [1, 1, 0, 1, 0]

```
* [2, 0, 0, 2, 0] - [0, 1, 3, 0, 0]; 250 - 250; sw=500
* [1, 0, 0, 2, 0] - [1, 1, 3, 0, 0]; 200 - 300; sw=500
* [1, 1, 0, 1, 0] - [1, 0, 3, 1, 0]; 250 - 250; sw=500
* [2, 1, 0, 0, 0] - [0, 0, 3, 2, 0]; 250 - 250; sw=500
* [1, 1, 0, 0, 0] - [1, 0, 3, 2, 0]; 200 - 300; sw=500
```

Figure 3.8: Example of a lying agent in our GUI.

The initiator, who has lied about its goal location, now receives the offer from the responder and accepts the offer, because the initiator can now reach its goal location 0 and have one leftover chip, thereby reaching a total score of 950 points. Without the lie of the initiator, the responder might not have given an additional chip to the initiator.

3.5 Experiments

The goal of this thesis is to contribute to research on lying and deceiving by AI systems by investigating the influence of lying in a negotiation setting, i.e., the Colored Trails game. In particular, we seek an answer to what extent agents capable of lying and misleading outperform similar agents that are not capable of lying and misleading. In the previous sections of this chapter, we discussed agents that are capable of different levels of theory of mind and agents that are capable of sending and receiving goal location messages. In this section, we discuss the experiments performed in this thesis.

Before delving into the specifics of each experiment, we provide a parameter table in Section 3.5.1. The remainder of Section 3.5 discusses different experiments using the agents described in this chapter. Starting with Section 3.5.2, we describe an experiment to set the probability p_0 of a *ToM*$_0$ agent sending a goal location message together with an offer. This probability is then used in the remaining experiments. Section 3.5.3 follows by describing an experiment where our different agents negotiate with each other. In addition, we outline an experiment with the agents of De Weerd et al. (2017) in Section 3.5.4 to compare the performance of our agents that are capable of sending goal location messages with the agents of De Weerd et al. (2017). Since the learning speed λ influences the results (see De Weerd et al., 2017), we vary the learning speed in the experiment described in Section 3.5.5. Finally, we will look at specific games where a Pareto improvement is possible in the initial situation in an experiment that is described in Section 3.5.6.

Since previous research has shown that making the first offer is influential because it serves as an anchor for the entire negotiation round (Raiffa et al., 2002; Rosette et al., 2014; Van Poucke & Buelens, 2002), we differentiate between *initiators* who make the first offer in a negotiation round and *responders*. The results of the experiments discussed in this section can be found in Chapter 4.

3.5.1 Experiments settings and parameter table

Table 3.1 provides an overview of the parameters set in the experiments, their descriptions, and their values. Regarding our five-by-five game board of Colored Trails, we fix the number of possible goal locations to twelve, where each goal location is at least three tiles away from the start location, that is, the center square. There are five colors the chips and tiles of the board can take on.

In our experiments, we fix the learning speed λ of both agents to 0.5 unless mentioned otherwise. In particular, we vary the learning speed in Experiment 4 discussed in Section 3.5.5.

Note that a *ToM*$_0$ agent chooses each goal location that is not its actual goal location with probability $p_1 = 0.02$ when the *ToM*$_0$ agent sends a goal location message. Since our Colored Trails setting consists of twelve possible goal locations, this means that, when a *ToM*$_0$ sends a goal location message, the *ToM*$_0$ agent chooses its actual goal location with a probability of 0.78.

In each experiment, the first 100 negotiation rounds for all pairs of agents are considered to be a set-up phase for the zero-order beliefs of the *ToM*$_0$ agents and are not used for gathering results. Namely, at the start of a negotiation round, the zero-order beliefs of a *ToM*$_0$ agent are set to 1, that is, the agent believes that any offer will be accepted with probability 1. This is done by providing the agent with five positive encounters of each offer. After several negotiations, a *ToM*$_0$ agent will learn that an offer that offers zero chips

Parameter	Value	Description
$\|\mathcal{L}\|$	12	Number of possible goal locations.
$\|\mathcal{C}\|$	5	Number of possible colors the chips and the tiles of the board can take on.
λ	0.5	Agent-specific learning speed that represents the degree to which new information influences the beliefs of the agent. This value is changed in Experiment 4 (see Section 3.5.5).
p_0	[0,1]	The probability of a *ToM*$_0$ agent sending a goal location message with an offer. This value is set in Experiment 1 (see Section 3.5.2).
p_1	0.02	The probability of a *ToM*$_0$ agent choosing a specific goal location that is not its actual goal location when the *ToM*$_0$ agent sends a goal location message.
-	500	Points an agent obtains for reaching its goal location.
-	100	Points an agent obtains for each step in the shortest path to its goal location.
-	50	Points an agent obtains for each leftover colored chip.
-	-1	Points subtracted from the score for both agents for each offer made in the negotiation round.
-	5	Number of positive encounters of every offer type at instantiation of a new negotiation round to set the zero-order beliefs of a *ToM*$_0$ agent.

Table 3.1: A summary table containing parameters, their description, and their value. These values are fixed in this thesis, but they can be changed. The purpose of this table is to provide an overview of the parameters and their values in this thesis.

to the trading partner will not be accepted. Following De Weerd et al. (2017), we exclude games in which an agent can reach its goal location with its initial set of chips to ensure that both agents have the incentive to negotiate to increase their scores. However, we do not exclude these games in the warm-up phase, i.e., in the warm-up phase all possible game board variations are possible.

In theory, a negotiation round could never terminate because each agent could continue making an offer. Hence, we set a limit on the number of offers that can be made in a negotiation round: We set this limit to 100. However, agents are not aware of this limit and negotiate as if this limit does not exist. In case the limit is reached, the initial distribution of chips becomes final.

Our main measure will be the score gain for both agents. The score gain is calculated as the score of the agent after the negotiation round has ended (either an agent accepts an offer, withdraws from negotiation, or the limit of 100 offers has been reached) minus the initial score of the agent. In the results, we do not subtract the cost of making an offer, but the agents still reason as if one point is subtracted from their score for each offer they make. We can take the number of offers made in a negotiation round as a separate measure.

3.5.2 Experiment 1: Setting the probability of a *ToM*$_0$ agent sending messages

In Section 3.3.2, we discussed that a *ToM*$_0$ agent sends a goal location message with probability $p_0 \in [0,1]$. In this experiment, we determine whether probability $p_0 \in [0,1]$ influences the score gain for a *ToM*$_0$ agent against the average trading partner. Consequently, we set p_0 for the remaining experiments.

We let a *ToM*$_0$ agent as an initiator and as a responder negotiate with the five different agents as given in Section 3.3. We will vary $p_0 \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ in each of these interactions. When two *ToM*$_0$ agents negotiate with each other, they will be having the same probability p_0, that is, we do not gather results where *ToM*$_0$ agents have a different probability of sending a goal location message together with an offer. As a warm-up phase for the zero-order beliefs of the *ToM*$_0$ agents, we use 100 negotiation rounds for all pairs of agents, after which we gather one round of results. We repeat this 1000 times and measure the score gain of the *ToM*$_0$ agent.

We will average the score gain for each $p_0 \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ over the 1000 games, five different trading partners, and the *ToM*$_0$ agent as an initiator and responder. Consequently, we obtain for each $p_0 \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ a score gain for $(1000 \times$

$5 \times 2 =$) 10,000 games for the *ToM*$_0$ agent. Finally, p_0 is set in the remaining experiments such that p_0 is a good approximation of a possibly best probability for a *ToM*$_0$ agent to send a goal location message together with an offer against the average trading partner in terms of score gain.

3.5.3 Experiment 2: Does lying and misleading outperform honesty?

After setting p_0, that is, the probability that a *ToM*$_0$ agent sends a goal location message together with an offer, we perform an experiment where we let the five agents as discussed in Section 3.3 negotiate with each other, each performing the role as initiator and as responder. We will be using 100 negotiation rounds for all pairs of agents as a warm-up phase for the zero-order beliefs of *ToM*$_0$ agents, after which we gather one round of results. We repeat this 1000 times and measure, among other things, the score gain for both the initiator and the responder.

Agents that are capable of lying or misleading have an extra instrument, i.e., they can send goal location messages that contain a false goal location. Hence, in this experiment, we test the core hypothesis of whether lying and misleading agents outperform similar agents that are not capable of lying or misleading in terms of score gain.

3.5.4 Experiment 3: Is there a benefit to sending goal location messages in Colored Trails?

In this experiment, we let the agents of De Weerd et al. (2017) with different orders of theory of mind negotiate with each other and compare the results with the results of Experiment 2. For this, we let zero-order, first-order, and second-order theory of mind agents negotiate with each other, each performing the role of initiator and responder. We will be using 100 negotiation rounds for all pairs of agents as a warm-up phase for the zero-order beliefs of *ToM*$_0$ agents, after which we gather one round of results. Similar to the other experiments, we repeat this 1000 times and measure, among other things, the score gain for both the initiator and the responder. Moreover, if applicable, the settings of Table 3.1 also apply to these agents. In particular, the learning speed λ is set to 0.5.

We test the core hypothesis that agents with the ability to send goal location messages to each other obtain a higher average score gain compared to the agents of De Weerd et al. (2017). Moreover, we hypothesize that our agents will need fewer offers than the agents of De Weerd et al. (2017) to reach a mutually beneficial outcome.

3.5.5 Experiment 4: Varying the learning speed

In this experiment, we let the five agents as discussed in Section 3.3 negotiate with each other, each performing the role of initiator and responder, similar to Experiment 2. However, in this experiment, we also vary the learning speed λ of agents, which is an agent-specific learning speed that represents the degree to which new information influences the beliefs of the agent. With this experiment, we aim to shed light on the influence of the learning speed λ in the negotiation process of our agents.

3.5.6 Experiment 5: Games where a Pareto improvement exists

In the final experiment, we again let the five agents as discussed in Section 3.3 negotiate with each other, each performing the role of initiator and responder. De Weerd et al. (2017) only considered games where neither agent was initially able to reach its goal location with the purpose to give both agents an incentive to negotiate; however, this does not necessarily result in games where either agent has an incentive to negotiate. For example, consider a game board with only one color, say purple, and none of the agents has a purple-colored chip. In this case, neither agent can reach its goal location, but they both do not have an incentive to negotiate. Although this may be an extreme case, similar cases are not excluded.

In Experiment 5, we only consider games where neither agent can initially reach its goal location and where the initial state is Pareto inefficient, that is, there exists a Pareto improvement. Hence, we only consider games where, compared to the initial distribution of chips, there exists at least one offer where one of the agents can be better off in terms of the score without the other agent being worse off. In Chapter A in the appendix, one can revisit the general Pareto principle as well as a more in-depth explanation of the Pareto principle in the context of this thesis.

4
Results

This chapter presents the findings obtained from conducting the experiments described in Section 3.5 to address the research question and objectives outlined in the previous chapters. By examining the results, this thesis seeks to contribute to the existing research on lying and deceiving by AI systems. Hence, in the results, we mainly focus on whether there are differences between agents that are able to lie or mislead and agents that are honest.

Before delving into the results of each experiment, we would like to stress that the distribution of the score gain is discrete and takes on values that are multiples of 50. There might be cases where an agent makes an offer that results in a negative score gain because it expects a counteroffer, in which case a negative value (multiples of 50) is possible. Note that an agent would never accept an offer that yields a negative score gain, since the agent would always be better off withdrawing from negotiation. However, it is still possible that an agent makes an offer that results in the agent itself obtaining a negative score gain, but the agent expects its trading partner to make a counteroffer. Cases where an agent obtains a negative score gain are rare. The most common values of the score gain are in the set $\{0, 50, 100, 150, 200, 500, 550, 600, 650, 700, 750\}$. We did not encounter cases where an agent achieves a score gain of over 750 points. This is only possible when an agent secures six colored chips and, thus, the trading partner receives only two colored chips with which it cannot obtain a positive score gain. Furthermore, note that there is a gap between 200 and 500 points, which is mainly caused by the score gain of 500 points when an agent can reach its goal location and an agent not accepting offers with a nonzero score gain.

For example, a score gain of 200 is possible when an agent initially cannot make any step toward its goal location yielding 200 points (50 points for each chip) and after the final distribution, the agent can make three steps toward its goal location and has two left-over chips yielding 400 points, which results in a 200 points score gain. This distribution offers the trading partner three chips, which the trading partner can accept when the three chips yield a higher score than the four initial chips. With only two chips, an agent can obtain a score of at most 200 points (two steps toward its goal location). Hence, an agent will never accept or rarely make an offer that offers itself two chips.

To get a higher score gain than 200 points, an agent will need to reach its goal location, resulting in 500 points. Consequently, a score gain between 200 and 500 points is not possible with the trading partner also being better off. Furthermore, a score gain of over 750 points is not possible with the trading partner being better off since the lowest initial number of points is 200 and the highest final points with at most five chips (and thus at least three chips for the trading partner) is 950 points (four steps toward the goal location, reaching the focal agent's goal location, and having one left-over chip) yielding a score gain of 750 points.

As a final note, we did not encounter many cases in our experiments where the limit of 100 offers was reached in a negotiation round; hence, this limit did not influence the length of the negotiation in general. Exceptions, where we did find negotiations where agents reached this limit, are Experiments 4 and 5. In Experiment 4, for example, the limit of 100 offers is reached for low values for the learning speed.

4.1 Experiment 1: Setting the probability of a *ToM*$_0$ agent sending messages

The results in this section are obtained by conducting the experiment as described in Section 3.5.2 to set the probability p_0 of a *ToM*$_0$ agent sending a goal location message together with an offer.

Figure 4.1 shows the score gain of the *ToM*$_0$ agent in a box plot of the 10,000 data points for each value of probability p_0. The whiskers of this box plot capture at least 95%

of the data, so the outliers drawn in the plot are points that are not in reach of the whiskers and are either in the top or the bottom 2.5% of the data points. Moreover, the mean value for each of the probabilities $p_0 \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$ is indicated by a green diamond.

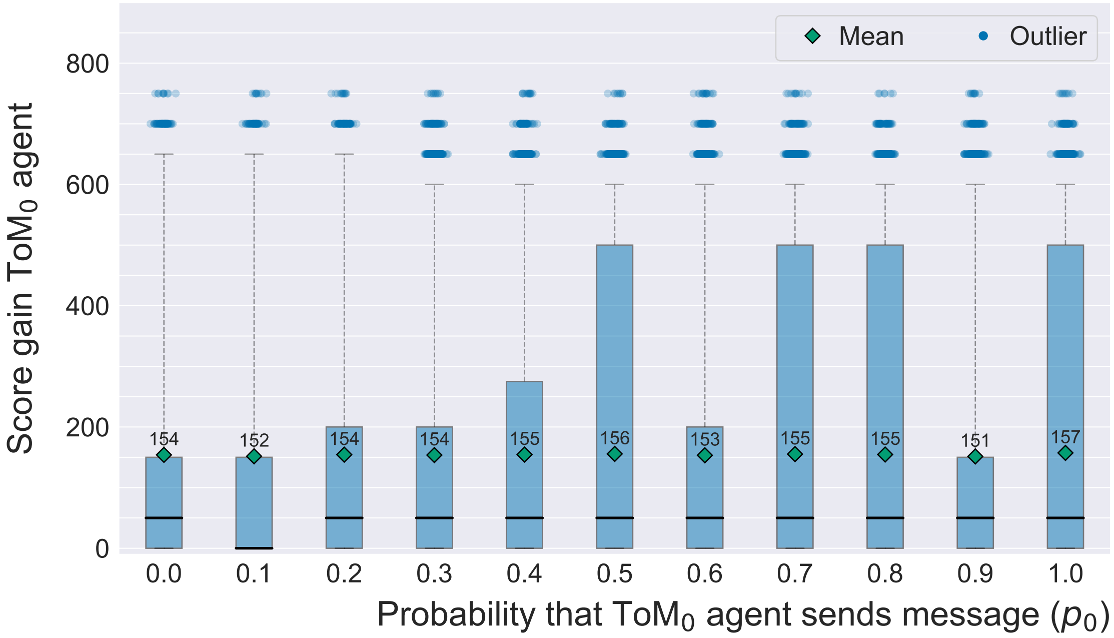


Figure 4.1: A box plot of the score gain per value of p_0, i.e., the probability of a *ToM*$_0$ agent sending a goal location message together with an offer. In this experiment, a *ToM*$_0$ agent negotiated as initiator and as responder in 1000 negotiation rounds with the five different agents as described in Section 3.3. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the x-axis to increase readability.

From Figure 4.1, it might seem that some of the probabilities p_0 yield higher scores than the other probabilities p_0 because of the height of the third quartile of the box plot indicating a different distribution of score gains. However, the seemingly big difference between the box plots is more likely to be caused by the odd distribution of the score gain as mentioned at the beginning of this chapter. In Figure B.1 in the appendix, we plotted the data points on top of this figure. It becomes clear that there are no results where the score gain is between 200 and 500 points. Hence, an extra data point above 200 points might change the view of the box plot drastically.

The score gain varies from 0 to 750 points. The median score gain is 50 for each of the probabilities p_0 except for $p_0 = 0.1$. We can test whether the data points for the eleven groups originate from the same distribution using the one-way analysis of variance (ANOVA). However, since we cannot assume normality (of the residuals) of data points for each of the probabilities p_0, we use the non-parametric alternative of ANOVA, that is, the Kruskal-Wallis test (Kruskal & Wallis, 1952). The Kruskal-Wallis test does not assume the normality of data and is less sensitive to (extreme) outliers, which are both problems in our data. Hence, a Kruskal-Wallis test was conducted to compare the score gain among the eleven groups with different p_0 values. The Kruskal-Wallis test did not reveal a significant difference between the groups (H=8.99, df=10, p=.53), indicating that there is insufficient evidence to conclude a difference in score gain among the groups. We might, therefore, argue that the probability of a *ToM*$_0$ agent sending a goal location message together with an offer does not influence the negotiation score gain of the *ToM*$_0$ agent in general.

The means and standard deviations for each of the probabilities p_0 are in the range $[151; 157]$ and $[231; 236]$, respectively. Figure 4.2 shows a bar plot of the sample means with their 99.91% Bonferroni-adjusted confidence intervals to ensure a family-wise error rate of less than 0.05 (Dunn, 1961). The error margins are calculated using the t-statistic with the following formula:

$$\text{margin of error} = t_{1-(\alpha/k),n-1} \cdot \frac{s}{\sqrt{n}}, \tag{4.1}$$

where s is the standard deviation, $\alpha = 0.05$ the significance level, $k = \binom{11}{2} = 55$ the number of comparisons for the Bonferroni adjustment, and $n = 10{,}000$ the number of observations per category. Note that the Bonferroni adjustment is rather conservative and other multiple comparison adjustments exist. However, we chose to err on the side of caution and use the intuitive Bonferroni adjustment.

The 99.91% (Bonferroni-adjusted) confidence intervals around the sample mean indicate that if we were to repeat this experiment an infinite number of times, we would expect that 99.91% of such calculated confidence intervals contain the true population mean. Note that when comparing two mean values at a significance level of, e.g., 0.05, we should use 83.4% confidence intervals around the mean instead of 95% confidence intervals (see, e.g., Goldstein & Healy, 1995; Knol, Pestman, & Grobbee, 2011). This means that, in contrast to expanding the confidence intervals to capture the multiple comparisons, one should decrease the confidence intervals around the mean to increase the power of the test. However, for interpretability and to err on the side of caution, we only used the Bonferroni adjustment.

The results might still be inconclusive as to which probability p_0 is best in terms of score gain against the average trading partner, and, thus, to which value we set p_0. Hence, we performed an additional experiment, in the same context, where we determined the fraction

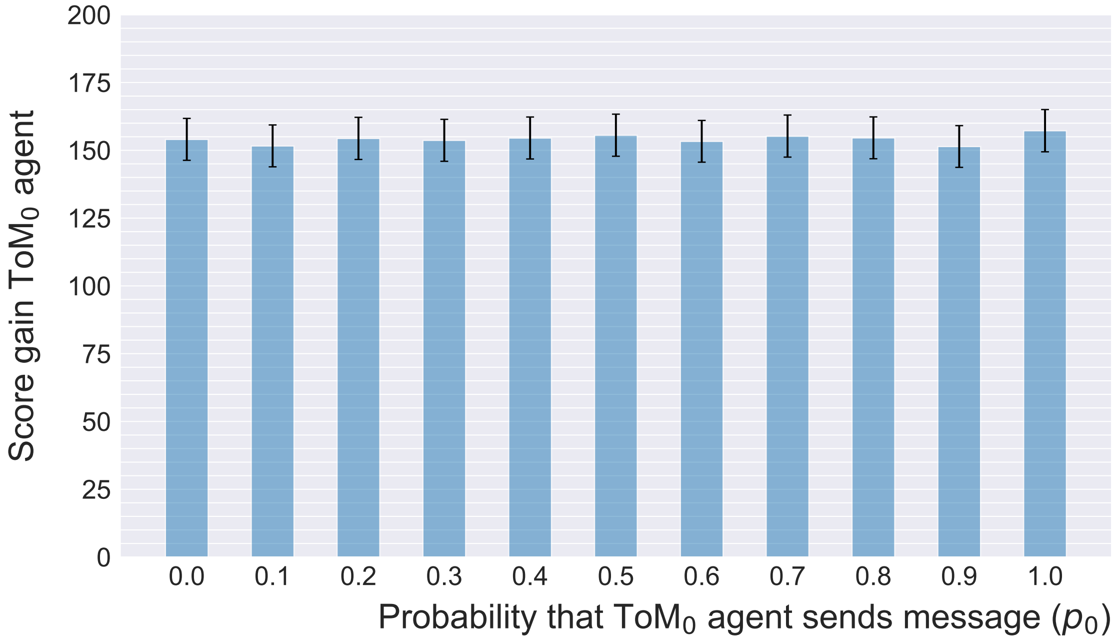


Figure 4.2: A bar plot of the mean score gain per value of p_0, i.e., the probability of a *ToM*$_0$ agent sending a goal location message together with an offer. The means contain 99.91% Bonferroni-adjusted confidence intervals (55 comparisons) to ensure a family-wise error rate of less than 0.05. The 99.91% confidence intervals are constructed using a t-statistic. If two confidence intervals do not overlap, we have evidence to conclude that the mean values significantly differ between groups.

of offers that are accompanied by a goal location message for the other four types of agents (see Section 3.3). We let each of the four types of agents (excluding *ToM*$_0$ agents) negotiate as initiator and responder with all five types of agents as trading partners. We set the probability of *ToM*$_0$ agents sending a goal location message together with an offer to zero, that is, *ToM*$_0$ agents were not able to send goal location messages. Note that the *ToM*$_0$ agents are still able to interpret goal location messages. As a warm-up phase for the zero-order beliefs of *ToM*$_0$ agents, we used 100 rounds of negotiations. After the warm-up phase, we collected results for one round of negotiation. This was repeated 1000 times for each of the four types of agents as an initiator and a responder with all five types of agents as trading partners.

The total number of offers made and the total number of goal location messages sent were recorded per agent type. We divided the total number of goal location messages sent by the total number of offers made to get an estimate of the fraction of offers that are accompanied by a goal location message. The results are shown in Figure 4.3. A chi-square test of independence was performed to examine the relationship between agent type and the proportion of offers that is accompanied by a goal location message. The relation between these variables was significant, χ^2(3, N=50,081)=3188.73, p<.00001, indicating that agent type influences the proportion of offers that are accompanied by a goal location message.

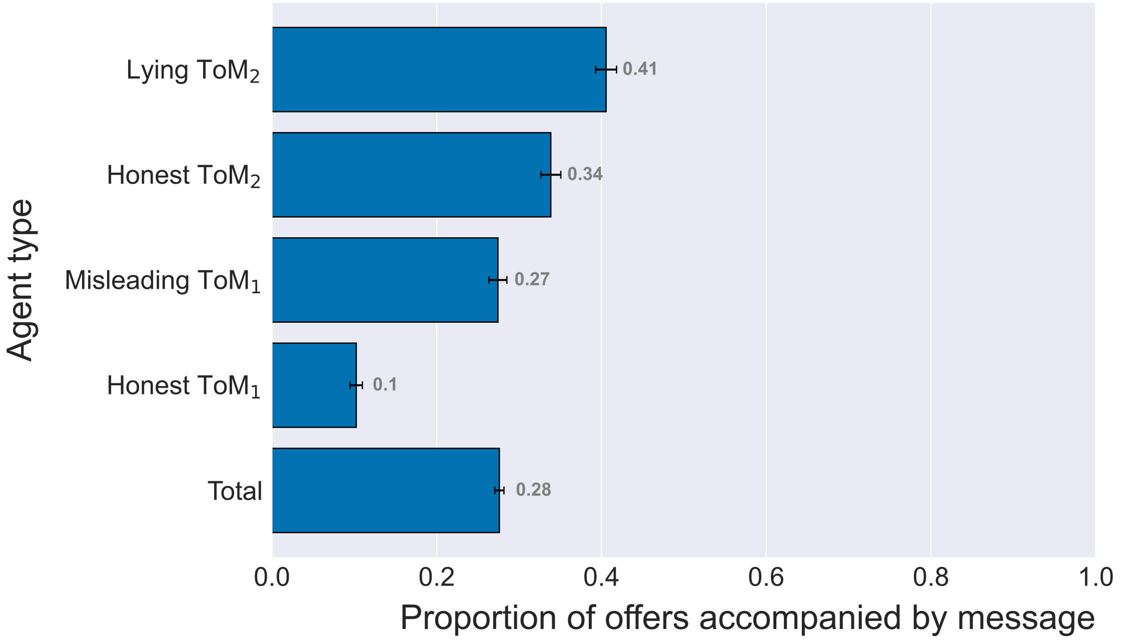


Figure 4.3: Bar plot of the fraction of offers that are accompanied by a goal location message for four types of agents (see Section 3.3). The *Total* bar indicates the weighted average fraction over all four types of agents. The proportions contain 99.5% Bonferroni-adjusted confidence intervals (10 comparisons) to ensure a family-wise error rate of less than 0.05. The 99.5% confidence intervals are constructed using the Wald method. If two confidence intervals do not overlap, we have evidence to conclude that the proportions significantly differ between agent types.

Figure 4.3 suggests that lying and honest ToM_2 agents send more goal location messages than misleading and honest ToM_1 agents. Moreover, misleading ToM_1 agents send more goal location messages than honest ToM_1 agents, and lying ToM_2 agents send more goal location messages than honest ToM_2 agents. This result is to be expected, since lying ToM_2 agents and misleading ToM_1 agents can send false goal location messages in addition to the capabilities of honest ToM_2 agents and honest ToM_1 agents, respectively.

While the average fraction of goal location messages sent with an offer is 0.28, we observe that ToM_1 agents send fewer goal location messages than ToM_2 agents. Hence, all points considered, we round probability p_0 down to 0.2, that is, agents with a zero-order theory of mind capability send, on average, goal location messages together with 20% of their offers. This value for p_0 will be used in the subsequent experiments.

4.1.1 Separating *ToM*$_0$ initiators and *ToM*$_0$ responders

Before discussing the main experiment in Section 4.2, it is interesting to note the differences between the results of the experiment discussed in this section for initiators and responders. Note that both the initiator and the responder negotiate with the five different agents. Figure 4.4 shows the same results as in Figure 4.1 but the results of the *ToM*$_0$ agent as initiator and as responder are separated.

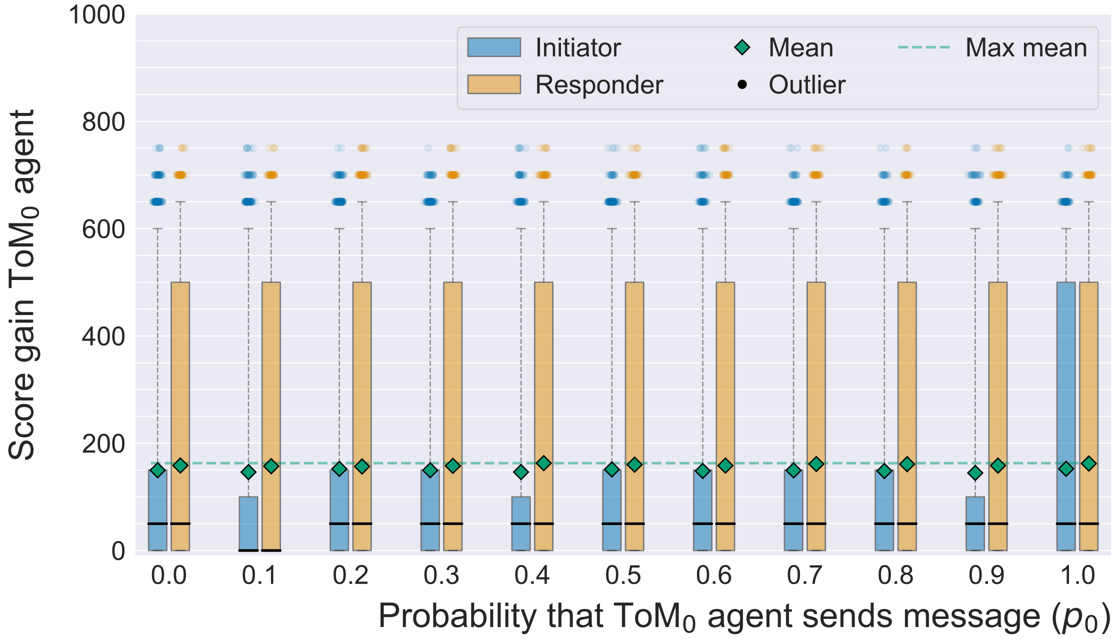


Figure 4.4: Box plot of the score gain per value of p_0, i.e., the probability of a *ToM*$_0$ agent sending a goal location message with an offer, separated into initiators and responders. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the x-axis to increase readability. The green dashed line indicates the value of the highest mean point and serves as a reference line.

While the median value is 50 for almost all values of p_0 (except for $p_0 = 0.1$) and for both the initiator and the responder, the mean values differ consistently between the *ToM*$_0$ agents as initiator and responder. The *ToM*$_0$ responder consistently obtains a higher mean score gain over the different probabilities p_0 than the *ToM*$_0$ initiator.

To test for each of the probabilities p_0 whether there is a difference in mean score gain between the initiator and responder, we plotted the differences in means in Figure B.2 in the appendix. We added 99.5% Bonferroni-adjusted confidence intervals (11 comparisons) to control that the family-wise error rate is lower than 0.05. We observe that the differences in means are significant only for some p_0 values (in particular $p_0 \in \{0.4, 0.9\}$). In general, we cannot reject the null hypotheses of the differences in means of the responder and initiator to be equal to zero.

A similar separation between the results of the initiator and responder is done for the results shown in Figure 4.3, where we plotted the estimated proportion of offers accompanied by a message for each type of agent (excluding ToM_0 agents). The bar plot where the results are separated for the initiator and responder for the four different types of agents is shown in Figure 4.5.

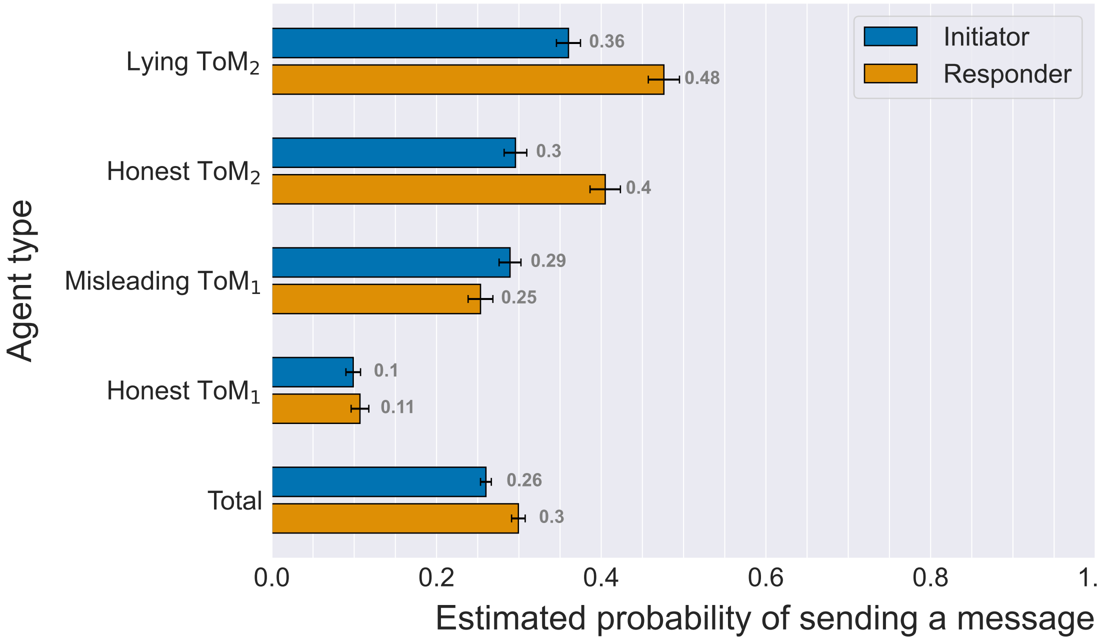


Figure 4.5: Bar plot of the fraction of offers that are accompanied by a goal location message for four types of agents (see Section 3.3) where the results are separated for initiators and responders. The *Total* bar indicates the weighted fraction over all four types of agents. The proportions contain 99% Bonferroni-adjusted confidence intervals (5 comparisons) to ensure a family-wise error rate of less than 0.05. The 99% confidence intervals are constructed using the Wald method. If the confidence intervals of the initiator and the responder (of similar agent types) do not overlap, we have evidence to conclude that the proportions significantly differ from each other.

From Figure 4.5 it becomes apparent that overall the responder sends a goal location message with a higher fraction of offers than the initiator. This result is especially caused by ToM_2 responders sending significantly more goal location messages together with their offers compared to ToM_2 initiators.

4.2 Experiment 2: Does lying and misleading outperform honesty?

The results in this section are obtained by conducting the experiment as described in Section 3.5.3, where the five different agents negotiate with each other. In total, we obtained results of 25,000 negotiations. In 11.0% of the negotiations, the negotiation process was terminated by the initiator before an initial offer was made. In 53.5% of the negotiations, a new distribution of colored chips became final.

Since each agent negotiated with every other agent in 1000 experiments, both as initiator and responder, we have 10,000 score gains for each agent type. A box plot of the score gain per agent type is given in Figure 4.6. The median values are 50 for each of the agent types but the mean values differ. Moreover, the distribution mass of the score gain is more located to the lower values for the *ToM*$_0$ agent, which can be seen from the location of the third quartile of the box plot. However, this can also be due to the odd distribution of the data as mentioned at the beginning of this chapter.

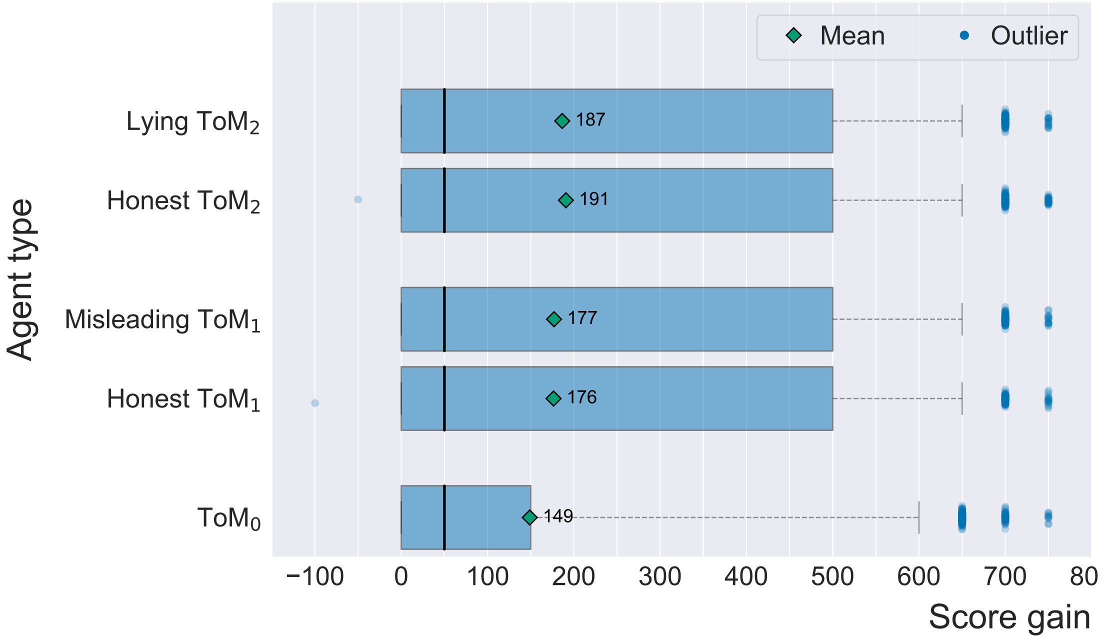


Figure 4.6: Box plot of the score gain for the five different types of agents (see Section 3.3). The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

A Kruskal-Wallis test was conducted to compare the score gain among the five agents. The Kruskal-Wallis test reveals a significant difference between the groups (H=124.53, df=4, p<.0001), indicating that there are differences in the distribution of the score gain between the groups. A post hoc test was done on the differences in means. The results are shown in Figure 4.7, where the differences in means of ten combinations of agents are shown together with 99.5% Bonferroni-adjusted confidence intervals (10 comparisons) such that the family-wise error rate is below 0.05. This graph indicates that the differences in mean score gain are significantly different from zero for all combinations of agents except for combinations where the agents have a similar theory of mind capability.

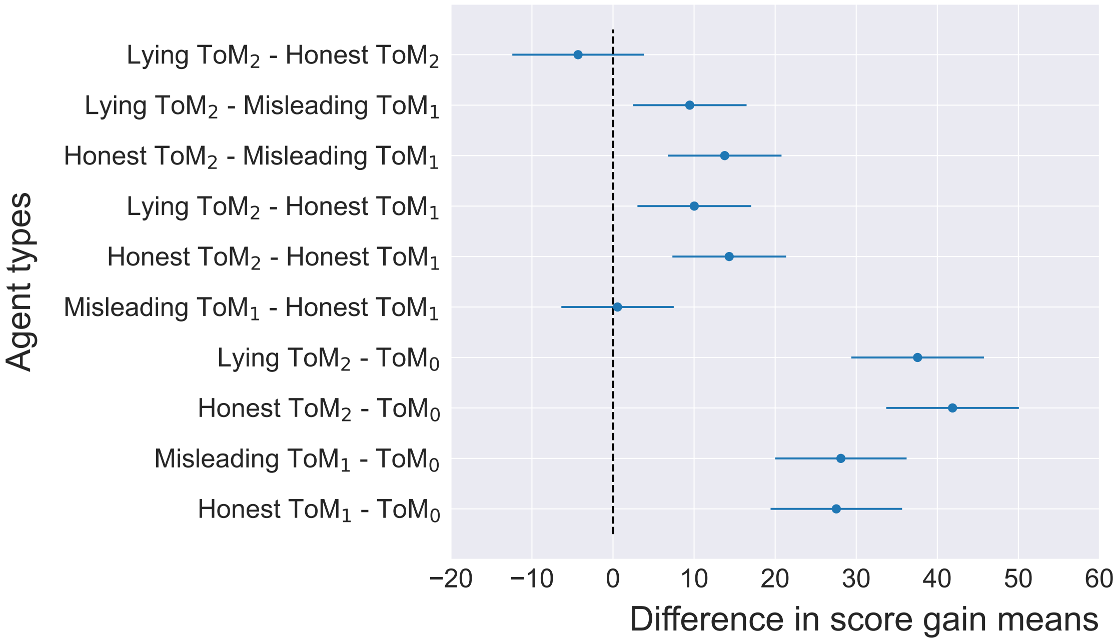


Figure 4.7: Difference in score gain mean values between the five different types of agents (see Section 3.3). The differences in means contain 99.5% Bonferroni-adjusted confidence intervals (10 comparisons) to ensure a family-wise error rate of less than 0.05. The 99.5% confidence intervals are constructed using a t-statistic. If a confidence interval contains 0, we do not have enough statistical evidence to conclude that the difference is unequal to zero, i.e., that there is a difference in score gain mean values between two different agents.

Interestingly, although a misleading ToM_1 agent can send false goal location messages, the misleading ToM_1 agent does not significantly outperform the honest ToM_1 agent. Since a misleading ToM_1 agent can send goal location messages containing a false goal location, it has more possible actions than an honest ToM_1 agent. While a misleading ToM_1 agent has this extra toolkit, it does not benefit the agent in terms of mean score gain. Moreover, when we compare a lying ToM_2 agent with an honest ToM_2 agent, we observe that the mean score gain is slightly lower for the lying ToM_2 agent. Even though this difference is not statistically significant, the extra toolkit of making false statements does not benefit a *ToM*$_2$ agent.

There are two cases where an agent achieves a negative score gain, as can be seen from Figure 4.6. Upon further inspection, these cases concern a *ToM*$_1$ agent and a *ToM*$_2$ agent who make an offer that decreases their score and gets accepted by their trading partners. A reason for a *ToM*$_2$ agent to send an offer that decreases its score is to deceive the trading partner into not believing that the actual goal location of the agent is the goal location mentioned. The *ToM*$_2$ agent then expects that the trading partner makes a better offer. However, in these two cases, the offer made by the focal agent was accepted by the trading partner, yielding the agent a negative score gain. While a *ToM*$_1$ agent does not have the capability to deceive the trading partner into believing a false goal location for the agent (recall Definition 2 of deceiving), it can still expect the trading partner to make a better counteroffer as a response to its own offer.

In Figure 4.8, we distinguish between the results with respect to the agents being initiators and responders. There are no differences between the median score gains for the initiators and responders, but there may be differences between the mean scores of the initiators and the responders.

In Figure 4.9, we plotted the differences in means with 99% Bonferroni-adjusted confidence intervals (5 comparisons) to control that the family-wise error rate is lower than 0.05. We observe that none of the differences in mean score gain are significantly different from zero. While not significant, the difference between the initiator and responder in terms of score gain for a *ToM*$_0$ agent is in favor of the responder. This result is in line with the results found in Section 4.1, where for each value of p_0 it was found that the *ToM*$_0$ responder slightly outperforms the *ToM*$_0$ initiator against the average trading partner in terms of mean score value. Another interesting point that may be highlighted from Figure 4.9 is that only an agent that has a first-order theory of mind capability is better off being the initiator.

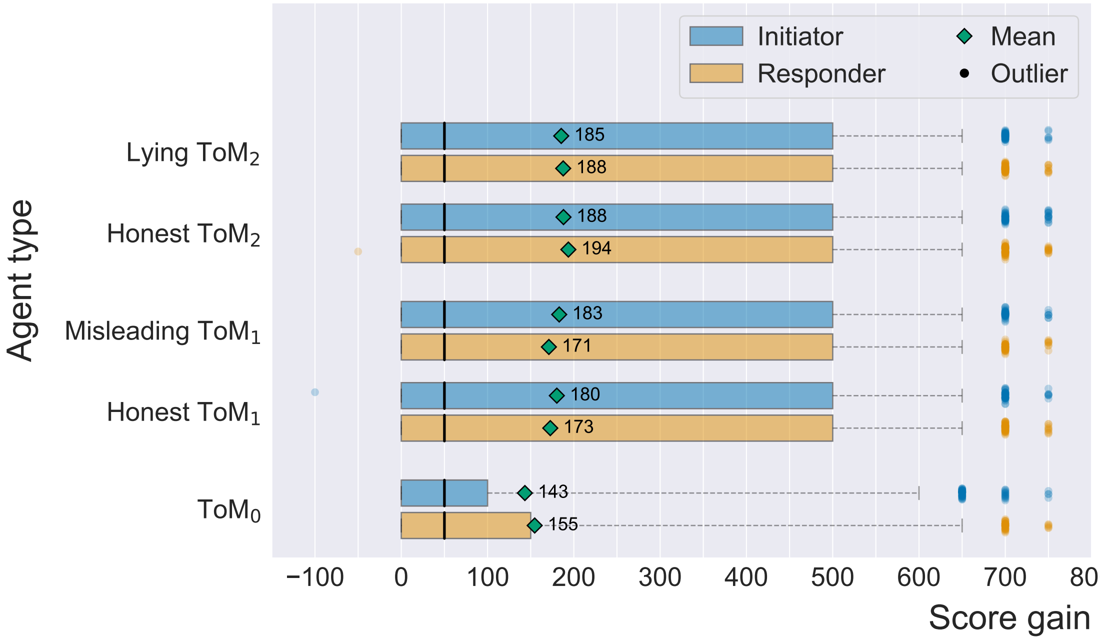


Figure 4.8: Box plot of the score gain for the five different types of agents (see Section 3.3) separated into initiators and responders. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

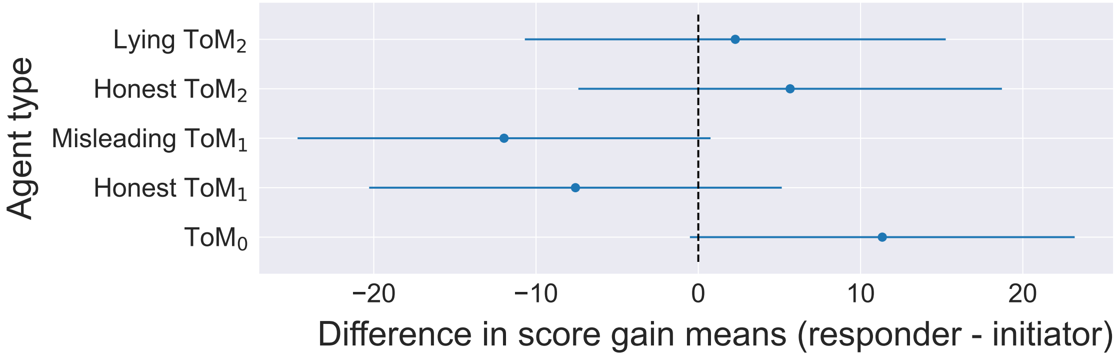


Figure 4.9: Difference in score gain mean values between the responder and initiator for different types of agents (see Section 3.3) with 99% Bonferroni-adjusted confidence intervals (5 comparisons) to ensure a family-wise error rate of less than 0.05. The confidence intervals are constructed using a t-statistic. If a confidence interval contains 0, we do not have enough statistical evidence to conclude that the difference is unequal to zero, i.e., that there is a difference in score gain mean values between the responder and the initiator.

4.2.1 Investigating other metrics

Figure 4.10 shows a box plot of the total number of offers in a negotiation round when a specific agent type takes part in the negotiation. Here, the number of offers per agent type depends on each other since each agent type negotiates with each agent type. However, we observe that in negotiations where ToM_0 agents are participating, the total number of offers does not reach high values. This can be seen from the maximum total number of offers being equal to 5 when a ToM_0 agent participates in the negotiation. In contrast, the mean total number of offers in a negotiation round is lower for lying and honest ToM_2 agents than for the other agent types. This result is more pronounced when we look at the total number of offers in a negotiation round when the initiator and responder are of the same agent type. In that case, we observe more clearly that lying and honest ToM_2 agents need, on average, fewer offers than ToM_1 and ToM_0 agents (see Figure B.3 in the appendix).

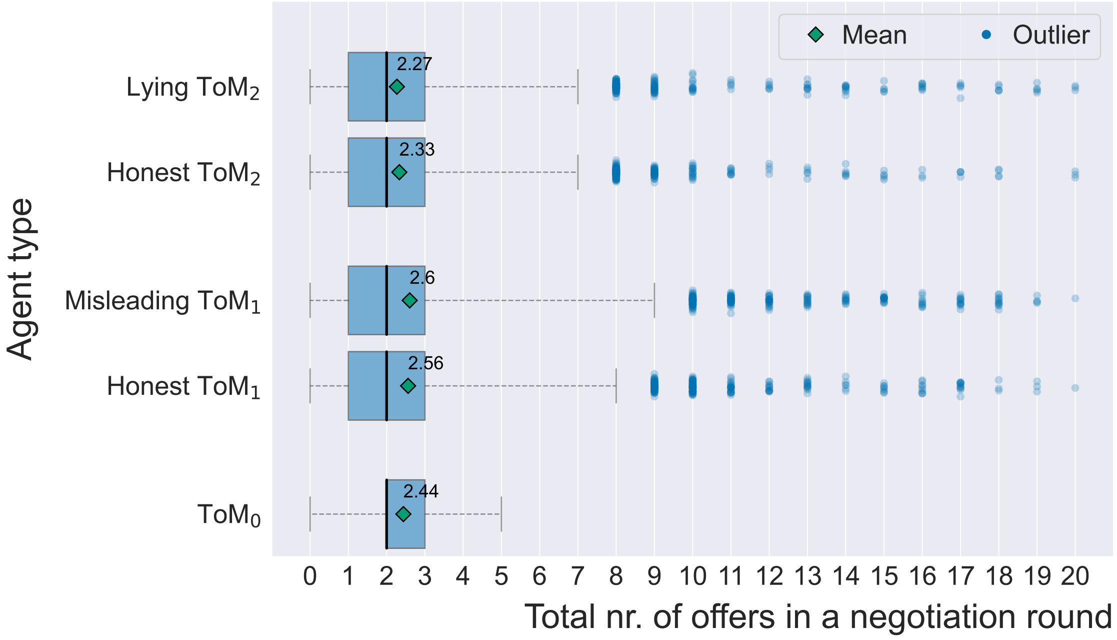


Figure 4.10: A box plot of the total number of offers in a negotiation round. In this plot, data points are gathered where each agent type negotiates with each agent type, both as responder and initiator. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

Figure 4.11 shows the fraction of offers that are accompanied by a true goal location message, a false goal location message, or no goal location message. Compared to the results in Figure 4.3, Figure 4.11 shows a similar fraction of offers that are accompanied by a goal location message (either true or false). This implies that the fraction of offers that are accompanied by a goal location message is not influenced by including negotiation results with a *ToM*$_0$ agent (that is capable of sending goal location messages).

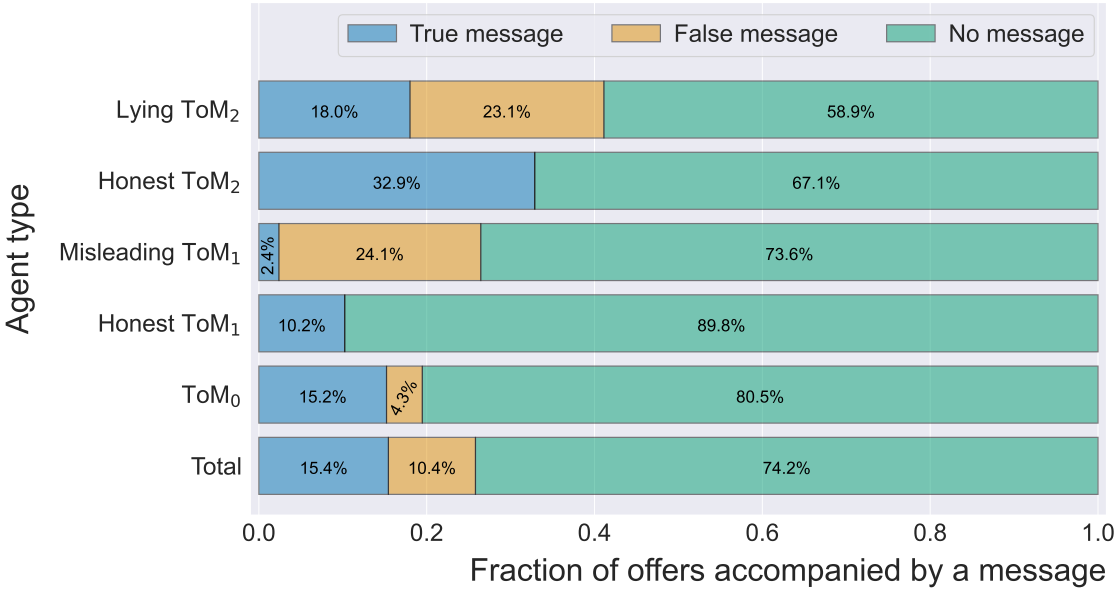


Figure 4.11: A stacked bar plot of the fraction of offers that are accompanied by a true goal location message, a false goal location message, or no goal location message. The *Total* bar indicates the weighted fraction of offers over all five types of agents.

A lying *ToM*$_2$ agent sends more goal location messages than an honest *ToM*$_2$ agent, which is to be expected since the capabilities of a lying *ToM*$_2$ agent extend the capabilities of an honest *ToM*$_2$ agent. Similarly, a misleading *ToM*$_1$ agent sends more goal location messages than an honest *ToM*$_1$ agent. Interestingly, the fraction of offers that are accompanied by a true goal location message is not the same for an honest agent compared to its lying or misleading counterpart. The number of true goal location messages increases significantly for honest *ToM*$_2$ agents compared to lying *ToM*$_2$ agents, χ^2(1, N=23,668)=687.18, p<.00001. Moreover, the number of true goal location messages increases significantly for honest *ToM*$_1$ agents compared to misleading *ToM*$_1$ agents, χ^2(1, N=25,800)=673.76, p<.00001. While a lying *ToM*$_2$ agent lies in approximately 56% of its sent goal location messages, a misleading *ToM*$_1$ agent misleads in approximately 91% of its sent goal location messages.

Figure 4.12 shows the fraction of offers that are accompanied by a true message, a false message, or no message, separated by initiators and responders. While lying *ToM*$_2$ initiators and responders send true messages with approximately the same fraction of offers, lying *ToM*$_2$ responders send false goal location messages with a higher fraction of offers than lying *ToM*$_2$ initiators. Honest *ToM*$_2$ responders send true goal location messages with a higher fraction of offers than honest *ToM*$_2$ initiators.

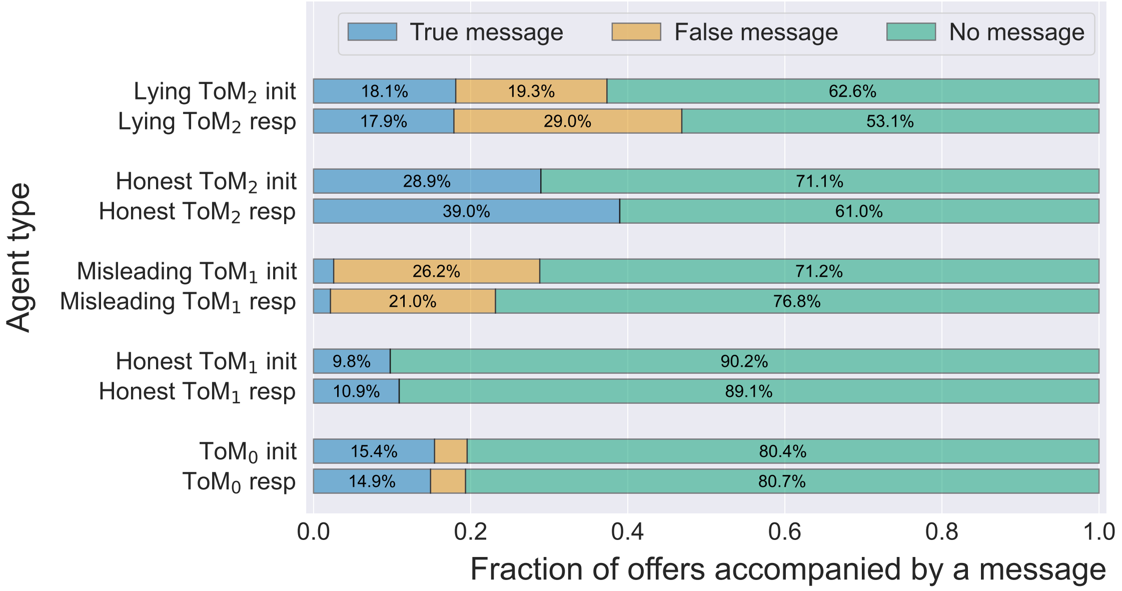


Figure 4.12: A stacked bar plot of the fraction of offers that are accompanied by a true goal location message, a false goal location message, or no goal location message with the results separated by initiators (init) and responders (resp).

Finally, Figure 4.13 shows a box plot of the social welfare gain of every negotiation an agent type participates in. As seen in Figure 4.6, the score gain data points are skewed to the right. This is mainly due to the many negotiations that yield a social welfare gain of zero and the odd distribution of the score gain. This right-skewness is also pronounced in Figure 4.13. From Figure 4.13, we observe that negotiations, wherein an honest *ToM*$_2$ agent participates, yield the highest mean social welfare gain. The results suggest that negotiations, where agents with a higher theory of mind capability participate, yield a higher mean social welfare gain. However, the difference between lying *ToM*$_2$ agents and honest *ToM*$_1$ agents in terms of the mean social welfare gain is rather small.

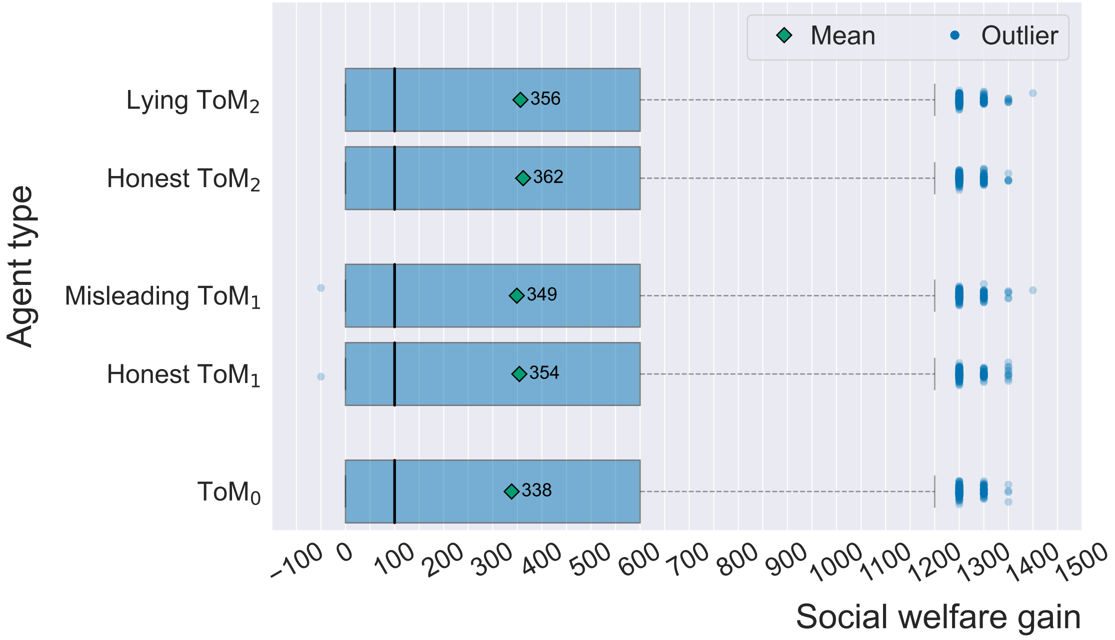


Figure 4.13: Box plot of the social welfare gain for the five different types of agents (see Section 3.3). The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

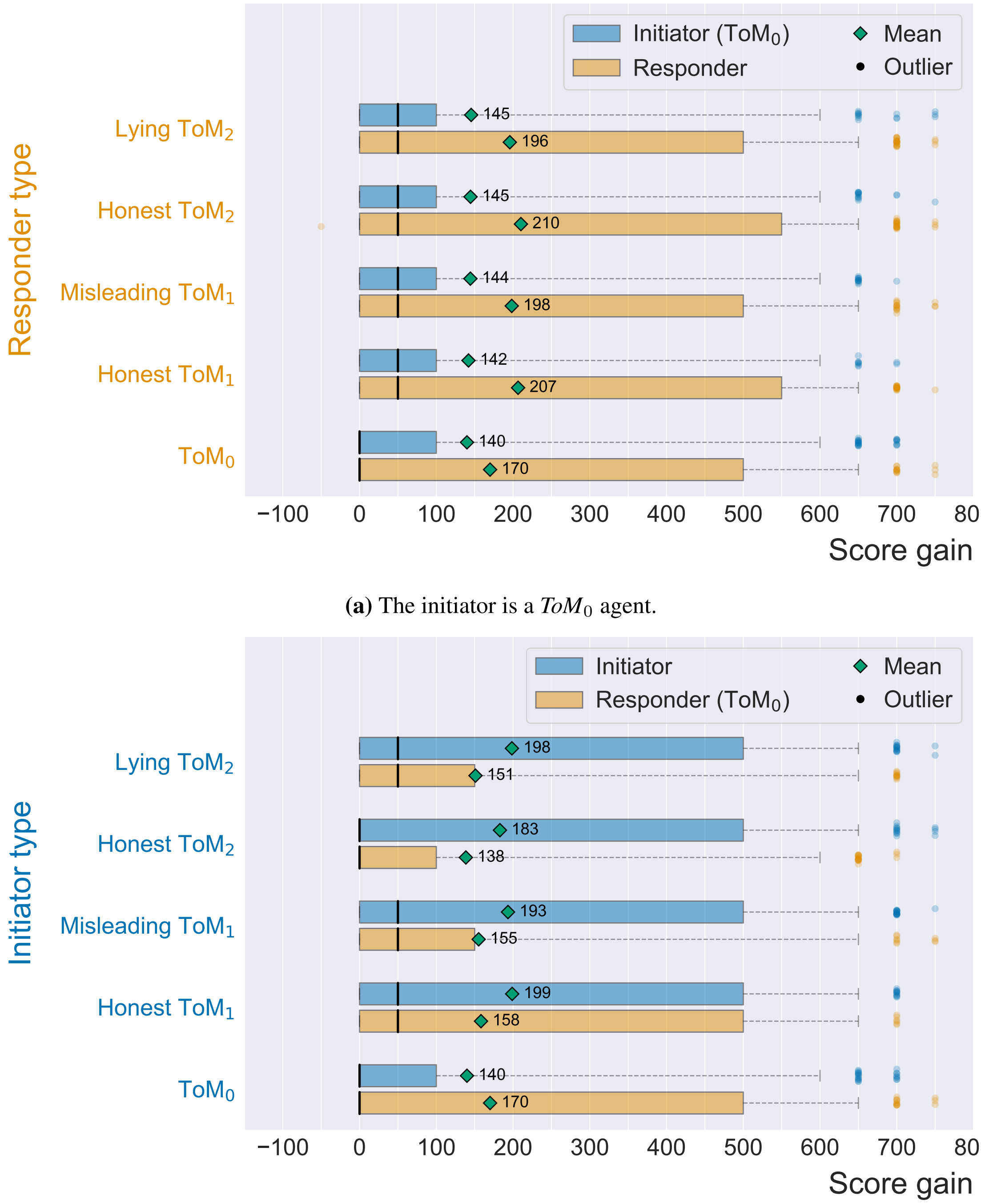
4.2.2 Comparing the score gain for each combination of agent types

We will now investigate each agent type by looking at the score gain results where we also separate the results by initiators and responders. The following figures fix one of the five agent types and show the results for that agent as initiator and responder.

Each of the following box plots follows from 1000 data points. The standard deviation deviates per combination of agent types and ranges between 230 and 270. This means that the standard error of a mean score gain is between 7 and 9. Practically this means that, since the standard error of a difference in means is even higher, differences in the mean score gain greater than 20 points might be relevant to be tested in order to reveal a significant difference. However, a multiple-comparison correction should also be applied such as the Bonferroni correction to ensure a family-wise error rate of less than 0.05. To err on the side of caution, in this section, we will not explicitly mention whether a difference is significant based on its p-value, but rather highlight some of the differences in mean score gain.

***ToM*$_0$ agent.** Figure 4.14 shows the results when we fix the initiator or responder to a *ToM*$_0$ agent. It becomes apparent that a *ToM*$_0$ agent is outperformed by every other agent in terms of mean score gain. A *ToM*$_0$ initiator performs rather homogeneously against each of the different trading partners, while the score gain of a *ToM*$_0$ responder varies considerably against different types of trading partners. A *ToM*$_0$ responder performs best against a *ToM*$_0$ initiator in terms of mean score gain, and the *ToM*$_0$ responder performs worst against an honest *ToM*$_2$ initiator.

When comparing lying with honest *ToM*$_2$ agents and misleading with honest *ToM*$_1$ agents, we observe that honest *ToM*$_1$ and honest *ToM*$_2$ agents perform better than their lying and misleading counterparts against a *ToM*$_0$ initiator. However, due to the high variance in the data, misleading *ToM*$_1$ agents (M=198.3, SD=258.92) did not significantly differ from honest *ToM*$_1$ agents (M=206.55, SD=264.31), both set against a *ToM*$_0$ initiator, in terms of score gain, t(1997.2)=0.70, p=.48. Moreover, there is no significant evidence that lying *ToM*$_2$ agents (M=195.8, SD=260.18) perform differently from honest *ToM*$_2$ agents (M=210.25, SD=267.23), both set against a *ToM*$_0$ initiator, in terms of score gain, t(1996.6)=1.23, p=.22. Another point of interest is that the results suggest that the honest *ToM*$_2$ agent (M=210.25, SD=267.23) set against a *ToM*$_0$ initiator performs better than an honest *ToM*$_2$ agent (M=182.9, SD=252.14) set against a *ToM*$_0$ responder, t(1991.3)=2.35, p=.019. Thus, the results suggest that an honest *ToM*$_2$ agent is better off being the responder against a *ToM*$_0$ agent than being the initiator.



(a) The initiator is a *ToM_0* agent.

(b) The responder is a *ToM_0* agent.

Figure 4.14: Box plots of the score gain for the five different agents where the initiator (a) or the responder (b) is fixed as a *ToM_0* agent. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

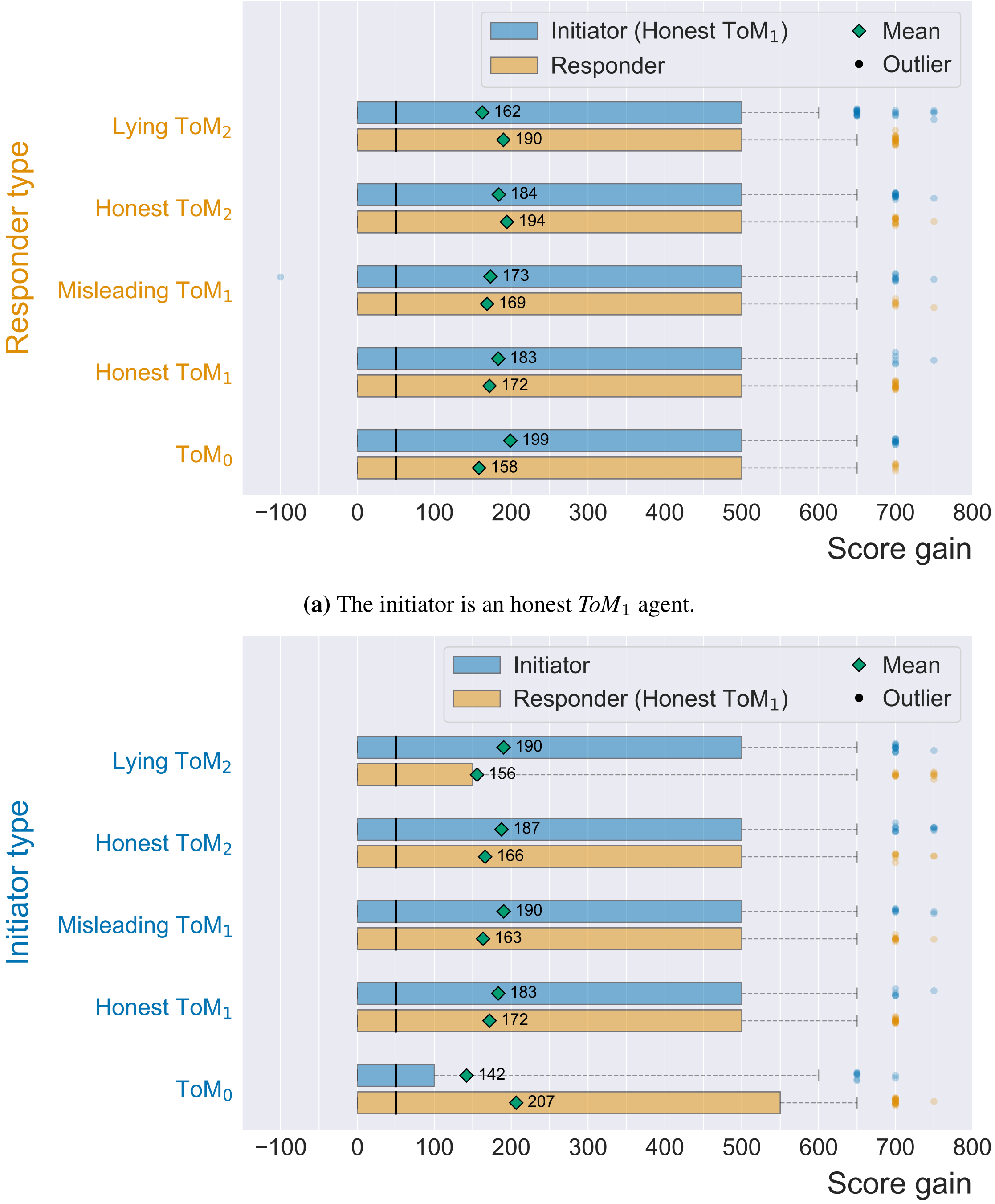
Honest *ToM*$_1$ agent. Figure 4.15 shows the results when we fix the initiator or responder to being an honest *ToM*$_1$ agent. An honest *ToM*$_1$ initiator performs best against a *ToM*$_0$ agent and worst against a lying *ToM*$_2$ agent. This difference is even more pronounced when the honest *ToM*$_1$ agent is the responder. Interestingly, the honest *ToM*$_1$ agent obtains a higher mean score gain against an honest *ToM*$_1$ agent compared to the mean score gain obtained against a misleading *ToM*$_1$ agent, both as initiator and responder. Similarly, the honest *ToM*$_1$ agent obtains a higher mean score gain against an honest *ToM*$_2$ agent compared to the mean score gain obtained against a lying *ToM*$_2$ agent, again both as initiator and responder.

Comparing an honest *ToM*$_1$ agent with a misleading *ToM*$_1$ agent, both set against an honest *ToM*$_1$ trading partner, we observe only slight (insignificant) differences in the mean score gain. Moreover, comparing an honest *ToM*$_2$ agent with a lying *ToM*$_2$ agent, both set against an honest *ToM*$_1$ trading partner, we observe again only slight (insignificant) differences in the mean score gain. Hence, while the honest *ToM*$_1$ agent performs slightly better against honest agents than against misleading/lying agents, its honest trading partner seems not to benefit from the negotiation in terms of score gain; however, there is no statistical evidence supporting these differences.

Misleading *ToM*$_1$ agent. Figure 4.16 shows the results when we fix the initiator or responder to being a misleading *ToM*$_1$ agent. Both as initiator and responder, the misleading *ToM*$_1$ agent performs best against a *ToM*$_0$ agent as a trading partner. A misleading *ToM*$_1$ initiator performs worst against a misleading *ToM*$_1$ agent as a trading partner, while a misleading *ToM*$_1$ responder performs worst against an honest *ToM*$_2$ agent as a trading partner. Except against a *ToM*$_0$ agent as a trading partner, the misleading *ToM*$_1$ agent performs worse as a responder than as an initiator against each agent type in terms of mean score gain.

Comparing an honest *ToM*$_1$ agent with a misleading *ToM*$_1$ agent, both set against a misleading *ToM*$_1$ trading partner, we observe only minor (insignificant) differences in the mean score gain. Moreover, comparing an honest *ToM*$_2$ agent with a lying *ToM*$_2$ agent, both set against a misleading *ToM*$_1$ trading partner, we observe again only slight (insignificant) differences in the mean score gain.

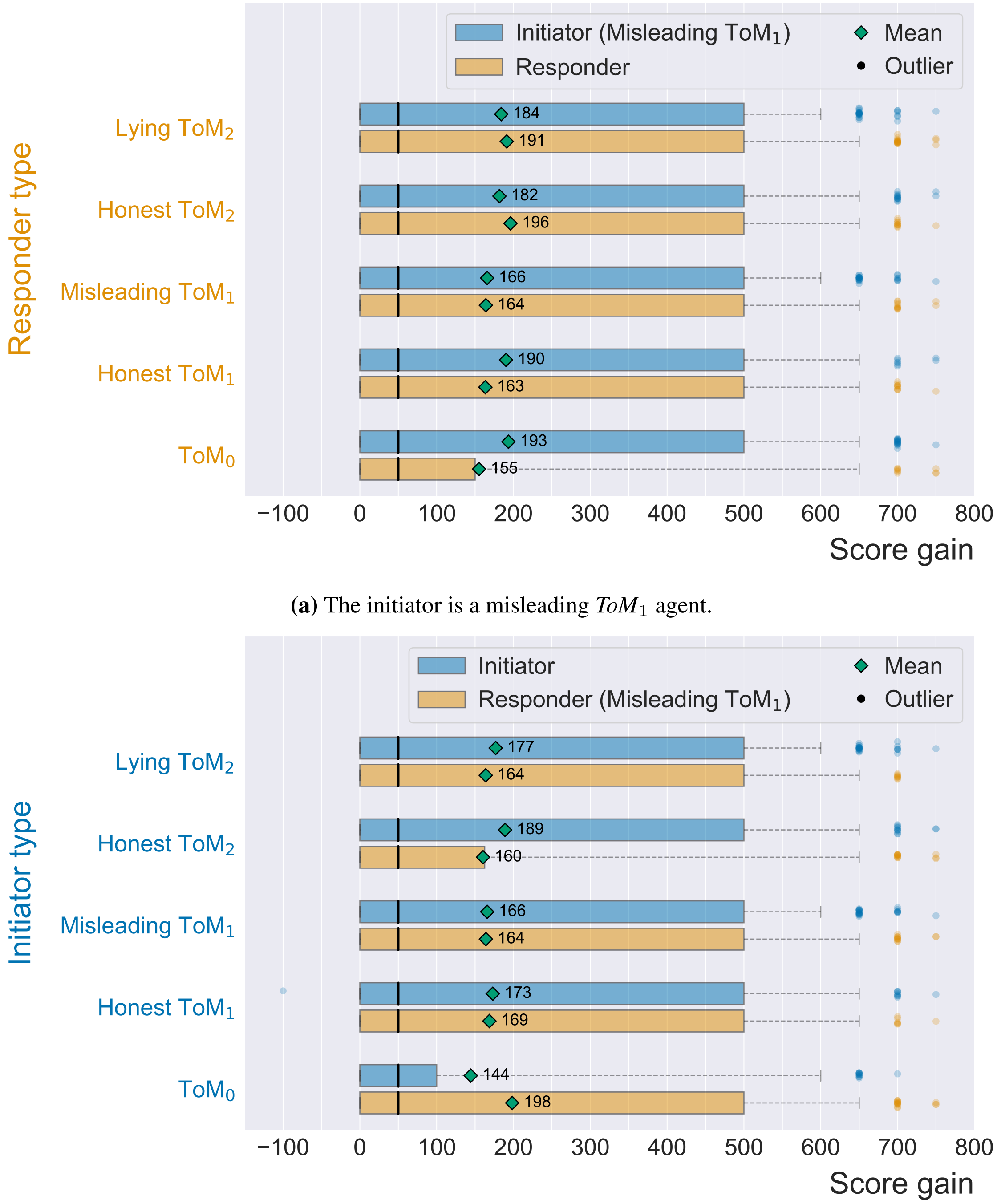
When comparing the results of an honest *ToM*$_1$ agent (see Figure 4.15) with the results of a misleading *ToM*$_1$ agent, we observe no significant differences. The results might suggest that a misleading *ToM*$_1$ agent (M=184.05, SD=247.99) performs better than an honest *ToM*$_1$ agent (M=162.35, SD=236.90), both set against a lying *ToM*$_2$ responder,



(a) The initiator is an honest *ToM*$_1$ agent.

(b) The responder is an honest *ToM*$_1$ agent.

Figure 4.15: Box plots of the score gain for the five different agents where the initiator (a) or the responder (b) is an honest *ToM*$_1$ agent. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.



(a) The initiator is a misleading ToM_1 agent.

(b) The responder is a misleading ToM_1 agent.

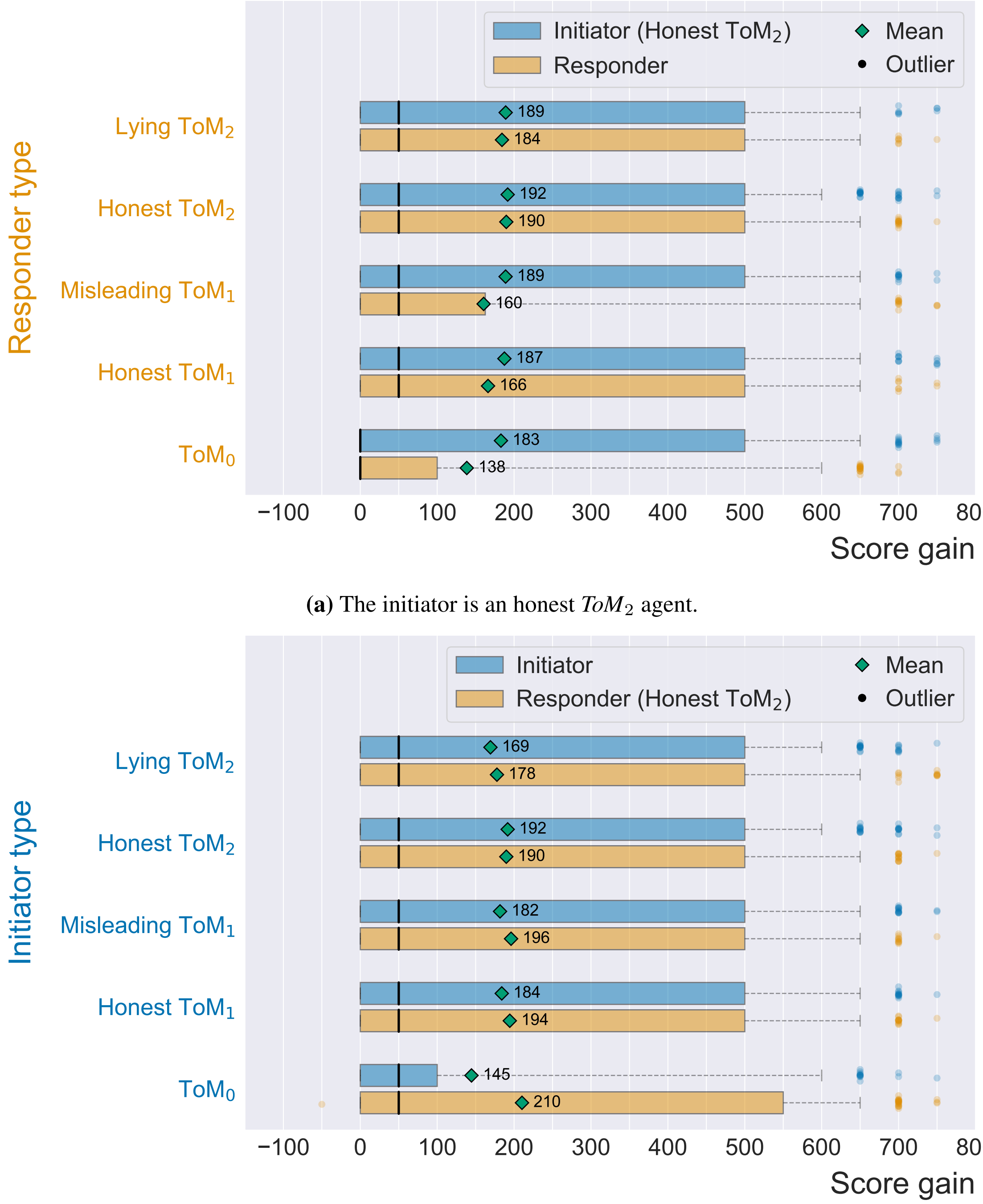
Figure 4.16: Box plots of the score gain for the five different agents where the initiator (a) or the responder (b) is a misleading ToM_1 agent. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

t(1993.8)=2.00, p=.046. However, in contrast, there is no significant difference between the mean score gains of an honest ToM_1 agent (M=155.6, SD=236.93) and a misleading ToM_1 agent (M=163.85, 240.05), both set against a lying ToM_2 initiator, t(1997.7)=0.77, p=.44.

Honest ToM_2 agent. Figure 4.17 shows the results when we fix the initiator or responder to being an honest ToM_2 agent. Interestingly, an honest ToM_2 initiator performs best set against an honest ToM_2 trading partner. However, an honest ToM_2 agent responder performs best set against a ToM_0 agent. The ToM_0 agent is on average the worst off against an honest ToM_2 agent compared to the other agents. There are no notable differences between honest and lying/misleading agents.

Lying ToM_2 agent. Figure 4.18 shows the results when we fix the initiator or responder to being a lying ToM_2 agent. In both cases where the lying ToM_2 agent is an initiator and responder, the mean score gain is highest against a ToM_0 agent as a trading partner. Compared to other types of agents, a ToM_0 agent performs worst against the lying ToM_2 agent. The lying ToM_2 initiator performs worst against an honest ToM_2 responder, and a lying ToM_2 responder performs worst against a lying ToM_2 initiator.

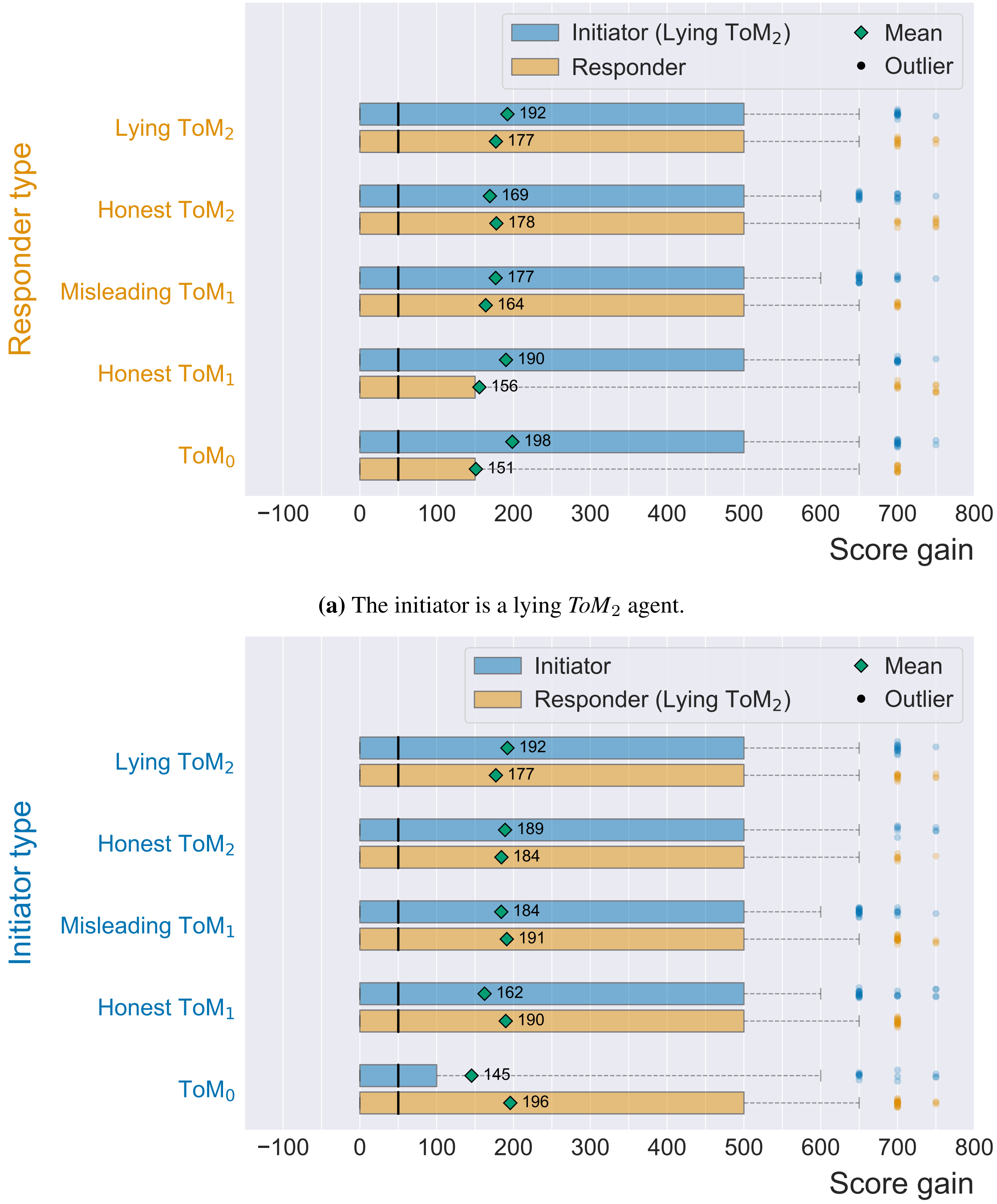
When comparing the results of an honest ToM_2 agent (see Figure 4.17) with the results of a lying ToM_2 agent, we observe no clear differences. In both cases, the ToM_2 agents perform well against a ToM_0 agent, and the ToM_2 responders perform worse against a ToM_2 initiator compared to other types of agents.



(a) The initiator is an honest *ToM*$_2$ agent.

(b) The responder is an honest *ToM*$_2$ agent.

Figure 4.17: Box plots of the score gain for the five different agents where the initiator (a) or the responder (b) is an honest *ToM*$_2$ agent. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.



(a) The initiator is a lying *ToM*$_2$ agent.

(b) The responder is a lying *ToM*$_2$ agent.

Figure 4.18: Box plots of the score gain for the five different agents where the initiator (a) or the responder (b) is a lying *ToM*$_2$ agent. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

4.3 Experiment 3: Is there a benefit to sending goal location messages in Colored Trails?

The results in this section are obtained by conducting the experiment as described in Section 3.5.4, where agents of De Weerd et al. (2017) with a zero-order, first-order, and second-order theory of mind capability negotiate with each other. In total, we obtained results of 9000 negotiations. In 9.1% of the negotiations, the negotiation process was terminated by the initiator before an initial offer was made. In 52.3% of the negotiations, a new distribution of colored chips became final. Compared to Experiment 2, these percentages lie within 2 percent points from each other.

Since each agent negotiated with every other agent in 1000 negotiation rounds, both as initiator and responder, we have 6000 score gains for each agent type. A box plot of the score gain is given in Figure 4.19. The median values are 50 for each of the agent types but the mean values differ. Here, the mean score gain increases with the theory of mind capability of agents.

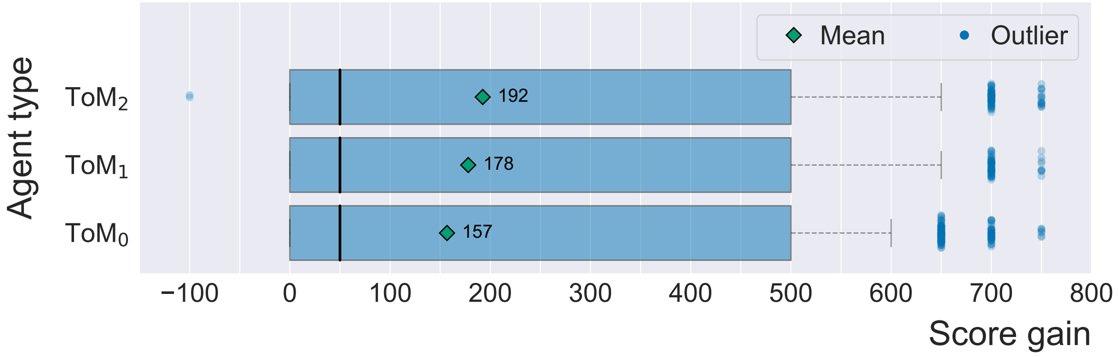


Figure 4.19: Box plot of the score gain for agents of De Weerd et al. (2017) with different theory of mind capabilities (see Section 3.2). The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

A Kruskal-Wallis test was conducted to compare the score gain among the three agents with different levels of theory of mind capability. The Kruskal-Wallis test reveals a significant difference between the groups (H=43.61, df=2, p<.00001), indicating that there are differences in the distribution of the score gain between the groups. A post hoc test was done on the differences in score gain mean values. The results are shown in Figure 4.20, where the difference in means of three combinations of agents are shown together with 98.3% Bonferroni-adjusted confidence intervals (3 comparisons) such that the family-wise error rate is below 0.05. This graph indicates that the differences in mean score gain are significantly different from zero for each combination of agents.

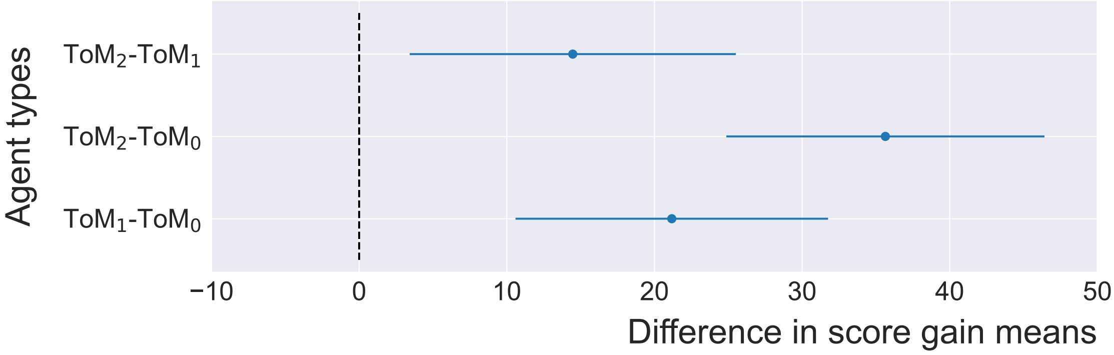


Figure 4.20: Difference in score gain mean values between agents from De Weerd et al. (2017) with different levels of theory of mind. The differences in means contain 98.3% Bonferroni-adjusted confidence intervals (3 comparisons) to ensure a family-wise error rate of less than 0.05. The 98.3% confidence intervals are constructed using a t-statistic. If a confidence interval contains 0, we do not have enough statistical evidence to conclude that the difference is unequal to zero, i.e., that there is a difference in score gain mean values between two different agents.

There are two cases where an agent achieves a negative score gain as can be seen from Figure 4.19. Upon further inspection, both cases concern a *ToM*$_2$ agent who makes an offer that decreases its score and gets accepted by the trading partner. Recall that a reason for a *ToM*$_2$ agent to send an offer that decreases its score is to deceive the trading partner into believing that the actual goal location of the *ToM*$_2$ is not its actual goal location. In return, the *ToM*$_2$ agent expects that the trading partner makes a better offer. However, in these two cases, the offer made by the *ToM*$_2$ agent was accepted by the trading partner, yielding the *ToM*$_2$ agents a negative score gain.

Compared to agents that are capable of sending goal location messages (see Figure 4.6), we observe that agents with an equal theory of mind capability obtain similar mean score gains. Moreover, the distributions seem to be indistinguishable between agents with a similar theory of mind capability, suggesting that the ability to send goal location messages does not benefit the score gain.

Figure 4.21 shows the total number of offers in a negotiation round where a specific agent type is participating. We observe that the median values have shifted to 3 instead of 2 when comparing the agents from De Weerd et al. (2017) with our agents that can send goal location messages (compare Figure 4.21 with Figure 4.10, respectively). In addition, while the mean value for a *ToM*$_0$ agent that is capable of sending goal location messages is rather similar to a *ToM*$_0$ agent that is not capable of sending goal location messages, a *ToM*$_2$ agent capable of sending goal location messages uses fewer offers, on average, than a *ToM*$_2$ agent that is not able to send goal location messages. These results suggest that agents being able to send goal location messages use fewer offers in a negotiation round, especially for *ToM*$_2$ agents.

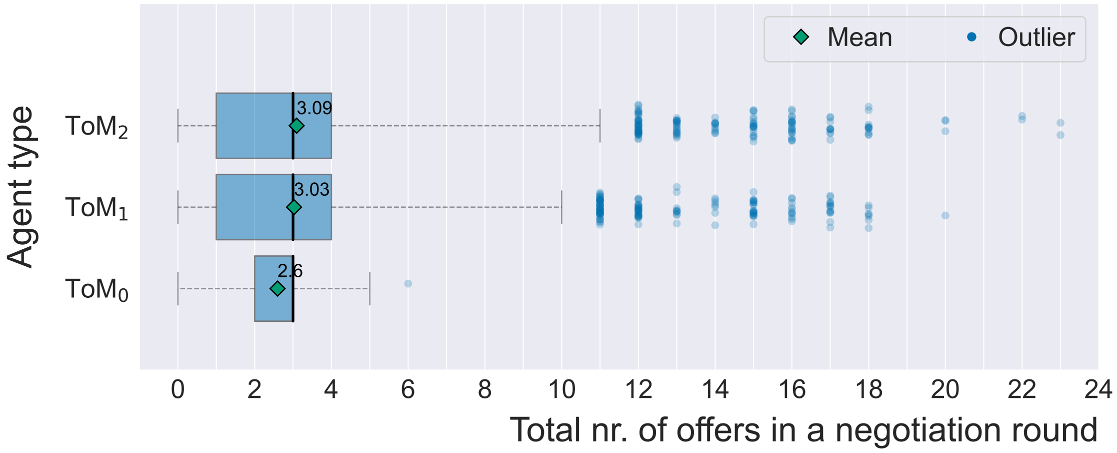


Figure 4.21: A box plot of the total number of offers in a negotiation round for agents of De Weerd et al. (2017). In this plot, data points are gathered where agents capable of different levels of theory of mind negotiate with each other, both as responders and initiators. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

4.4 Experiment 4: Varying the learning speed

The results in this section are obtained by conducting the experiment as described in Section 3.5.5, where the learning speed λ is varied. A box plot of the score gain for different learning speeds is given in Figure 4.22.

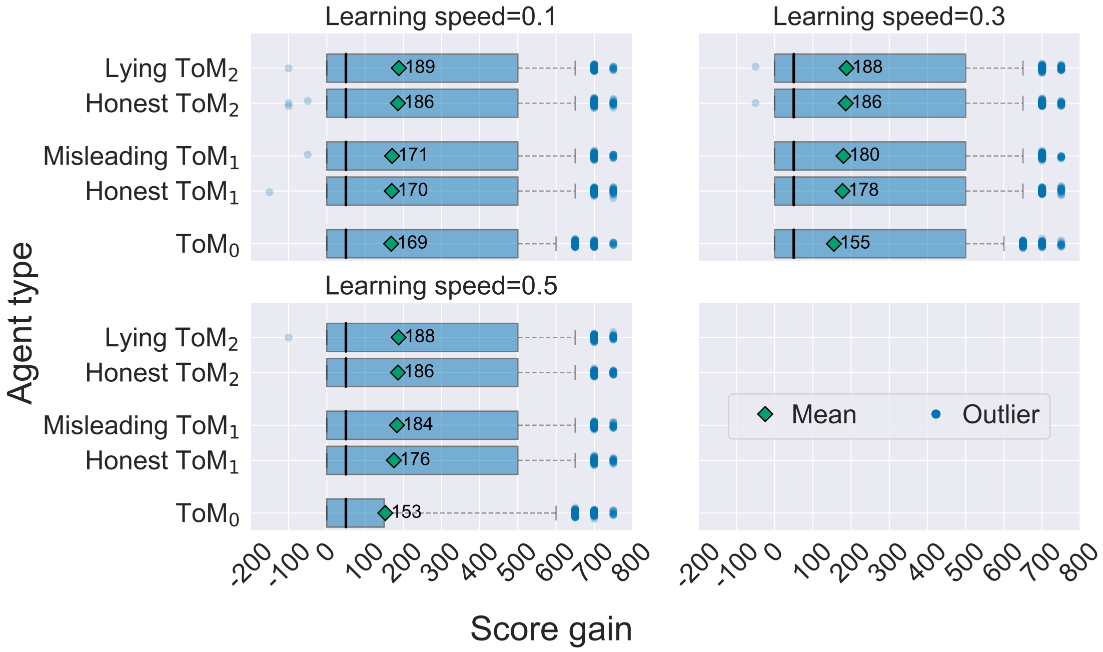


Figure 4.22: Box plot of the score gain for the five different agents (see Section 3.3) for different values for the learning speed λ. Both agents in a negotiation have the same learning speed. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

First of all, Figure 4.22 shows an increase in the mean score gain for a ToM_0 agent. A Kruskal-Wallis test reveals that the learning speed indeed significantly changes the score gain of a ToM_0 agent (H=25.04, df=2, p<.00001). Moreover, a Kruskal-Wallis test suggests that the learning speed changes the score gain of a misleading ToM_1 agent (H=7.87, df=2, p=.020); however, when we use a Bonferroni correction for the 5 comparisons (agent types) we can make, this result is not significant ($p > 0.01$). No significant differences for the other types of agents have been found. These results suggest that a ToM_0 agent benefits from a lower learning speed, which might be in favor of a misleading ToM_1 agent. Another point of interest is that there are more occurrences where an agent obtains a negative score gain when the learning speed is lower.

Figure B.4 shows the results separated by initiators and responders. However, there are no notable differences between the results of the initiators and responders similar to what we found in Experiment 2 (see Section 4.2). Nevertheless, it is noticeable that for lower learning speeds, the responder seems to consistently outperform the initiator for all agent types against the average trading partner.

Figure 4.23 shows the total number of offers in a negotiation round for each agent type. Note that each agent type has negotiated with each agent type, so the total number of offers may also depend on its trading partner. We observe that for lower learning speeds, the distribution of the total number of offers gets a heavier right tail. Moreover, the mean value increases when the learning speed decreases for all agent types. In all three plots, the mean total number of offers in a negotiation round is the smallest for lying and honest ToM_2 agents and the largest for misleading and honest ToM_1 agents.

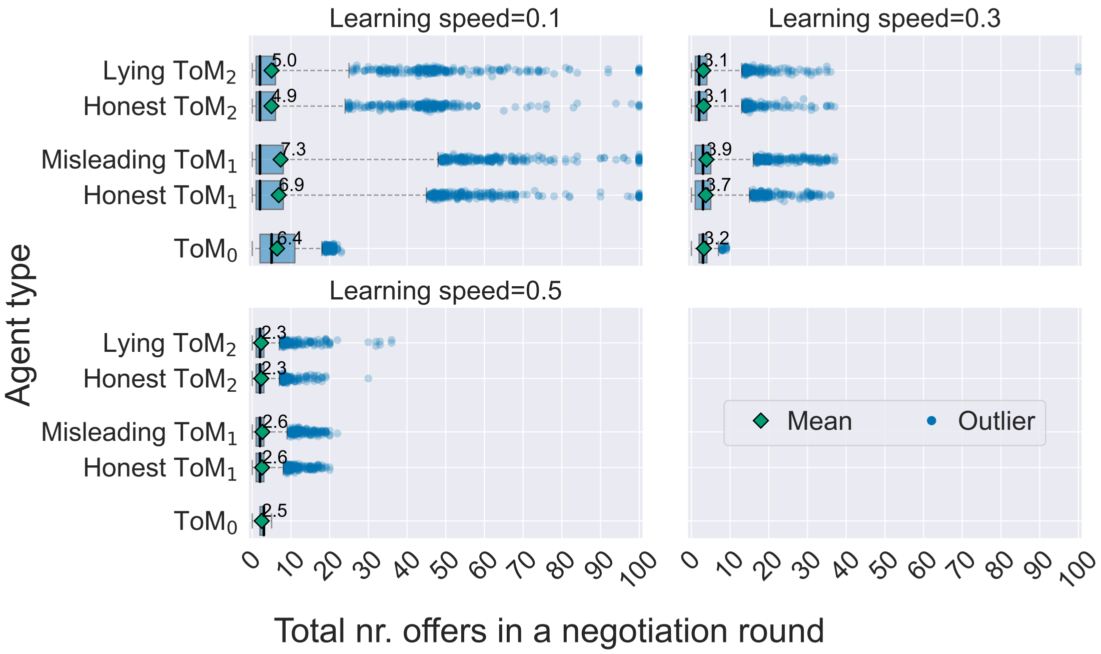


Figure 4.23: A box plot of the total number of offers in a negotiation round for different levels of learning speeds λ. In this plot, data points are gathered where each agent type (see Section 3.3) negotiates with each agent type with a similar learning speed, both as responder and initiator. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

Figure 4.24 shows the fraction of offers that are accompanied by a true goal location message, false goal location message, or no goal location message. For lower values of the learning speed, we observe that fewer goal location messages are sent. An exception is the ToM_0 agent, whose probability p_0 of sending a goal location message together with an offer is fixed to 0.2.

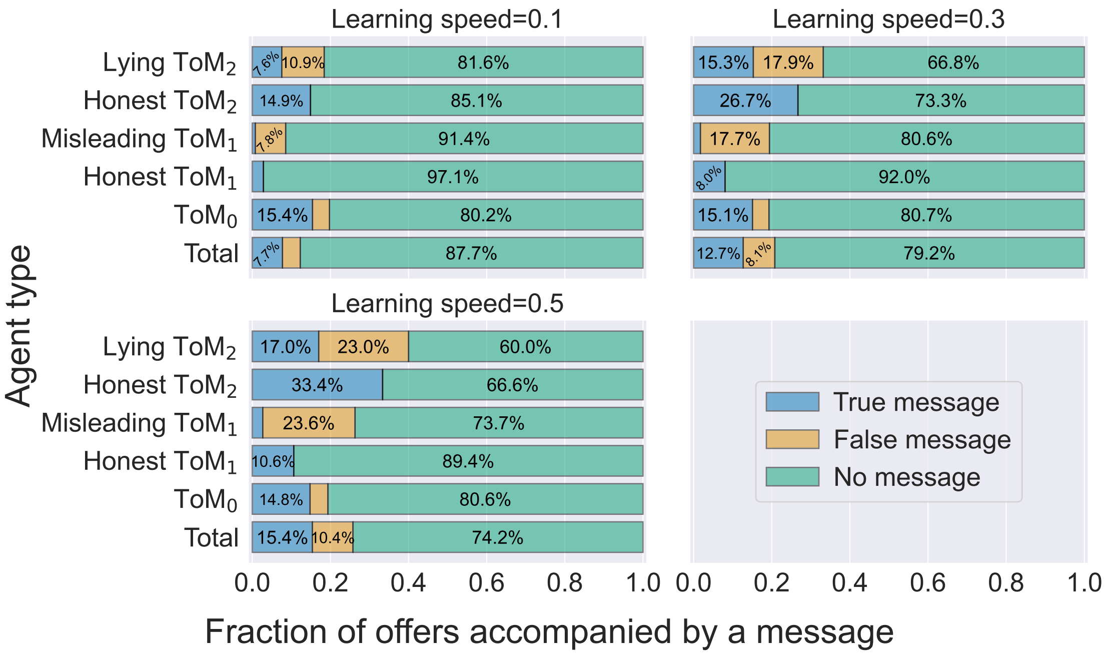


Figure 4.24: A stacked bar plot of the fraction of offers that are accompanied by a true goal location message, a false goal location message, or no goal location message for different learning speeds λ. Percentages are given for reference provided that they are greater than 5%. The *Total* bar indicates the weighted fraction of offers over all five types of agents.

4.5 Experiment 5: Games where a Pareto improvement exists

In Experiment 2, we found that in 77.4% of the negotiations, a Pareto improvement is possible from the initial situation, that is, there exists a distribution of colored chips where at least one agent is better off than the initial situation without hurting the trading partner in terms of the score gain. In only 55.1% of the negotiations of Experiment 2, there was a strict Pareto improvement, which means that there is a distribution where both agents can simultaneously be better off than the initial situation in terms of the total score. When we consider only the negotiations of Experiment 2 where a Pareto improvement was possible, the percentage of negotiations with a new final distribution increases from 53.5% to 69.1%. If we consider only the negotiations of Experiment 2 with a strict Pareto improvement, the percentage of negotiations with a new final distribution is 96.8%. Because in all negotiations of Experiment 2 where a new distribution was accepted a Pareto improvement was possible from the initial situation, the results in this section are obtained by conducting the experiment as described in Section 3.5.6, where only games with a possible Pareto improvement are considered. In total, we have results of 25,000 negotiations.

In 4.5% of the negotiations, the negotiation process was terminated by the initiator before an initial offer was made. In Experiment 2, we found that 11.0% of the negotiations were terminated by the initiator before an initial offer was made. The percentage found in Experiment 5 is thus 6.5 percent point lower than found in Experiment 2, indicating that the initiator starts a negotiation process more often. In 68.8% of the negotiations, a new distribution of colored chips became final. This percentage is 15.3 percent point higher compared to Experiment 2, where we found a new distribution of colored chips became final in only 53.5% of the negotiations.

A box plot of the score gain per agent type for Experiment 5 is given in Figure 4.25. The median values are 50 for each of the agent types but the mean values differ.

A Kruskal-Wallis test was conducted to compare the score gain among the five agents. The Kruskal-Wallis test reveals a significant difference between the groups (H=218.56, df=4, p<.00001), indicating that there are differences in the distribution of the score gain between the groups. A post hoc test was done on the differences in means. The results are shown in Figure 4.26, where the difference in means of ten combinations of agents are shown together with 99.5% Bonferroni-adjusted confidence intervals (10 comparisons) such that the family-wise error rate is below 0.05. The graph indicates that the differences in mean score gain are significantly different from zero for almost all combinations of

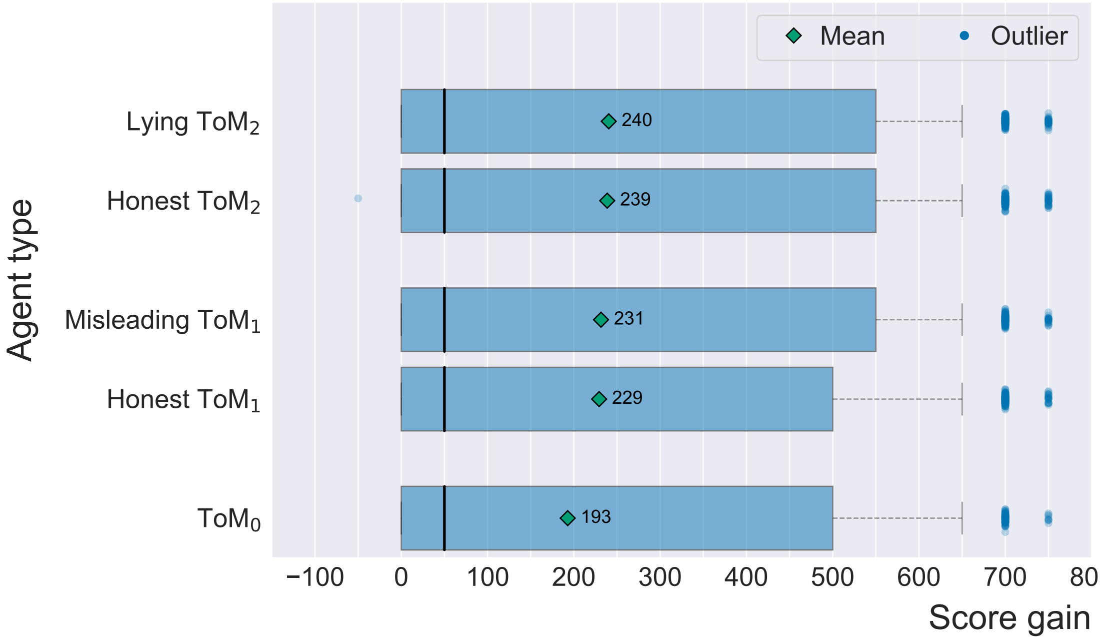


Figure 4.25: A box plot of the score gain for the five different types of agents (see Section 3.3). The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

agents except for combinations where the agents have a similar theory of mind capability or the combination of an honest ToM_2 with a misleading ToM_1 agent.

Compared to Experiment 2 (see Figure 4.7), we now have no statistical evidence of the difference in score gain means between an honest ToM_2 agent and a misleading ToM_1 to be unequal to zero. Hence, when we only consider games where there is a possible Pareto improvement, the difference in performance between an honest ToM_2 agent and a misleading ToM_1 agent is not significant anymore.

Even though we only consider games with a possible Pareto improvement, there is a case where a ToM_2 agent obtains a negative score gain. There thus still exists cases where the expected value of making an offer that yields the agent a negative score gain is the highest among all other offers.

Figure 4.27 shows the experiment results separated by initiators and responders. Compared to Experiment 2 (see Figure 4.8), the mean score gain is higher for each agent type, both as initiator and responder.

In Figure 4.28, we plotted the differences in means with 99% Bonferroni-adjusted confidence intervals (5 comparisons) to control that the family-wise error rate is lower than 0.05. Compared to the results in Experiment 2, the differences between the initiator

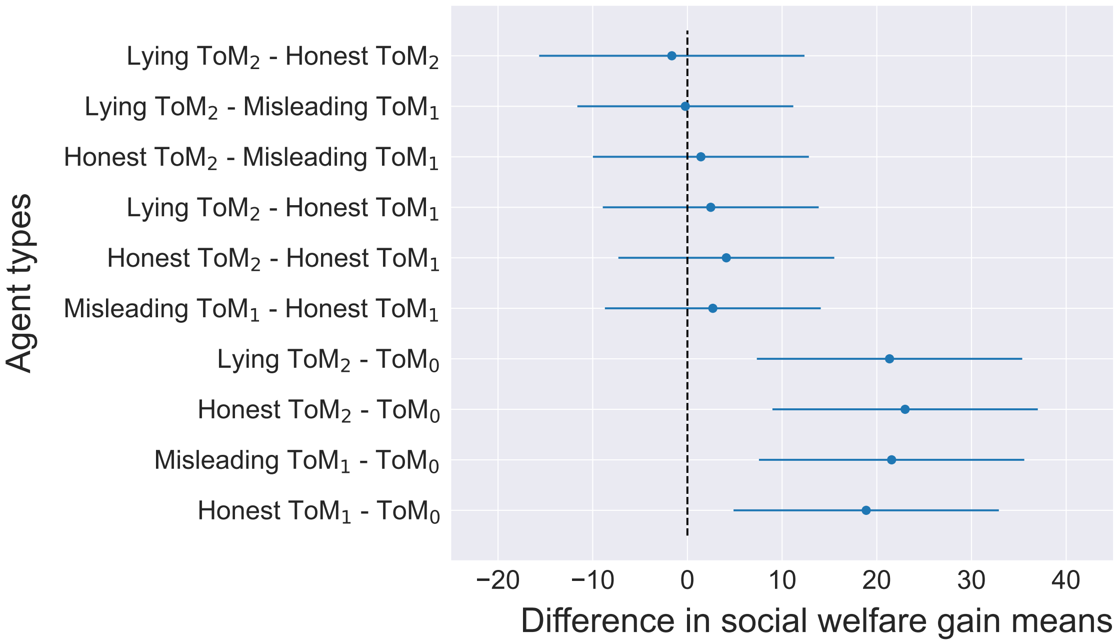


Figure 4.26: Difference in score gain mean values between the five different types of agents (see Section 3.3). The differences in means contain 99.5% Bonferroni-adjusted confidence intervals (10 comparisons) to ensure a family-wise error rate of less than 0.05. The 99.5% confidence intervals are constructed using a t-statistic. If a confidence interval contains 0, we do not have enough statistical evidence to conclude that the difference is unequal to zero, i.e., that there is a difference in score gain mean values between two different agents.

and responder remain insignificant except for the *ToM*$_0$ agent. In games where there is a possible Pareto improvement, the *ToM*$_0$ agent is significantly better off being the initiator than the responder against the average trading partner.

Figure 4.29 shows the distribution of the total number of offers in a negotiation round when a specific agent type takes part in the negotiation. In comparison with Experiment 2 (see Figure 4.10), there are some outliers with a total number of offers in a negotiation round greater than 20. Interestingly, most of these outliers are created by lying and honest *ToM*$_2$ agents. The mean values are also higher for each agent type. Furthermore, the mean total number of offers in a negotiation round is now higher for lying and honest *ToM*$_2$ agents compared to *ToM*$_0$ agents, which was not the case in the previous experiments. This may be partly due to the extreme outliers for lying and honest *ToM*$_2$ agents.

Figure 4.30 shows the fraction of offers that are accompanied by a true goal location message, a false goal location message, or no goal location message. In general, the number of offers without a goal location message increases when we only consider games with a possible Pareto improvement.

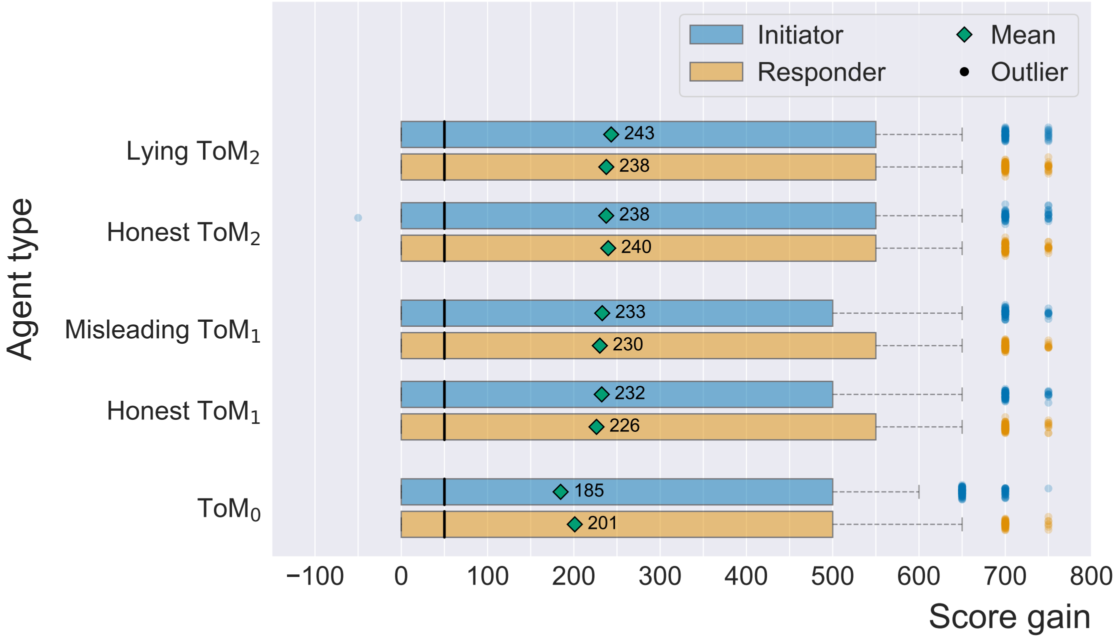


Figure 4.27: A box plot of the score gain for the five different types of agents (see Section 3.3) separated into initiators and responders. The lower and upper whiskers of the box plot reach the bottom 2.5% and 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

Finally, Figure 4.31 shows a box plot of the social welfare gain of every negotiation an agent type participates in where only games with a possible Pareto improvement are considered. Compared to Experiment 2 (see Figure 4.13), the social welfare gain has increased. While there might have been a difference between the social welfare gain of ToM_1 and ToM_2 agents, this difference is not pronounced in Figure 4.31. Interestingly, the median values have shifted from 100 to 550 for each of the agent types.

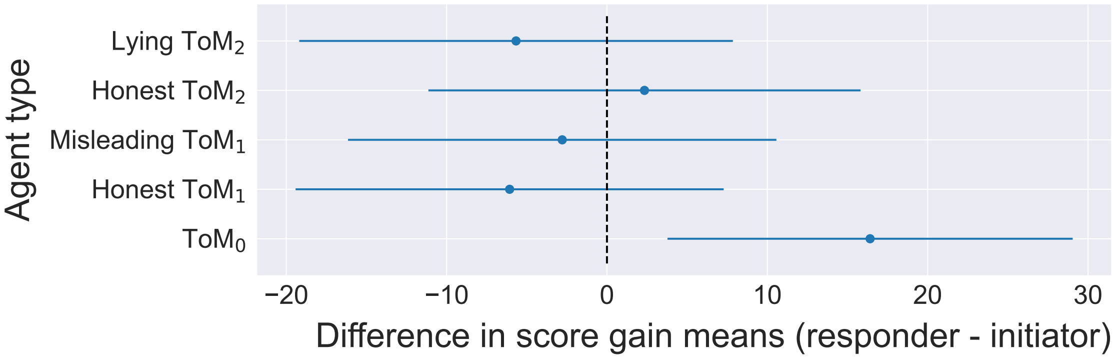


Figure 4.28: Difference in score gain mean values between the responder and initiator for different types of agents (see Section 3.3) with 99% Bonferroni-adjusted confidence intervals (5 comparisons) to ensure a family-wise error rate of less than 0.05. The confidence intervals are constructed using a t-statistic. If a confidence interval contains 0, we do not have enough statistical evidence to conclude that the difference is unequal to zero, i.e., that there is a difference in score gain mean values between the responder and the initiator.

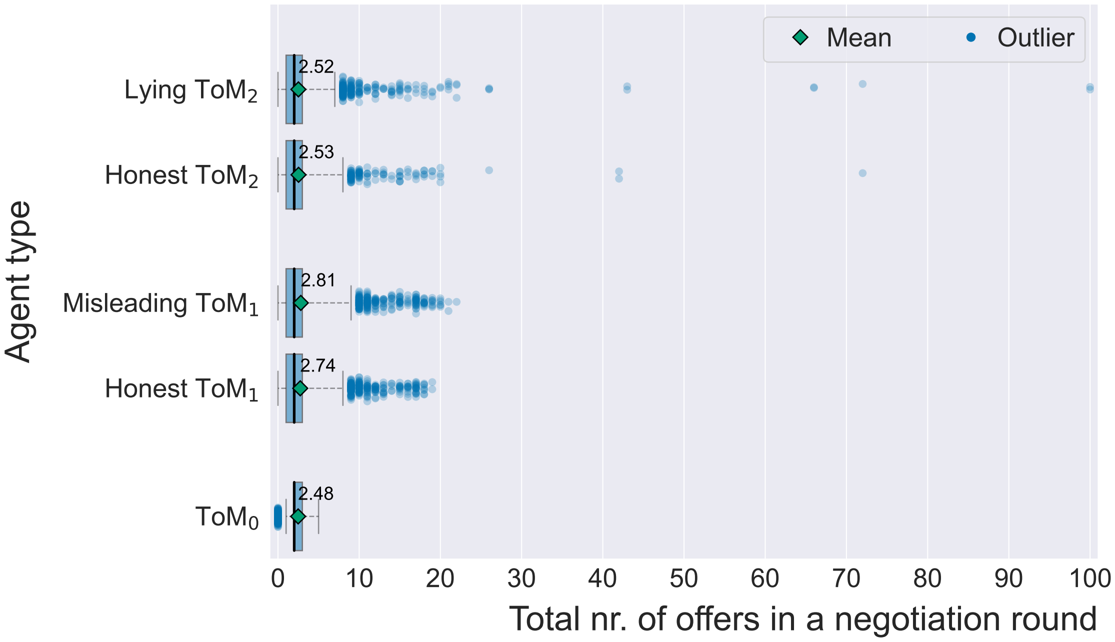


Figure 4.29: A box plot of the total number of offers in a negotiation round. In this plot, data points are gathered where each agent type negotiates with each agent type, both as responder and initiator. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

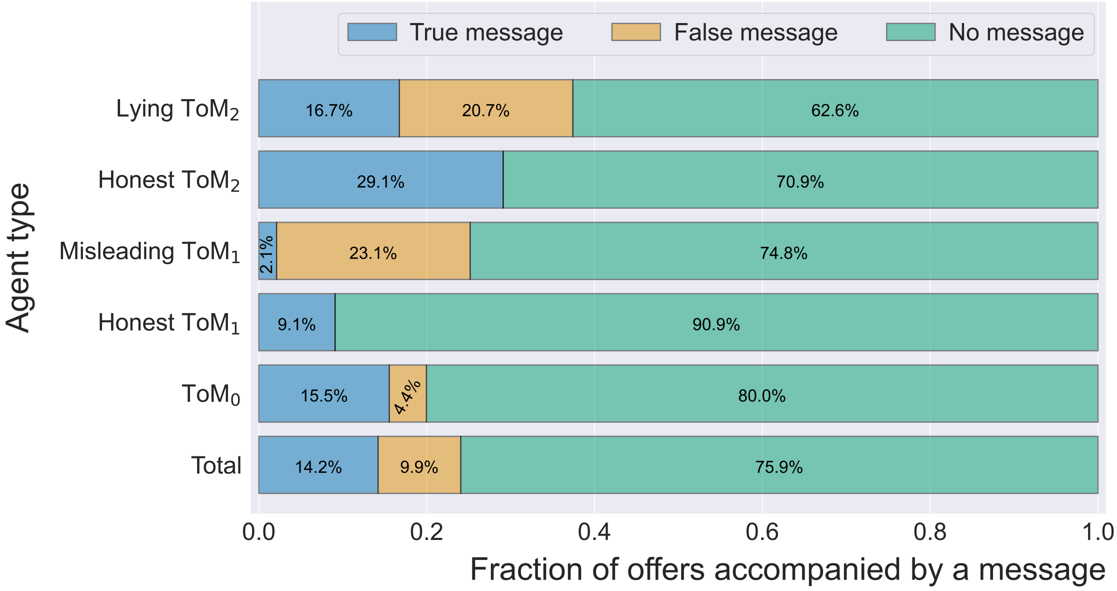


Figure 4.30: A stacked bar plot of the fraction of offers that are accompanied by a true goal location message, a false goal location message, or no goal location message. The *Total* bar is the weighted fraction of offers over all five types of agents.

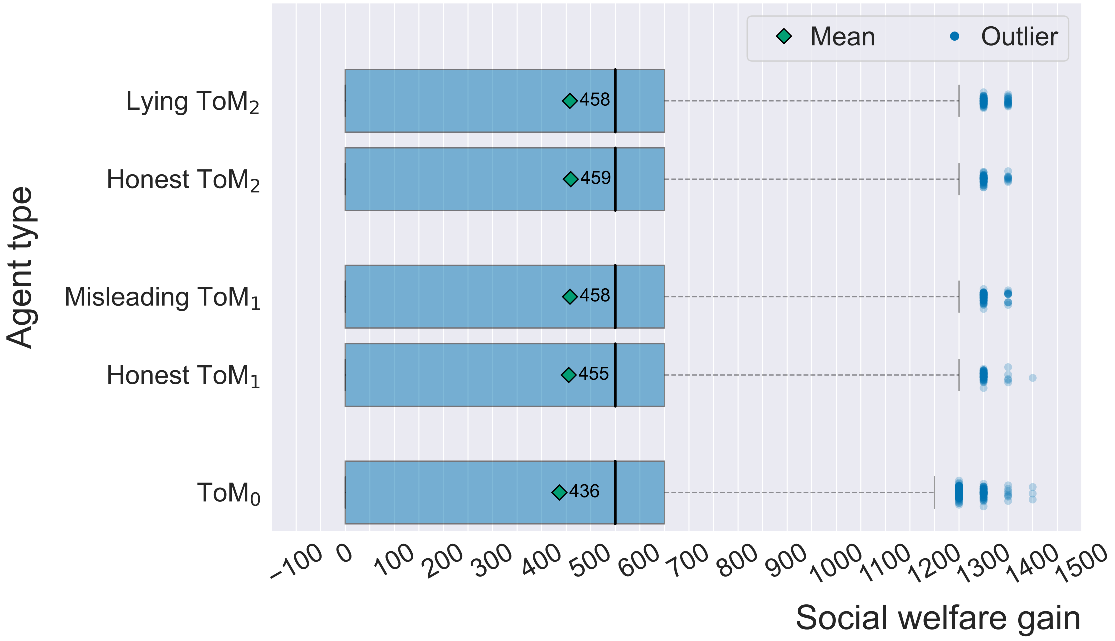


Figure 4.31: Box plot of the social welfare gain for the five different types of agents. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

5 Discussion & Conclusion

The aim of this thesis was to contribute to the existing research on lying and deceiving by AI systems in order to contribute to trustworthy AI. More specifically, the goal was to examine the influence of lying in the negotiation setting of Colored Trails by answering the following research question:

> *What is the influence of lying by artificial agents in the multi-agent negotiation setting of Colored Trails?*

The Colored Trails game, introduced by Grosz, Kraus, and colleagues (2004; 2010) is a framework that is commonly used to investigate decision making in mixed-motive situations, that is, situations where the agents have conflicting motives to cooperate or to compete with each other. In the Colored Trails game, agents negotiate over the distribution of the available colored chips with which they aim to get the highest number of points possible.

By adapting the definition of lying by Van Ditmarsch et al. (2020) to the context of Colored Trails, we defined lying to be an agent that makes a statement that the agent believes to be false with the intent that its trading partner believes that the statement is true and that its trading partner believes that the lying agent also believes said statement is true. First, in order for an agent to lie according to this definition, it needs a theory of mind capability, i.e., the ability to attribute mental content such as beliefs, desires, and goals to another individual (Premack & Woodruff, 1978). To provide our agents with a theory of mind capability, we used the agents of De Weerd et al. (2017). We used their notation where a $\textit{ToM}_k$ agent means that the agent is capable of theory of mind up to the kth-order but not beyond. Second, in order for an agent to lie in Colored Trails, an agent needs to be able to send false messages. We determined that this was not possible in the original setting of Colored Trails by only making offers, so we introduced agents that were able to send a goal location message, that is, a message that states to the receiver that the sender has the goal location mentioned.

While we adopt the definition of lying as given by Van Ditmarsch et al. (2020), various other definitions or ideas of lying exist. In this thesis, we distinguish between the notions of deceiving (Definition 2), lying (Definition 3), and misleading (Definition 4). In our definitions, for each of these concepts, a theory of mind capability is needed. While agents capable of only a first-order theory of mind can mislead and not deceive or lie, agents capable of a second-order theory of mind can use all three concepts. Hence, we constructed five different agents, all capable of sending and receiving goal location messages: a $\textit{ToM}_0$ agent, an honest $\textit{ToM}_1$ agent, a misleading $\textit{ToM}_1$ agent, an honest $\textit{ToM}_2$ agent, and a lying $\textit{ToM}_2$ agent. Using these agents, we contributed to the existing literature in multiple ways.

We first made a graphical user interface (GUI) where a user can observe two agents negotiate in Colored Trails. This GUI is used to examine the behavior of the agents in Colored Trails and to provide examples of agents lying or misleading. The user has various options to choose the negotiation setting. Although the user cannot participate in the negotiation itself, the GUI can be extended such that humans can play against agents.

As a second contribution to this thesis, we performed several experiments to determine the effect of lying in the negotiation setting of Colored Trails. Here, we tested the core hypothesis that agents that are capable of lying or misleading outperform similar agents that are not capable of lying or misleading in terms of score gain. However, we found that this is not the case. Although agents that are capable of lying or misleading have an extra instrument, i.e., they can send goal location messages that contain a false goal location in addition to the capabilities of their honest counterparts, they seem not to benefit from this extra toolkit.

An explanation for the lack of improved performance of lying agents compared to honest agents may be the uncertain environment in which the agents negotiate. Agents do not know the goal location of the trading partner and do not know the agent type of the trading partner. Other studies have suggested that theory of mind allowed us to survive and deal with more complex and unpredictable environments (De Weerd et al., 2017, 2022). Moreover, higher-order theory of mind (that includes second-order theory of mind) has been associated with better negotiation skills (De Weerd et al., 2017). These results are also pronounced in our experiment results and might explain why theory of mind influences the results to a greater extent than the ability to lie or mislead.

Although agents with a theory of mind capability form beliefs about the goal location of the trading partner and about the theory of mind capability of the trading partner, the first offer must be made with absolute uncertainty about the goal location of the trading partner. Hence, it is interesting to observe that ToM_2 responders send more goal location messages than ToM_2 initiators. ToM_2 responders are likely to have obtained some information about the trading partner because they received an offer from the initiator. Consequently, ToM_2 responders can use information deduced from this offer to lie to the trading partner such that the expected value for the ToM_2 agent of a counteroffer increases. Note, that making the environment less uncertain by making the goal locations known to both agents is also no option because in that case, agents cannot lie about their goal location.

While we found that lying is not beneficial for artificial agents in our negotiation setting, there is no consensus among researchers on whether lying is beneficial for humans in negotiations. Howard Raiffa's work (1982; 2002) primarily focused on negotiation analysis and decision making, and he acknowledged that parties may choose to misrepresent information to gain a negotiating advantage. Humans may choose to use various kinds of lies, deception, or misrepresentations of information in negotiations. One can lie about a subject matter or price, but one can also make pro-social lies. In this work, we focused on lying on a subject matter, i.e., lying about your goal. Although some (researchers) argue that lying is always wrong (Sherwood, 2022), others argue that not all forms of lying are selfish (Levine & Schweitzer, 2015). Lying is a short-term strategy, but it can have an impact on longer relationships; hence, negotiators should consider the long-term consequences of using lying and deception. Whether or not lying is beneficial for humans, it is clear that it is used in negotiations intentionally.

Comparing our agents with the agents of De Weerd et al. (2017) suggested that the ability to send goal location messages does not benefit the score gain. In general, the distributions of the score gain seem to be indistinguishable between agents with a similar theory of mind capability. However, the negotiation length, i.e., the total number of offers in a negotiation, is reduced by the ability of agents to send goal location messages. While the number of new distributions of colored chips that became final is about equal, the length of the negotiation reduces when agents can send goal location messages. These results indicate that the negotiation becomes more efficient without decreasing performance when agents are given the ability to send goal location messages.

Furthermore, we found that the learning speed, an agent-specific parameter that represents the degree to which new information influences the beliefs of the agent, influences some of the results significantly. In particular, we observed that the mean score gain significantly increases for a *ToM*$_0$ agent when the learning speed decreases. Moreover, the total number of offers in a negotiation round increases when the learning speed decreases. These two results are evident since more offers are needed to change the beliefs of the agents. This also explains why there were more occurrences where an agent obtained a negative score gain when the learning speed is lower; namely, an agent might expect the trading partner to make a counteroffer instead of accepting the offer.

The game of Colored Trails has many initial configurations. These configurations influence the results of the experiments. Following De Weerd et al. (2017) in our experiments, we only considered games where neither agent could initially reach their goal location, in order to ensure that both agents have the incentive to negotiate to increase their score. However, even with these restrictions, we found that in 22.6% of the games, neither agent could improve their score without hurting the trading partner, that is, there was no Pareto improvement possible. Moreover, we found large variances in the score gains. In our final experiment, we considered only initial settings in which a Pareto improvement was possible. Considering only these games, we observed more games with a new distribution of colored chips. Consequently, the mean score gain increased significantly for each of the agent types. However, the variance of the score gain remained about the same. These results suggest that due to the large number of settings of Colored Trails, many negotiation results should be obtained to draw conclusions. Alternatively, one can investigate only certain settings so that the results are less susceptible to chance. While it might be worthwhile to investigate the influence of lying in certain settings only, this could jeopardize generalizability of the results.

It should be noted that the interpretation of goal location messages by *ToM*$_0$ agents is rather limited. When a *ToM*$_0$ agent receives a goal location message, it considers offers that do not contain the color of the mentioned goal location less likely to be accepted. This might have resulted in the misleading *ToM*$_1$ agent sending false goal location messages in many cases to increase the likelihood of its offer going to be accepted or to increase the likelihood of a better counteroffer in terms of score gain. A *ToM*$_1$ agent might simply aim to get a chip color of another goal location by its misleading message. While a *ToM*$_0$ agent cannot model the trading partner having a goal, another interpretation of a goal location message by a $\textit{ToM}_0$ agent would be using the length of the path to the mentioned goal location in determining offers that are less likely to be accepted.

Finally, due to the agents being able to adopt lower orders of theory of mind, it was hard to tell whether an agent intended to lie, mislead, or simply send a false goal location. Another idea was to fix the order of theory of mind, but then there were still problems with second-order theory of mind agents intending to change the beliefs of the trading partner but based on intentions that are not considered lying. For example, a lying agent might send a second goal location message that contradicts the previously sent goal location message in order to change the beliefs of the trading partner. Hence, it is important to keep in mind that what is considered lying in the results, might not actually be a lie.

5.1 Conclusion

In the year 2000, Castelfranchi (2000) predicted that there will be problems of deception and trust between humans and artificial entities and among artificial agents themselves. Presently, researchers found that people ascribe intentions of lying and deception to robots and agents (Kneer, 2021; Rogers et al., 2023). With the rising use of artificial intelligence in advanced systems that include automated negotiations, we need transparency of the AI system and examine when and in what ways an AI system is “willing” to lie. When an AI system becomes sufficiently smart, nothing prevents it from lying. We might be able to mitigate unwanted deception by AI agents when we understand all the possible ways an AI agent can lie or deceive.

In this thesis, we found that the effect of lying by artificial agents in the multi-agent negotiation setting of Colored Trails is insignificant in terms of score gain compared to an honest agent, in general. Against the average trading partner, the results seem to be largely influenced by the theory of mind capability instead of the capability to lie. Since, in many other applications, there are consequences to (being caught) lying for an agent in possibly next encounters with the trading partner, the results may suggest that honesty is the best policy. These results are hopeful in the sense that, while agents are becoming increasingly smart, there are no benefits for them to lie in (mixed-motive) negotiation settings.

5.2 Future work

Various assumptions and considerations have led to the thesis as it is. Therefore, several improvements and suggestions for future directions can be derived from this thesis. We have split them into improvements to the settings of Colored Trails and to the lying of agents. We close this discussion with future research directions for applications of this work.

Colored Trails. The Colored Trails setting provides many possible initial situations that influence the negotiation process. Because the initial configuration (e.g., whether there exists a possible Pareto improvement from the initial situation) impacts the negotiation process, the number of data points that we considered (1000) might be insufficient to capture whether there are differences in the average performance of agents. Additionally, this has been noticed from the high standard deviations of the results leading possibly to many insignificant results. A reason for this high standard deviation is the odd score gain distribution, which is discrete with multiples of 50, highly skewed, and contains grossly two groups of points. While we employed the score calculation from De Weerd et al. (2017) and others, it might be interesting to look at the influence of the score calculation on the performance of the agents (see, e.g., de Jong, Hennes, Tuyls, & Gal, 2011, for another score calculation). Similarly, we can compare the performance of agents in the same setting instead of averaging the performance over random games as we did in this thesis.

Another consideration in this thesis was the static positions in which the initiator and the responder negotiated in the warm-up phase. The warm-up phase was meant for ToM_0 agents to learn across games. However, these fixed positions of initiators and responders might yield overfitting of the zero-order beliefs in a particular position, meaning that these zero-order beliefs perform worse in the other position. Although we suspect the influence of this to be insignificant, a possible alteration of the fixed positions of the initiator and the responder is to randomly place the agent in the position of the initiator or responder at the start of each negotiation. However, agents might still adapt to the trading partner type. Hence, an additional idea is to make a pool of agents with different types of agents and let them negotiate with each other through random encounters.

In our Colored Trails game, we fixed the learning speed of the agents to 0.5. Since the learning speeds are similar for both agents, it is possible that an agent models the beliefs of its trading partner correctly. Another idea would be to make the learning speed differ between agents or to make it normally distributed. This difference in learning speed might result in the agents modeling the trading partner incorrectly and obtaining counteroffers that the agent would not expect. Consequently, the negotiation process and results might change (see, e.g., De Weerd et al., 2017, for results on agents with differing levels of theory of mind and varying learning speeds).

Lying. While there are many controversies with the definitions of lying and deceiving, we distinguished between the two definitions in this thesis and added a definition for misleading.

While, in this thesis, agents were given the capability to send goal location messages in order to provide a possibility for lying, other alternatives exist. In Section 3.3.1, we already discussed that an agent in the Colored Trails game could also have been given the ability to send a preference order of offers or to send the number of chips needed to reach its goal location. It might be interesting to investigate these options for communication too.

Lying may be beneficial (see, e.g., Levine & Schweitzer, 2015) or can have undesirable consequences. In this thesis, agents can lie about their goal location without direct consequences for themselves. When the trading partner of a lying/misleading agent stumbles upon a contradiction in its beliefs, it revokes its beliefs and does not believe the lying/misleading agent in subsequent statements in the current negotiation round. While this might influence the current negotiation round, the lying/misleading agent is only implicitly caught to lie/mislead. An interesting addition to the current model is to add an explicit award when the receiver catches a liar by, for example, receiving points for pointing out liars and deducting points for being caught lying.

Furthermore, in this thesis, agents believe the trading partner instantly when a goal location message is sent (under the condition that the message does not contradict its beliefs), even if the agent believes that the probability of that goal location being the actual goal location of the trading partner is extremely small (but nonzero). However, agents do consider the possibility that the goal location mentioned is not the actual goal location of the trading partner by being able to revoke their beliefs. Another option would have been that the trading partner believes the agent only when the believed probability of the mentioned goal location being the actual goal location of the trading partner is highest among the goal location probabilities. Another alternative could be to construct agents with a trust parameter that models the relation of trust and is used to update the goal location beliefs. Namely, if agents are caught lying, trust by the other party may be damaged and could lead to retaliation. Agents might also be more suspicious if they found out that the trading partner can lie about its goal location.

The results showed that ToM_2 agents sent, on average, more goal location messages with their offers than ToM_1 agents. An interesting addition to this thesis would be to let agents reason about why the trading partner did (not) send a particular goal location message. While we exclude this reasoning, an agent could be able to reason about what it would have messaged if it were in the position of the trading partner. Consequently, the agent can adapt its belief in the goal location of the trading partner and its confidence in the theory of mind capability of its trading partner.

Moreover, we found that, for higher learning speeds, many goal location messages are sent with offers. An interesting alternative would be to add a cost (e.g., 1 point) to sending a goal location message. We expect that fewer goal location messages will be sent with an offer.

While we found that there is no general benefit to agents being capable of lying, given the opportunity, agents did lie anyway. A separate study may shed light on more specific situations where it may (not) be optimal for an artificial agent to lie in a negotiation setting. On the one hand, the potential benefits and risks of artificial agents lying in negotiation settings may vary depending on the specific context, and, thus, specific situations that foster lying should be determined. On the other hand, we might be more interested in the benefits of artificial agents in negotiations in general such as whether they can achieve a better outcome for themselves, or whether it is better to build trust with the other party.

Applications. The results in this thesis are gathered using the Colored Trails framework. Colored Trails is a useful research test-bed for investigating the decision making of agents in a negotiation setting. Besides artificial agents, a heterogeneous group of humans and agents can play the Colored Trails game (Kraus et al., 2004). Previous research used this framework to show that theory of mind agents can encourage the use of higher-order theory of mind in human participants (De Weerd, Broers, & Verbrugge, 2015). Our research could be extended by letting humans negotiate with lying agents to train them to expose liars in negotiations. This could help them to be aware of situations that encourage agents to lie, which might help them to even catch people lying in negotiations. Moreover, since we added an extra communication option besides making an offer, this research could be used to train people to recognize situations where it is beneficial to communicate more than simply an offer, or maybe when it might be beneficial to lie (or when not to lie).

We found that lying is not beneficial for artificial agents in the Colored Trails, a mixed-motive setting. De Weerd (2015) found that mixed-motive settings are more likely to be the main contributor to the emergence of higher-order theory of mind in humans than either purely competitive or purely cooperative settings. Lying is seen as a Machiavellian tactic (Christie & Geis, 2013) and may only be effective in strictly competitive settings.

Raveenthiran (2023) analyzed human behavior in human-human and human-agent play in the mixed-motive setting of the Mod-Signal Game. Raveenthiran (2023) found that humans tend to play competitively and be dishonest with their signal despite there being a higher payoff of playing cooperatively. Future research may also look at the influence of lying in strictly competitive settings or strictly cooperative settings, for example, to explain which of these settings contributed to the emergence of lying from an evolutionary perspective as De Weerd (2015) did for the emergence of higher-order theory of mind in humans.

Besides Colored Trails, other frameworks where agents negotiate over the distribution of resources exist. Ebrahimnezhad and Fujita (2023) introduce NegoSim, a new negotiation simulator that includes protocols, negotiating parties, and analytic tools. Ebrahimnezhad and Fujita (2023) claim to have made an appropriate platform for researchers to investigate negotiations. Relating to our research, one can add agents with and without a lying strategy to investigate whether there is no benefit to lying in other negotiation settings too. Another example where agents can be analyzed in a negotiation setting is introduced by Facebook Artificial Intelligence Research (Lewis, Yarats, Dauphin, Parikh, & Batra, 2017). Similar to this thesis, they studied negotiation on a multi-issue bargaining task. In contrast to this thesis, they used reinforcement learning to train their dialog agents to use effective natural language in negotiations. It might be interesting to observe whether these agents are capable of lying and if so, whether it is beneficial to lie in more complex negotiations involving natural language and, possibly, humans.

References

Ågotnes, T., van Ditmarsch, H., & Wang, Y. (2018). True lies. *Synthese*, *195*(10), 4581–4615.

Alam, M., Gerding, E. H., Rogers, A., & Ramchurn, S. D. (2015). A scalable interdependent multi-issue negotiation protocol for energy exchange. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (pp. 1098–1104).

Augustine, S. (1956). De Mendacio. In P. Schaff (Ed.), *A Select Library of the Nicene and Post-Nicene Fathers of the Christian Church* (Vol. 3). (Written by St. Augustine of Hippo around A.D. 395.)

Baarslag, T., Fujita, K., Gerding, E. H., Hindriks, K., Ito, T., Jennings, N. R., … Williams, C. R. (2013). Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence*, *198*, 73–103.

Baarslag, T., Kaisers, M., Gerding, E., Jonker, C. M., & Gratch, J. (2017). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4684–4690).

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., … Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074.

Beall, J., Glanzberg, M., & Ripley, D. (2020). Liar Paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2020/entries/liar-paradox/`.

Bugnyar, T., & Heinrich, B. (2006). Pilfering ravens, Corvus corax, adjust their behaviour to social context and identity of competitors. *Animal Cognition*, *9*, 369–376.

Bugnyar, T., & Kotrschal, K. (2002). Observational learning and the raiding of food caches in ravens, Corvus corax: Is it 'tactical' deception? *Animal Behaviour*, *64*(2), 185–195.

Caminada, M. (2009). Truth, lies and bullshit; Distinguishing classes of dishonesty. In *Proceedings of the Social Simulation Workshop at the International Joint Conference on Artificial Intelligence* (pp. 39–50).

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

Center for Information Technology. (2023). *Hábrók Documentation.* (`https://wiki.hpc.rug.nl/habrok/start`)

Chen, S., & Weiss, G. (2012). An efficient and adaptive approach to negotiation in complex environments. In *Proceedings of the 20th European Conference on Artificial Intelligence* (pp. 228–233).

Chevalier-Skolnikoff, S. (1986). An exploration of the ontogeny of deception in human beings and nonhuman primates. *Deception, Perspectives on Human and Nonhuman Deceit*, 205–220.

Christie, R., & Geis, F. L. (2013). *Studies in Machiavellianism.* Academic Press.

Corporate Finance Institute. (2021). *Pareto Improvement.* Retrieved July 11, 2023, from `https://corporatefinanceinstitute.com/resources/economics/pareto-improvement/`

Dance, F. E. (1970). The "concept" of communication. *Journal of Communication*, *20*(2), 201–210.

de Jong, S., Hennes, D., Tuyls, K., & Gal, Y. K. (2011). Metastrategies in the Colored Trails game. In *The 10th International Conference on Autonomous Agents and Multi-Agent Systems - Volume 2* (pp. 551–558). International Foundation for Autonomous Agents and Multi-Agent Systems.

Dennett, D. C. (1989). *The Intentional Stance.* MIT press.

DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, *74*(1), 63–79.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). Theory of mind: Did evolution fool us? *PloS One*, *9*(2), Article number: e87619.

de Weerd, H. (2015). *If You Know What I Mean: Agent-Based Models for Understanding the Function of Higher-Order Theory of Mind.* (PhD thesis, University of Groningen) de Weerd, H., Broers, E., & Verbrugge, R. (2015). Savvy software agents can encourage the use of second-order theory of mind by negotiators. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)* (pp. 542–547).

de Weerd, H., Verbrugge, R., & Verheij, B. (2013a). Higher-order theory of mind in negotiations under incomplete information. In *International Conference on Principles and Practice of Multi-Agent Systems* (pp. 101–116).

de Weerd, H., Verbrugge, R., & Verheij, B. (2013b). How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, *199*, 67–92.

de Weerd, H., Verbrugge, R., & Verheij, B. (2015). Higher-order theory of mind in the Tacit Communication Game. *Biologically Inspired Cognitive Architectures*, *11*, 10–21.

de Weerd, H., Verbrugge, R., & Verheij, B. (2017). 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.

de Weerd, H., Verbrugge, R., & Verheij, B. (2022). Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, *36*(2), 1–33.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64.

Ebrahimnezhad, A., & Fujita, K. (2023). Negosim: A modular and extendable automated negotiation simulation platform considering euboa. *Applied Sciences*, *13*(1). Retrieved from `https://www.mdpi.com/2076-3417/13/1/642`

El-Hani, C. N., Queiroz, J., & Stjernfelt, F. (2010). Firefly femmes fatales: A case study in the semiotics of deception. *Biosemiotics*, *3*(1), 33–55.

Fallis, D. (2009). What is lying? *The Journal of Philosophy*, *106*(1), 29–56.

Ficici, S. G., & Pfeffer, A. (2008). Modeling how humans reason about others with partial information. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 315–322).

Frankfurt, H. G. (2005). *On Bullshit*. Princeton University Press.

Gal, Y., Grosz, B., Kraus, S., Pfeffer, A., & Shieber, S. (2010). Agent decision-making in open mixed networks. *Artificial Intelligence*, *174*(18), 1460–1480.

Gerbrandy, J., & Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language and Information*, *6*(2), 147–169.

Goldstein, H., & Healy, M. J. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *158*(1), 175–177.

Gray, W. D. (2020). Introduction to volume 12, issue 2 of topiCS. *Topics in Cognitive Science*, *12*(2), 464–465.

Green, S. D., Duarte, R. C., Kellett, E., Alagaratnam, N., & Stevens, M. (2019). Colour change and behavioural choice facilitate chameleon prawn camouflage against different seaweed backgrounds. *Communications Biology*, *2*(1), Article number: 230.

Grice, H. P. (1975). Logic and conversation. In *Speech Acts* (pp. 41–58). Brill.

Hurst, L. (2023). *Rapid growth of new sites using AI tools like ChatGPT is driving the spread of misinformation.* Retrieved July 14, 2023, from `https://www.euronews.com/next/2023/05/02/rapid-growth-of-news-sites-using-ai-tools-like-chatgpt-is-driving-the-spread-of-misinforma`

Hyman, R. (1989). The psychology of deception. *Annual Review of Psychology*, *40*(1), 133–154.

IBM. (n.d.). *Building trust in AI.* Retrieved July 14, 2023, from `https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html`

Isaac, A., & Bridewell, W. (2017). Why robots need to deceive (and how). *Robot Ethics*, *2*, 157–172.

Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Sierra, C., & Wooldridge, M. (2001). Automated negotiation: Prospects, methods and challenges. *International Journal of Group Decision and Negotiation*, *10*(2), 199–215.

Johnson, E., Gratch, J., & DeVault, D. (2017). Towards an autonomous agent that provides automated feedback on students' negotiation skills. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems* (pp. 410–418).

Kneer, M. (2021). Can a robot lie? Exploring the folk concept of lying as applied to artificial agents. *Cognitive Science*, *45*(10), Article number: e13032.

Knol, M. J., Pestman, W. R., & Grobbee, D. E. (2011). The (mis) use of overlap of confidence intervals to assess effect modification. *European Journal of Epidemiology*, *26*, 253–254.

Kramár, J., Eccles, T., Gemp, I., Tacchetti, A., McKee, K. R., Malinowski, M., ... Bachrach, Y. (2022). Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, *13*(1), 1–15.

Kraus, S. (1997). Negotiation and cooperation in multi-agent environments. *Artificial Intelligence*, *94*(1), 79–97.

Kraus, S., Grosz, B. J., Talman, S., Havlin, M., & Stossel, B. (2004). The influence of social dependencies on decision-making: Initial investigations with a new game. In *Autonomous Agents and Multi-Agent Systems, International Joint Conference* (Vol. 3, pp. 782–789). IEEE Computer Society.

Kraus, S., Sycara, K., & Evenchik, A. (1998). Reaching agreements through argumentation: A logical model and implementation. *Artificial Intelligence*, *104*(1), 1–69.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Lavoie, J., & Talwar, V. (2020). Care to share? Children's cognitive skills and concealing responses to a parent. *Topics in Cognitive Science*, *12*(2), 485–503.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88–106.

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). Deal or no deal? End-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Li, W., Logenthiran, T., Phan, V.-T., & Woo, W. L. (2017). Intelligent housing development building management system (HDBMS) for optimized electricity bills. In *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)* (pp. 1–6).

Livet, P., & Varenne, F. (2020). Artificial intelligence: Philosophical and epistemological perspectives. In *A Guided Tour of Artificial Intelligence Research* (pp. 437–455). Springer.

Mahon, J. E. (2016). The definition of lying and deception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2016/entries/lying-definition/`.

Masters, P., Smith, W., Sonenberg, L., & Kirley, M. (2021). Characterising deception in AI: A survey. In S. Sarkadi, B. Wright, P. Masters, & P. McBurney (Eds.), *Deceptive AI. First International Workshop, DeceptECAI 2020, and Second International*

Workshop, DeceptAI 2021. Communications in Computer and Information Science (Vol. 1296, pp. 3–16).

Mehrabian, A. (1971). *Silent Messages*. Wadsworth Belmont, CA.

Meibauer, J. (2019). *The Oxford Handbook of Lying*. Oxford Handbooks.

Merriam-Webster. (n.d.). Lie. In *Merriam-Webster.com dictionary.* Retrieved December 17, 2022, from `https://www.merriam-webster.com/dictionary/lie`

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*(5), 749–773.

Panisson, A. R., Sarkadi, Ş., McBurney, P., Parsons, S., & Bordini, R. H. (2018). Lies, bullshit, and deception in agent-oriented programming languages. In *20th International Trust Workshop (Co-Located with AAMAS/IJCAI/ECAI/ICML 2018)* (Vol. 14, pp. 50–61).

Panisson, A. R., Sarkadi, Ş., McBurney, P., Parsons, S., & Bordini, R. H. (2019). On the formal semantics of theory of mind in agent communication. In *International Conference on Agreement Technologies* (pp. 18–32).

Paquette, P., Lu, Y., Bocco, S. S., Smith, M., O-G, S., Kummerfeld, J. K., ... Courville, A. C. (2019). No-press Diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, *32*, 4474–4485.

Parsons, S., Sierra, C., & Jennings, N. (1998). Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, *8*(3), 261–292.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." Attribution of second-order beliefs by 5-to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471.

Peterson, D. (2011). *Sex, Lies, and Fireflies: Does deceit happen naturally?* Retrieved January 23, 2023, from `https://www.psychologytoday.com/us/blog/the-moral-lives-animals/201109/sex-lies-and-fireflies`

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Rahwan, I., Ramchurn, S. D., Jennings, N. R., McBurney, P., Parsons, S., & Sonenberg, L. (2003). Argumentation-based negotiation. *The Knowledge Engineering Review*, *18*(4), 343–375.

Raiffa, H. (1982). *The Art and Science of Negotiation*. Harvard University Press.

Raiffa, H., Richardson, J., & Metcalfe, D. (2002). *Negotiation Analysis: The Science and Art of Collaborative Decision Making*. Belknap Press of Harvard University Press.

Ranger, S. (2020). *What is the IoT? Everything you need to know about the Internet of Things right now.* Retrieved February 24, 2023, from `https://www.zdnet.com/article/what-is-the-internet-of-things-everything-you-need-to-know-about-the-iot-right-now/`

Raveenthiran, A. (2023). *Human Behaviour Analysis of Human-Human and Human-Agent Interactions in the Mod-Signal Game.* (Bachelor's thesis, University of Groningen)

Roff, H. (2020). *AI Deception: When Your Artificial Intelligence Learns to Lie.* Retrieved January 15, 2023, from `https://spectrum.ieee.org/ai-deception-when-your-ai-learns-to-lie`

Rogers, K., Webber, R. J. A., & Howard, A. (2023). Lying about lying: Examining trust repair strategies after robot deception in a high-stakes HRI scenario. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 706–710). Association for Computing Machinery.

Rosette, A. S., Kopelman, S., & Abbott, J. L. (2014). Good grief! Anxiety sours the economic benefits of first offers. *Group Decision and Negotiation*, *23*, 629–647.

Sarkadi, Ş. (2018). Deception. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 5781–5782).

Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., & Parsons, S. (2019). Towards an approach for modelling uncertain theory of mind in multi-agent systems. In *Agreement Technologies: 6th International Conference, AT 2018, Bergen, Norway, December 6-7, 2018, Revised Selected Papers 6* (pp. 3–17).

Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., & Chapman, M. (2019). Modelling deception using theory of mind in multi-agent systems. *AI Communications*, *32*(4), 287–302.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226–233.

Sherwood, C. N. (2022). A lie is a lie: The ethics of lying in business negotiations. *Business Ethics Quarterly*, *32*(4), 604–634.

Sierra, C., Jennings, N. R., Noriega, P., & Parsons, S. (1997). A framework for argumentation-based negotiation. In *International Workshop on Agent Theories, Architectures, and Languages* (pp. 177–192).

Sklar, E., Parsons, S., & Davies, M. (2005). When is it okay to lie? A simple model of contradiction in agent-based dialogues. In *Argumentation in Multi-Agent Systems: First International Workshop, ArgMAS 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers 1* (pp. 251–261).

Soon, G. K., On, C. K., Anthony, P., & Hamdan, A. R. (2019). A review on agent communication language. In *Computational Science and Technology. Lecture Notes in Electrical Engineering* (Vol. 481, pp. 481–491). Springer.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, *8*(4), 339–356.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810.

Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development*, *26*(5), 436–444.

Talwar, V., & Lee, K. (2008). Social and cognitive correlates of children's lying behavior. *Child Development*, *79*(4), 866–881.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

van Ditmarsch, H. (2014). Dynamics of lying. *Synthese*, *191*(5), 745–777.

van Ditmarsch, H., Hendriks, P., & Verbrugge, R. (2020). Editors' review and introduction: Lying in logic, language, and cognition. *Topics in Cognitive Science*, *12*(2), 466–484.

van Ditmarsch, H., van Eijck, J., Sietsma, F., & Wang, Y. (2012). On the logic of lying. *Games, Actions and Social Software: Multidisciplinary Aspects*, 41–72.

van Ditmarsch, H., van Eijck, J., & Verbrugge, R. (2009). Common knowledge and common belief. In J. van Eijck & R. Verbrugge (Eds.), *Discourses on Social Software* (pp. 99–122). Amsterdam University Press.

van Poucke, D., & Buelens, M. (2002). Predicting the outcome of a two-party price negotiation: Contribution of reservation price, aspiration price and opening offer. *Journal of Economic Psychology*, *23*(1), 67–76.

van der Wall, S. B. (1990). *Food Hoarding in Animals*. University of Chicago Press.

Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, *38*(6), 649–680.

Verbrugge, R., & Mol, L. (2008). Learning to apply theory of mind. *Journal of Logic, Language and Information*, *17*, 489–511.

Vygotsky, L. S., & Cole, M. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.

Weiss, G. (1999). *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT press.

Wellman, M. P., Greenwald, A., & Stone, P. (2007). *Autonomous Bidding Agents: Strategies and Lessons from the Trading Agent Competition*. MIT Press.

Whiten, A., & Byrne, R. W. (1988). The Machiavellian intelligence hypotheses. In W. A. Byrne R.W (Ed.), *Machiavellian Intelligence: Social Complexity and the Evolution of Intellect in Monkeys, Apes and Humans* (pp. 1–9). Clarendon Press/Oxford University Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wooldridge, M. (1999). Intelligent agents. In *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence* (Vol. 1, pp. 27–73).

Wooldridge, M. (2009). *An Introduction to Multiagent Systems*. John Wiley & Sons.

A
The Concept of Pareto Efficiency

Vilfredo Pareto (1848 – 1923), an Italian economist, introduced the concept of Pareto efficiency (Corporate Finance Institute, 2021). Pareto efficiency is a state where it is not possible to reallocate goods such that at least one individual is better off without making another individual worse off. A Pareto inefficient state is thus a state where a reallocation of goods is possible and where an individual can be better off without making another individual worse off. This introduces the concept of a Pareto improvement, which is a condition in which a reallocation of goods makes at least one individual better off without making another individual worse off. Ultimately, by Pareto improvements, we can move from a Pareto inefficient state to a Pareto efficient state.

In the context of this thesis, Pareto efficiency is reached when no offer can be made such that the new distribution of colored chips yields a higher score for one agent without decreasing the score for another agent. Here, we do not subtract the penalty for making an offer from the final score. A Pareto improvement in our context thus means that an offer exists such that the new distribution yields a higher score for one agent without decreasing the score of the trading partner. A Pareto inefficient state is consequently a distribution of chips in a particular game where a Pareto improvement is possible.

B
Additional Plots

B.1 Experiment 1: Setting the probability of a *ToM*$_0$ agent sending messages

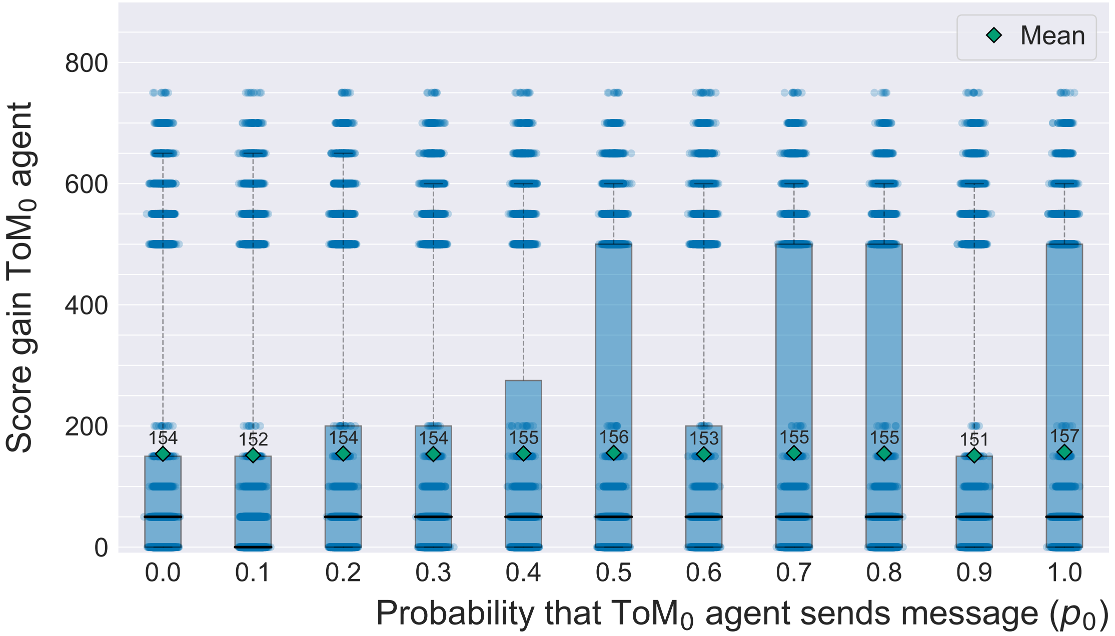


Figure B.1: Box plot of the score gain per value of p_0, i.e., the probability of a ToM_0^* agent sending a goal location message together with an offer. Compared to Figure 4.1, the data points are plotted on top of the box plot to show the distribution of the score gain. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points are scattered normally along the x-axis to increase readability.

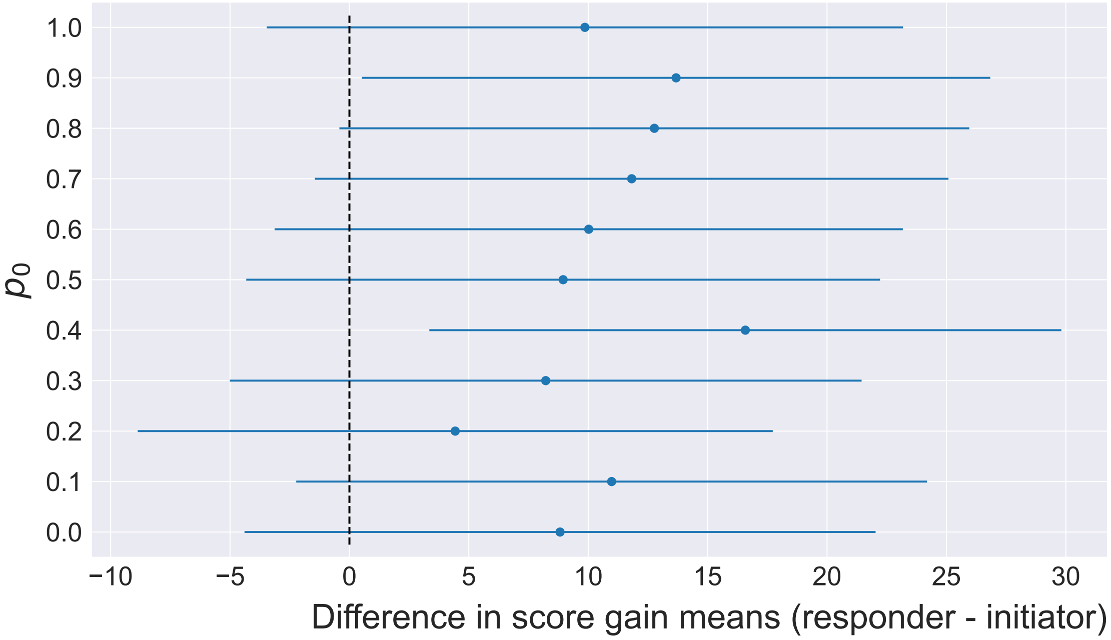


Figure B.2: Difference in score gain mean values between the responder and initiator for different values of p_0, i.e., the probability that a *ToM*$_0$ agent sends a goal location message together with an offer, with 99.5% Bonferroni-adjusted confidence intervals (11 comparisons) to ensure a family-wise error rate of less than 0.05. The confidence intervals are constructed using a t-statistic. If a confidence interval contains 0, we do not have enough statistical evidence to conclude that the difference is unequal to zero, i.e., that there is a difference in score gain mean values between the responder and the initiator. Note that we cannot use matched samples here since we consider the score gain means of the *ToM*$_0$ agents, and the *ToM*$_0$ agents also negotiate with other types of agents.

B.2 Experiment 2: Does lying and misleading outperform honesty?

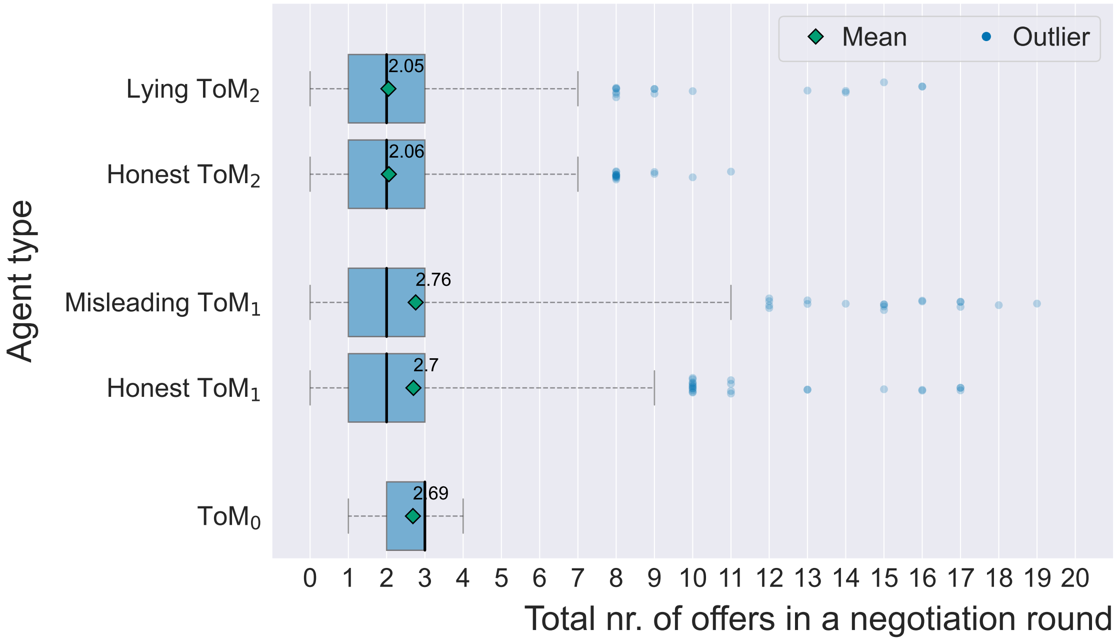


Figure B.3: A box plot of the total number of offers in a negotiation round. In this plot, data points are gathered where each agent type (see Section 3.3) only negotiates with a similar agent type, both as responder and initiator. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

B.3 Experiment 4: Varying the learning speed

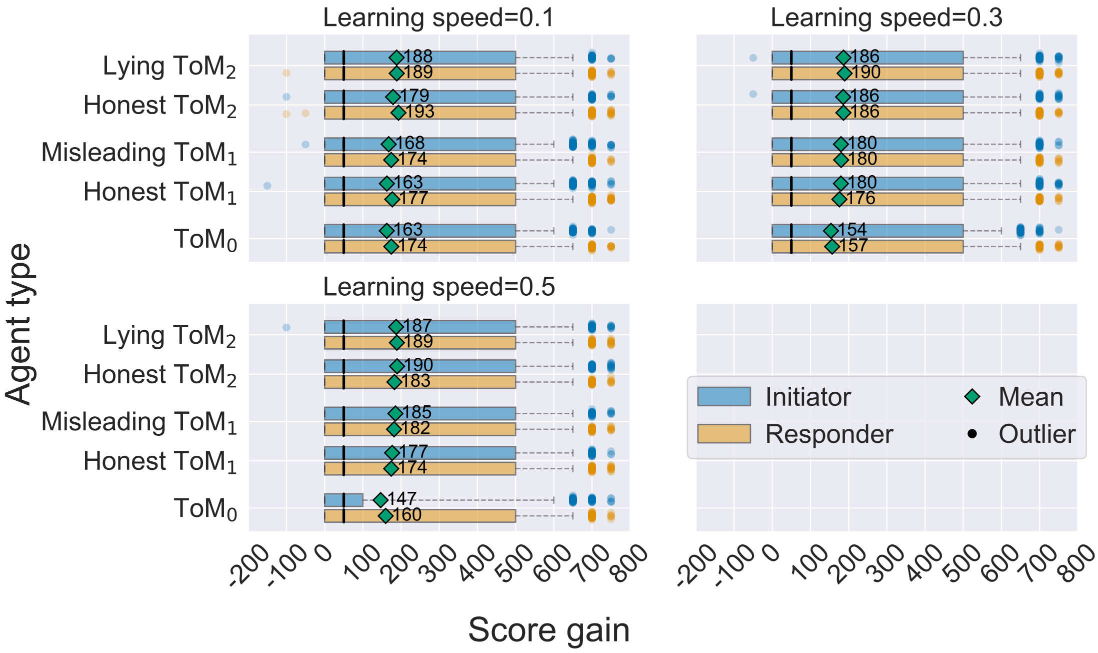


Figure B.4: Box plot of the score gain for the five different types of agents (see Section 3.3) for different levels of learning speeds λ separated into initiators and responders. Agents in a negotiation have the same learning speed. The lower and upper whiskers of the box plot reach the bottom 2.5% and the top 97.5% data points, respectively. Data points (outliers) are scattered normally along the y-axis to increase readability.

C

Installation

This appendix includes the instructions to run the GUI in Java. The GUI has been tested on Windows using IntelliJ IDEA and Eclipse and Java versions 17 and 19. Moreover, we ran experiments on the Hábrók high-performance computing cluster (Center for Information Technology, 2023) that uses a Linux operating system, where we tested the code on Java versions 11 and 17.

Below, I explain the steps to download Eclipse and IntelliJ to run the GUI on a Windows operating system. Moreover, I provide an intuition on how to run the GUI in a Linux environment. One may also prefer to use its own Java tool to run the GUI.

Might the problem arise that the GUI does not fit your screen, try the following. Go to `Settings` on your laptop by selecting `Start` > `Settings` > `System` > `Display`. Then, to change the size of your text and apps, choose an option (either 125% or 100% (recommended)) from the drop-down menu next to `Scale`. Now, the GUI should fit your screen!

C.1 Windows

C.1.1 Eclipse

C.1.1.1 Installation Eclipse

One can follow the steps below to install Eclipse to run Java:

1. One can install Eclipse via the following link: `https://www.eclipse.org/downloads/`.

2. Click on `Download` under *Get Eclipse IDE*, and then again on `Download` on the next page for installation of Eclipse.

3. After the files have been downloaded, click on the downloaded executable and follow the steps of the eclipseinstaller.

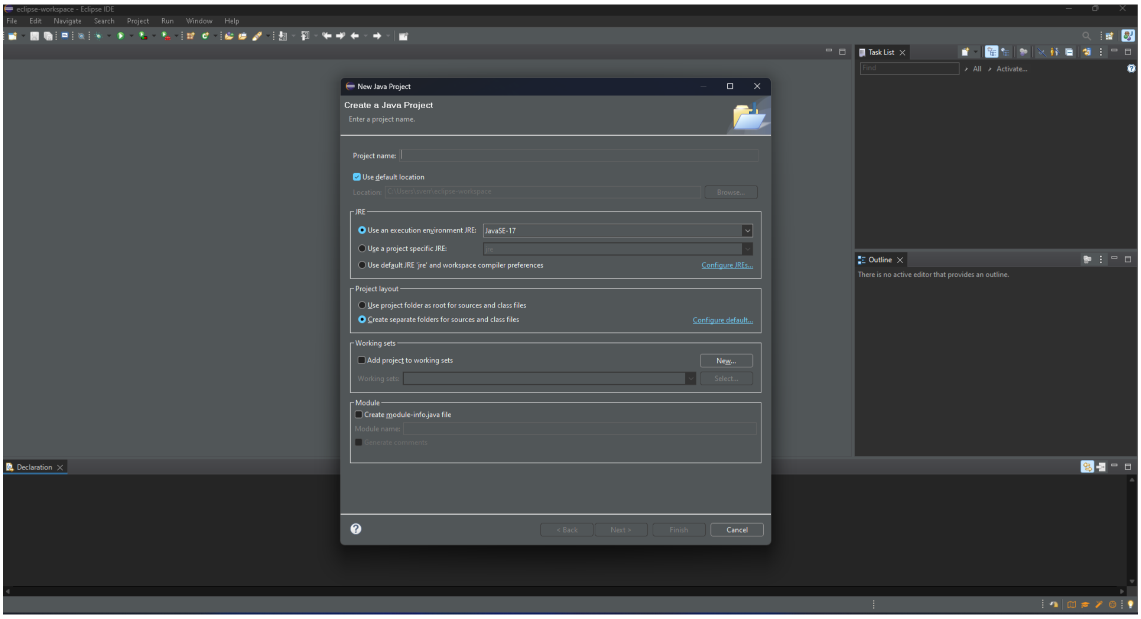
4. Once you finished downloading Eclipse, open the executable of Eclipse to start Eclipse.

C.1.1.2 Run the graphical user interface

1. Download the code from `https://github.com/SverreBr/MasterThesis` (a zip file) or clone the git repository.

2. Place the zip file in the preferred folder and extract all files. You should now have a folder that is named `MasterThesis-master` (or any other folder name where you placed the files) and contains at least two subfolders: `fig` and `src`.

3. In your Eclipse workspace, click on `file` > `new` > `Java Project`. The following should appear:


eclipse-workspace - Eclipse IDE
File Edit Navigate Search Project Run Window Help
Task List
Find
All
Activate...
Outline
Declaration
Back
Next
Finish
Cancel

New Java Project

Create a Java Project

Enter a project name.

Project name:

☑ Use default location

Location: C:\Users\sverr\eclipse-workspace

JRE

● Use an execution environment JRE: JavaSE-17

○ Use a project specific JRE: jre

○ Use default JRE 'jre' and workspace compiler preferences

Project layout

○ Use project folder as root for sources and class files

● Create separate folders for sources and class files

Working sets

□ Add project to working sets

Working sets:

Module

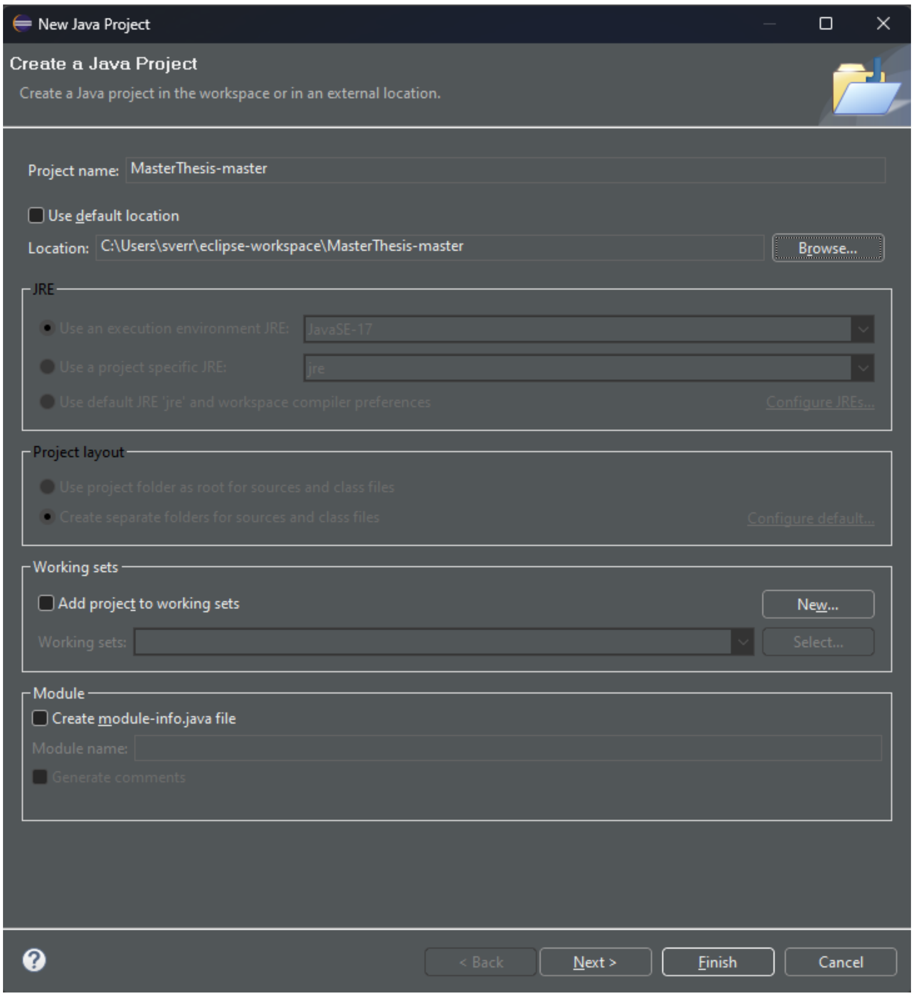
□ Create module-info.java file

Module name:

□ Generate comments

There is no active editor that provides an outline.

4. Untick `Use default location` and browse to the folder named `MasterThesis-master`. Further, untick `Create module-info.java file`. It should look like the following:



Create a Java Project

Create a Java project in the workspace or in an external location.

Project name: MasterThesis-master

□ Use default location

Location: C:\Users\sverr\eclipse-workspace\MasterThesis-master

JRE

● Use an execution environment JRE: JavaSE-17

○ Use a project specific JRE: jre

○ Use default JRE 'jre' and workspace compiler preferences

Project layout

○ Use project folder as root for sources and class files

● Create separate folders for sources and class files

Working sets

□ Add project to working sets

Working sets:

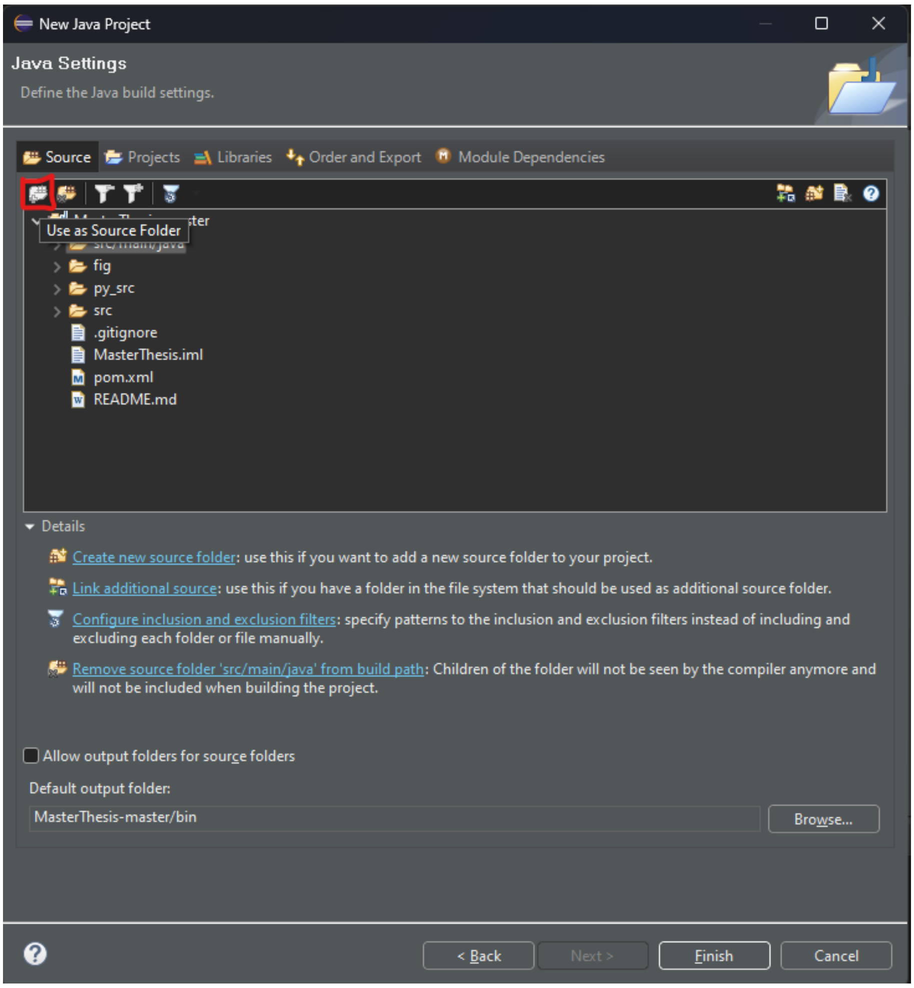
Module

□ Create module-info.java file

Module name:

□ Generate comments

5. Click on `Next` and ensure that `src/main/java` is labeled as a source folder. If this is not the case, it can be done by clicking on the icon indicated by a red rectangle in the following figure. Then, click on `Finish`.



Java Settings

Define the Java build settings.

- fig
- py_src
- src
- .gitignore
- MasterThesis.iml
- pom.xml
- README.md

Details

Create new source folder: use this if you want to add a new source folder to your project.

Link additional source: use this if you have a folder in the file system that should be used as additional source folder.

Configure inclusion and exclusion filters: specify patterns to the inclusion and exclusion filters instead of including and excluding each folder or file manually.

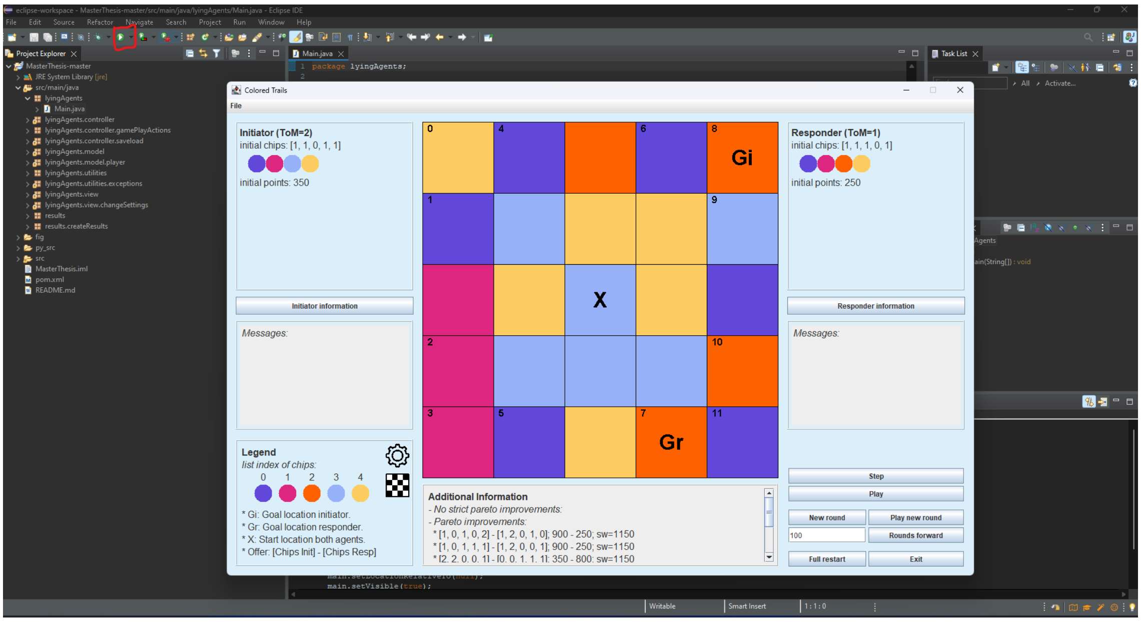
Remove source folder 'src/main/java' from build path: Children of the folder will not be seen by the compiler anymore and will not be included when building the project.

☐ Allow output folders for source folders

Default output folder:

MasterThesis-master/bin

6. First open the file `Main.java` that is located in the folder `src/main/java/lyingAgents`. Then, run the GUI by clicking on the icon indicated by a red rectangle in the following figure. This should look at follows:



Initiator (ToM=2)
initial chips: [1, 1, 0, 1, 1]
initial points: 350

Responder (ToM=1)
initial chips: [1, 1, 1, 0, 1]
initial points: 250

Messages:

Legend
list index of chips:
* Gi: Goal location initiator.
* Gr: Goal location responder.
* X: Start location both agents.
* Offer: [Chips Init] - [Chips Resp]

Additional Information
- No strict pareto improvements:
- Pareto improvements:
* [1, 0, 1, 0, 2] - [1, 2, 0, 1, 0]; 900 - 250; sw=1150
* [1, 0, 1, 1, 1] - [1, 2, 0, 0, 1]; 900 - 250; sw=1150
* [2. 2. 0. 0. 1] - [0. 0. 1. 1. 1]: 350 - 800: sw=1150

C.1.2 IntelliJ IDEA

C.1.2.1 Installation IntelliJ IDEA

One can follow the steps below to install the IntelliJ IDEA (community edition) to run Java:

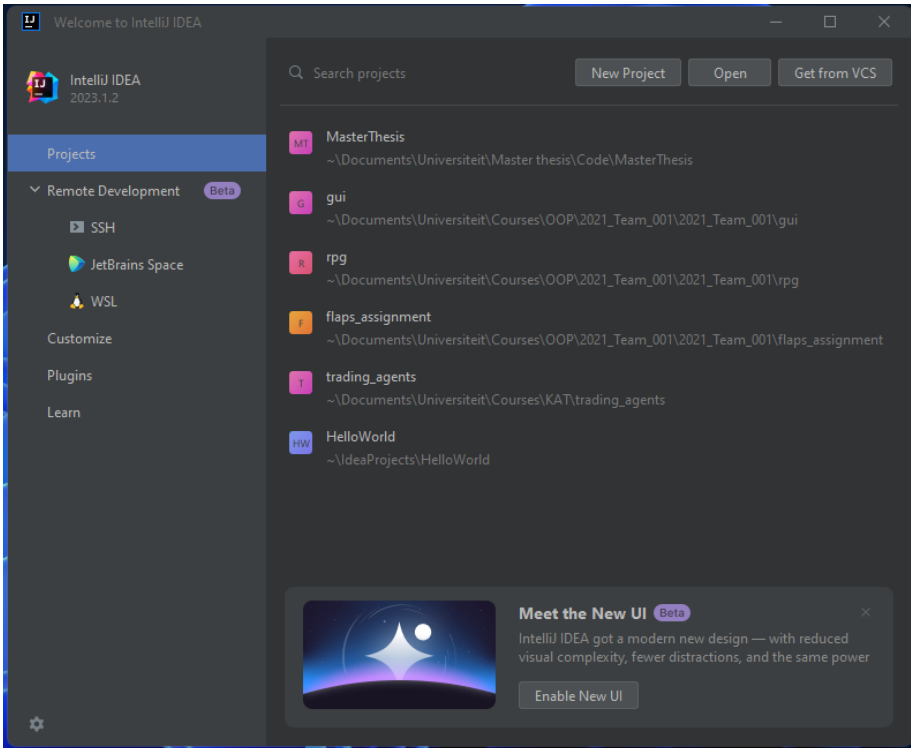
1. One can follow the steps of the following link to install the IntelliJ IDEA: `https://www.jetbrains.com/help/idea/installation-guide.html`
2. Once you finished downloading the IntelliJ IDEA by following, open the IntelliJ IDEA.

C.1.2.2 Run the graphical user interface

After opening IntelliJ IDEA, one can choose to either download the code from `https://github.com/SverreBr/MasterThesis` (a zip file) or clone the git repository. In case one chooses to download the code, you can follow these steps:

1. Download the code from `https://github.com/SverreBr/MasterThesis` (a zip file).
2. Place the zip file in the preferred folder and extract all files. You should now have a folder that is named `MasterThesis-master` (or any other folder name where you placed the files) and contains at least two subfolders: `fig` and `src`.

3. In your IntelliJ IDEA workspace, click on `open` (see Figure C.1) and search for the folder where you placed the code. You might be asked to trust the project. After that, one can simply run the code by clicking on the `run` button.



MasterThesis
~\Documents\Universiteit\Master thesis\Code\MasterThesis

gui
~\Documents\Universiteit\Courses\OOP\2021_Team_001\2021_Team_001\gui

rpg
~\Documents\Universiteit\Courses\OOP\2021_Team_001\2021_Team_001\rpg

flaps_assignment
~\Documents\Universiteit\Courses\OOP\2021_Team_001\2021_Team_001\flaps_assignment

trading_agents
~\Documents\Universiteit\Courses\KAT\trading_agents

HelloWorld
~\IdeaProjects\HelloWorld

Meet the New UI

IntelliJ IDEA got a modern new design — with reduced visual complexity, fewer distractions, and the same power

Figure C.1: IntelliJ IDEA workspace

In case one chooses to clone the git repository, one can click on `Get from VCS` in the IntelliJ IDEA workspace (see Figure C.1) and enter the repository URL.

C.2 Linux

We only tested running experiments in a Linux environment. Since we used the Hábrók high-performance computing cluster, we were not able to test the GUI. Nevertheless, an example of the commands to run the GUI in a Linux environment may be as follows:

1. `javac --source-path src/main/java -d bin src/main/java/lyingAgents/Main.java`

2. `java -cp bin lyingAgents/Main`

Make sure that Java is installed or loaded.