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

The Evolution of Theory of Mind in a Simulated Population: a Mixed-Motive Setting

Graduation Project

Sanne Berends

October 2024

Internal Supervisor(s):
Dr. Harmen de Weerd
Prof. dr. Rineke Verbrugge

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

Abstract

One of the characteristics that distinguishes humans from other species is our advanced awareness that other people have goals, intentions, and beliefs: our theory of mind. We can apply this theory of mind recursively, which is known as higher-order theory of mind. There exist various hypotheses as to why humans developed this skill, while other species do not show evidence of this. One of these hypotheses, the Mixed-motive interaction hypothesis, suggests that mixed-motive social situations like negotiations drove the evolution of theory of mind in humans. Simulations of pairwise interactions have shown that indeed, agents with (higher-order) theory of mind performed better than agents without theory of mind. However, pairwise interactions alone do not capture the complexity of social group settings. Whether the use of theory of mind also provides advantages on a population level has not been studied yet.

To further test the validity of the Mixed-motive interaction hypothesis, we constructed a simulated evolutionary process where agents with various orders of theory of mind (zero, one, or two) trade resources to survive. The negotiations needed for a successful trade form the mixed-motive aspect of the environment. We adapted an existing model of theory of mind to fit this new setting. Two experiments were conducted, one where agents did not distinguish between trading partners, and one where agents remembered partner-specific information. In the latter, agents rejected negotiations with incompatible partners and remembered what they learned from previous negotiations with this partner. In both experiments, the number of agents per theory of mind order was tracked to see which species survived in this environment.

The results of the experiments show that agents benefited from their theory of mind in the pairwise negotiations, but that the agents without theory of mind outlasted these species in the evolutionary process. This was the case for both experiments. The time it took agents to find a consensus in a negotiation when using theory of mind took significantly longer than it took agents with no theory of mind. As a consequence, the former more often failed to reach their resource threshold. In the long run, this caused the skill to go extinct.

These findings shed new light on the possible evolutionary advantages of higher-order theory of mind in mixed-motive settings.

Acknowledgments

This graduation project has been a year-long journey, filled with extensive reading, coding, and writing. Now that I have completed this research, I know that writing this acknowledgments section is the simplest part of the project, due to the vast amount of support I have received from so many people. First of all, I want to thank my supervisor Harmen de Weerd. Every time we had a progress meeting, I left feeling a lot more secure about my project, as well as having renewed enthusiasm to work on it. The feedback and guidance that I received helped me gain a fresh perspective on the challenges I was facing. I additionally would like to thank Rineke Verbrugge, my second supervisor, who lent me the book Thinking Big: How the Evolution of Social Life Shaped the Human Mind, which was not only interesting but also very useful for my literature review. I am grateful that I got to work under the supervision of two researchers who contributed so much to the field I have dived into these past months.

Apart from my supervisors, I would also like to express my appreciation for Hábrók, which is the high-performance computing cluster of the RUG. I was granted permission to use this cluster by Center for Information Technology (2024), which saved me an incredible amount of time.

Finally, I would like to express my appreciation for all the others who supported me during the course of this project. The daily walks in the park to discuss new ideas, the lunch breaks, but also the quality time off on the weekends really made this a wonderful experience.

Contents

Chapter 1

Introduction

One of the characteristics that distinguishes humans from other animals is our advanced awareness that other people have goals, intentions, and beliefs. We may not always know or understand them, but we are able to reason with them while making decisions. This Theory of Mind (ToM; Premack and Woodruff, 1978) can help us choose a desirable action in social settings. It is also useful in games, where we can use ToM to try to predict what the other person will do to decide our own action. Take for example the 'rock-paper-scissors' game: in order to win the game, you try to predict which of the three items your opponent will pick, and choose the item that beats him. Without ToM, we might simply choose the action that would have worked best in the last game(s). As an example, consider getting your favorite snack from an unlabeled vending machine: you do not even consider there to be a competitive situation because you cannot reason about the goals of others (the 'other' being the vending machine in this analogy). With theory of mind, however, we try to predict what the opponent will do as a reaction to our own previous actions. Thus, considering other's intentions (i.e., using ToM) can impact our own decisions.

Humans are not only capable of the described ToM, which is known as *first-order* ToM. We are also capable of second-order Theory of Mind (ToM2). This means that we are aware that others can reason about our own goals, intentions, and beliefs. We can thus consider that their actions may be impacted by them considering what our intentions are. This kind of reasoning is recursive, going back and forth between agents: an agent models the model that another has of this agent (Verbrugge, 2009). The level of recursion is the order of the ToM. When an agent has no ToM, it only concerns observable facts. This is known as zero-order ToM (ToM0). Orders of ToM above ToM1 are considered higher-order ToM (Verbrugge, 2009). ToM2 agents, for example, model others as ToM1 agents, which they think model them as ToM0 agents. Adults can generally understand higher-order ToM up to ToM4 (Kinderman et al., 1998). ToM4 allows us to (vaguely) understand sentences like "Andrew knows that Bob thinks that Andrew believes that Chris intends to go to the party". In practice, ToM4 may not

necessarily provide an additional social benefit compared to ToM3. Instead, it depends on the setting: the use of ToM4 is more commonly seen in story tasks than in strategic game settings, where increasingly higher orders of ToM typically exhibit diminishing return (e.g., De Weerd and Verheij, 2011; De Weerd et al., 2012; De Weerd et al., 2022).

These examples show that ToM is a useful skill that humans use on a daily basis. This could mean that the use of ToM would also be beneficial to animals. The ability of animals to use ToM is an area of research that initially focused mainly on primates like chimpanzees since these species are closely related to us. Premack and Woodruff (1978) and Povinelli et al. (1996), for example, investigated whether these animals have a notion of the mental states of others by using various experimental setups. Research on ToM in animals is not limited to just primates, it also focuses on other species like corvids (Emery et al., 2004). The research by Emery and colleagues describes how corvids seem to attribute mental content to other corvids and use this to hide their food, suggesting the use of ToM. However, there is no consensus about whether non-human animals are capable of using ToM because the conclusions that are drawn based on research in this area are conflicting (Krupenye and Call, 2019; Van der Vaart and Hemelrijk, 2014; Arre and Santos, 2021). It is difficult to assess the mental processes of an animal when it is impossible to verbally communicate with them. The main argument against animals having ToM is therefore that the behavior of animals in the experiments can be caused by other cognitive processes than ToM: it is not possible to rule these alternative explanations out. Current research gives us reason to believe that the 'ToM' of animals is not as advanced as in humans, and that higher-order ToM is unique to humans (Krupenye and Call, 2019; Van der Vaart and Hemelrijk, 2014). This raises the question of which social settings caused humans to develop this advanced cognitive skill, whilst other species did not do this to this extent.

To explore this question, research focuses on the settings in which ToM can be useful. Agent-based simulation research has shown that the use of ToM can be beneficial in competitive, cooperative, and mixed-motive settings (De Weerd et al., 2013b; De Weerd, Verbrugge, and Verheij, 2015; De Weerd et al., 2017). A competitive setting is a setting in which the gain of one individual is the loss of another individual. Individuals compete with each other to reach their own goal. A cooperative setting on the other hand is a setting in which the gain of an individual is also the gain of another individual. They work together towards a shared goal. A mixed-motive setting involves a combination of cooperation and competition. An example is a negotiation, where individuals need each other to reach their personal goal, but they want the best possible outcome for themselves. The ability to take the goals of others into account allows an individual to increase their personal gain in each of these settings. However, since animals also encounter competitive and cooperative settings, this alone does not explain why humans have uniquely advanced (higher-order) ToM.

The relevance of ToM across different settings offers insights into its evolution in humans. The Machiavellian intelligence hypothesis indicates competitive settings as the cause of the development of ToM in humans (Byrne and Whiten, 1988, as cited in De Weerd et al., 2022), the

Vygotskian intelligence hypothesis indicates cooperative settings as the cause (Vygotsky and Cole, 1978, as cited in Moll and Tomasello, 2007), and the Mixed-motive interaction hypothesis indicates negotiation settings as the cause of the development of ToM in humans (Verbrugge, 2009). Research by De Weerd et al. (2014, 2017) suggests that the Mixed-motive interaction hypothesis is more likely than the Machiavellian intelligence hypothesis and the Vygotskian intelligence hypothesis, because higher-order ToM was shown to be more beneficial to the agents than in other settings. De Weerd et al. (2022) additionally state that we may have developed ToM due to the complex social settings and unpredictable environments encountered during evolution.

Given the lack of (higher-order) ToM evidence in animals despite their exposure to competitive and cooperative interactions and the advantages of higher-order ToM in mixed-motive settings, this study will explore the Mixed-motive interaction hypothesis further. The conclusions that were drawn from De Weerd et al. (2014, 2017, 2022) that support the hypothesis are all based on pairwise interactions. Whether or not the advantages of (higher-order) ToM reasoning found in pairwise mixed-motive interactions translate into advantages on a population level is an open question.

1.1 The Current Research

This thesis aims to provide new insights into the evolution of ToM by simulating a multi-agent setting and establishing an evolutionary process in a population of agents. Individual agents use different orders of ToM, ranging from zero to two. The agents are positioned in an environment that requires negotiation to gather sufficient resources. Agents can 'die' when failing to gather all resources, and are then replaced by a new agent, which is a copy of a randomly chosen agent of the population. Some randomness (mutation) is added to the 'reproduction': with a small probability, the agent that was chosen to 'reproduce' may not be copied exactly, but instead altered a bit. The order of ToM of the new agent may thus differ from that of its 'parent'. This process contains all the main components of an evolutionary process, which are replication, selection and mutation (see Nowak (2006)).

The aim of the research is to see how ToM in the agents evolves over time, and if in such complex social settings the use of higher-order ToM increases an agent's chances to survive. The research question is thus:

How do various orders of Theory of Mind evolve in a population of agents placed in a mixed-motive environment?

Sub-questions that will be used to answer the research question are:

1. *Is there an order of ToM that is the 'winner' in this environment, or is a dynamic equilibrium reached?*

2. *Do lower orders of ToM (ToM0, ToM1) go extinct over time? In other words: does higher-order ToM provide an evolutionary benefit in this negotiation environment?*

The structure of this thesis is as follows: Chapter 2 provides an overview of relevant background literature. It includes a discussion of research in the field of ToM, agent-based modeling, and evolution theory. Chapter 3 details the methods of the research. It describes the two experiments that were performed, as well as the GUI that can be used to run these experiments. The results of the experiments are presented in Chapter 4. Finally, Chapter 5 discusses the implications of the findings, provides suggestions for future research, and concludes with the main insights derived from this study.

Chapter 2

Background Literature

Artificial Intelligence aims to create computer systems that mimic human intelligence. Intelligence in this sense is not just brain power, but also comprehension of surroundings. Theory of Mind is a skill that humans use to understand the motives of others and to guess their actions. What Theory of Mind is, whether we are the only species capable of Theory of Mind and how we developed this skill is discussed in the first section of this chapter (Section 2.1). The current research aims to explain the emergence of Theory of Mind in a mixed-motive (negotiation) setting. For this reason, an agent-based model was created. Section 2.2 presents information on agent-based models, which can be used as a research tool to simulate societies. This section also discusses why agent-based models can be useful to study Theory of Mind. The model that we present simulates an evolutionary process that is based on the principles of Darwinian evolution. Literature about the Darwinian Evolution is discussed in Section 2.3. This chapter thus gives an overview of previous studies that are relevant for the topic of the current research.

2.1 Theory of Mind

Theory of Mind (ToM; Premack and Woodruff, 1978) denotes the ability to reason about the mental content of other beings. This includes the goals, intentions, and beliefs of others. This skill is useful in a variety of social settings: think about playing the game poker, working in teams, or negotiating about your salary. In these scenarios, considering the thoughts of others to predict their actions may benefit your personal gain. Humans develop the ability to use ToM when they are between three and five years old (Wellman et al., 2001). This has been empirically established by using the so-called *false belief task*, an experimental paradigm to assess ToM (Wimmer and Perner, 1983; see Section 2.1.2).

Most individuals learn how to use *higher-order* ToM, meaning that their use of ToM is recursive (Verbrugge, 2009). People are thus aware that others can reason about their own mental state. The level of recursion is the order of the ToM. Thus, *zero-order* ToM (ToM0)

refers to the inability to use ToM, while *fourth-order* ToM (ToM4) refers to people's ability to understand sentences like "Andrew knows that Bob thinks that Andrew believes that Chris intends to go to the party". Adults are generally able to understand and use ToM4 (Kinderman et al., 1998).

There is no consensus on whether we are the only species capable of ToM, but it is generally accepted that humans are the only species capable of higher-order ToM. This raises the question as to why humans did develop such a complex cognitive skill, whilst other species most likely did not. This section first reviews a selection of studies regarding ToM in animals (Section 2.1.1). Then, in Section 2.1.2 the capacity of ToM reasoning in humans is discussed, followed by a discussion of the evolutionary development of ToM in humans in Section 2.1.3.

2.1.1 Theory of Mind in Animals

There is no definite answer to the question if animals can reason about the mental state of others, even though it has been a topic of research for several decades. The question holds significance because knowing if other animals developed this skill allows us to determine what types of social settings may cause species to develop ToM. The first researchers to address ToM in animals were Premack and Woodruff in their paper titled *Does the chimpanzee have a theory of mind?* (Premack & Woodruff, 1978). A chimpanzee was presented with videos of a human that faced a problem like not being able to reach a banana, or being locked in a cage. The chimpanzee was given several images and had to decide which image provided a solution to the problem of the video. The results indicated that the chimpanzee was capable of solving the problem for the human in many cases. Based on these results, the question arose whether the behavior of the chimpanzee was caused by the chimpanzee projecting mental contents onto the human and thus using ToM to solve the problem, or if other interpretations are more plausible.

Several studies followed, many of them focusing on primates. Examples are Povinelli et al. (1996), who studied if chimpanzees understand visual perception, Call and Tomasello (1999), who studied chimpanzee's and orangutan's performance on false belief tasks, and Kummer et al. (1996), who studied Macaque's ability to take the perspective of the experimenter. Each of these three studies showed some evidence of ToM in primates. Critics argue that the behavior of primates that some researchers interpret as evidence of ToM, may as well be achieved by associatism (Premack and Woodruff, 1978; Povinelli and Vonk, 2003; Penn and Povinelli, 2007). This means that primates choose a certain action because they recognize the situation and choose the next step of the familiar sequence. The inability to verbally communicate with primates withholds us from distinguishing between the two interpretations.

A recent study (Arre & Santos, 2021) concludes that primates have some kind of ToM to represent the vision and knowledge of others. They note that current research does not provide enough evidence that primates can impute other's beliefs. This conclusion is in line with a study by Povinelli and Vonk (2004) and Call and Tomasello (2008). The latter reflects on past

research regarding ToM in primates. Call and Tomasello conclude that chimpanzees have some understanding of other's goals, intentions, perception, and knowledge, even though there is not enough evidence to believe that they have ToM as advanced as humans that allows them to understand that others have a mental representation of the world. A similar conclusion as Call and Tomasello was drawn by Tomasello et al. (2003). Puga-Gonzalez et al. (2009) introduce an Agent Based Model (ABM; see Section 2.2) of primates with an anxiety level which motivates the grooming of others, generating affiliative effects also observed in real primates. This research shows that complex primate behavior can be explained by other cognitive processes than ToM, and thus shows an alternative interpretation of research results suggesting primate (higher-order) ToM.

Non-primates have also been the subject of ToM research, especially corvids. An example is a research by Van der Vaart et al. (2012), who investigated the re-caching phenomenon of the scrub jay as described in Emery et al. (2004). Empirical evidence showed that scrub jays, a type of corvid, relocate their caches when they are watched. Emery and colleagues interpret this finding as evidence supporting that scrub jays have a ToM, and use this to hide their food from others by reasoning about the mental content of competitors. Van der Vaart and colleagues provide an alternative explanation: they simulate corvids that contain a stress level that is impacted by the presence of others and the inability to relocate their caches. Their model shows results similar to empirical evidence, solely caused by stress levels. This research contributes to the evidence that other cognitive processes than ToM can explain the behavior of animals that was previously assumed to be the result of the use of ToM.

Thus, so far the evidence on the capability of non-human animals to use ToM is conflicting, but there is no evidence that non-human animals use higher-order ToM. A demanding cognitive skill like (higher-order) ToM evolves in a species when there is an advantage to using the skill (see Section 2.3). There exist various hypotheses (see Section 2.1.3) that offer possible explanations as to which types of situations foster the development of a ToM. Primates encounter both competitive and cooperative social situations in their daily activities, and both types of settings have been used in experiments (e.g., Hare and Tomasello, 2004). However, they do not show behavior consistent with (higher-order) ToM. Various theories may explain why this is the case (e.g., they may use an alternative to ToM, or the competitive and cooperative settings that they encounter are not competitive or cooperative enough to trigger the development of ToM). A possible explanation could be that some other type of social setting requires this cognitive skill. Research has not yet shown why the competitive and cooperative settings that non-human animals encounter were not enough to develop (the use of) higher-order reasoning. Section 2.1.3 further explores the evolutionary development of ToM.

2.1.2 Theory of Mind in Humans

The *false belief task* is a paradigm designed to test the theory of mind reasoning of children (Wimmer & Perner, 1983). The original false belief task focuses on ToM1, but there exist adaptations to test ToM2 and ToM3 (e.g., Flobbe et al., 2008; Liddle and Nettle, 2006). To test for ToM1, the Sally-Anne test is often used (Baron-Cohen et al., 1985). A participant is shown an image that shows a cartoon. In this cartoon, a doll named Sally puts marbles in a basket. She then walks away, and another doll, Anne, takes the marbles and places them in a box. The participant and the dolls cannot see the marbles. The question is where Sally will look for her marbles when she returns. To answer correctly, the participant needs to understand that Sally could not see the marbles being moved to the box, and thus that she still has the false belief that they are in the basket. So, to succeed in this false belief task, the participant needs to use ToM1 to infer the beliefs of Sally. Variations on the ToM1 false belief test exist, such as the Chocolate Bar Story (Hogrefe et al., 1986) and the Birthday Puppy Story (Tager-Flusberg & Sullivan, 1994). False belief tasks showed that children develop the ability to use ToM when they are between three and five years old (Wellman et al., 2001).

Second-order false belief tasks showed that ToM2 does not develop until a few years after ToM1 (Flobbe et al., 2008; Liddle & Nettle, 2006). An example of a second-order false belief task is an extension of the Sally-Anne test by Baron-Cohen et al. (1985). Again, Sally puts the marbles in a basket and Anne replaces them to the box. However, in this test, Sally observes Anne through a window, without Anne realizing. Sally thus knows that the marbles have been relocated. The question is where Anne thinks Sally will look for the marbles. The participant now needs to understand that Anne thinks that Sally thinks that the marbles are still in the basket. It thus requires second-order reasoning. Flobbe et al. (2008) used a similar task in their research, an adapted version of the Chocolate Bar first-order false belief task by Hogrefe et al. (1986), and concluded that children develop the ability to use ToM2 when they are between seven and eight years old. However, note that research by Arslan et al. (2020) showed that children between five and six years old can already be trained to correctly apply ToM2 at second-order false belief tasks. Flobbe et al. (2008) note that the use of ToM2 is harder for children than the use of ToM1, and that this also holds for adults, although adults perform better than children. These findings suggest that ToM keeps improving in humans after the age of eight. These conclusions are in line with research by Liddle and Nettle (2006), who found that children aged ten and eleven master the ability to use ToM1 and ToM2. They also manage to apply some ToM3, since they perform above chance on tasks requiring ToM3. Adults can reason using ToM3, better than children, and are additionally capable of using ToM4 (Kinderman et al., 1998).

That adults can reason using up to ToM4 does not mean we always use this ability. Research suggests that we first omit using ToM in interactions, and only apply it when necessary (Meijering et al., 2014). Other research suggests that we use ToM1 by 'default', and only occasionally use ToM2 (Hedden & Zhang, 2002). However, the latter research used a misleading

training session (see Verbrugge et al., 2018). Much of the research on the use of higher-order ToM uses strategic games as a test bed, since these games do not only require the understanding of higher-order ToM, but also its application (e.g., Colman, 2003; De Weerd, Broers, and Verbrugge, 2015; Hedden and Zhang, 2002; McKelvey and Palfrey, 1992; Verbrugge et al., 2018). Previous research shows that humans may adjust their use of ToM to that of others, even if they are not aware of the order of ToM that their opponent/trading partner is using (Colman, 2003; De Weerd, Broers, & Verbrugge, 2015; Hedden & Zhang, 2002). For a detailed review of research on higher-order ToM in games, see Verbrugge (2009).

ToM is thus a useful skill that humans develop at a young age, but that we do not always use to our maximum capability. A plausible explanation for our modest use of this capability is its high cognitive demand (Aiello & Wheeler, 1995; Schneider et al., 2012). To avoid spending more cognitive resources than necessary, people only use higher-order ToM when the situation requires so. These kinds of situations, where higher-orders of ToM like ToM4 give us a social benefit, are likely to have contributed to our evolutionary development of higher-order ToM (De Weerd et al., 2013b) (see Section 2.1.3). However, it is still an open question why some situations, like story tasks, cause humans to use ToM automatically, whilst in others we fail to apply it correctly.

2.1.3 Evolutionary Development of Theory of Mind

What caused humans to develop higher-order ToM, whilst other species did not develop this skill has been a topic of research for a long time. Three hypotheses that offer possible explanations are the Machiavellian intelligence hypothesis (Byrne and Whiten, 1988, as cited in De Weerd et al., 2022 and Moll and Tomasello, 2007), the Vygotskian intelligence hypothesis (Vygotsky and Cole, 1978, as cited in Moll and Tomasello, 2007), and the Mixed-motive hypothesis (Verbrugge, 2009). In the past 15 years, new insights about the evolutionary development of ToM in humans have been gained through the use of agent-based models (ABMs) (see Section 2.2). ABMs have been used to simulate different types of environments to test hypotheses for the development of ToM. Each of the hypotheses is detailed in this section, followed by empirical results from ABMs that suggest which hypothesis is most likely to be true.

According to the Machiavellian intelligence hypothesis, ToM in humans evolved due to competitive settings that humans faced (Byrne and Whiten, 1988, as cited in De Weerd et al., 2022 and Moll and Tomasello, 2007). The hypothesis builds upon the ideas of Humphrey (1976), who argued that social competition in daily life required animals to develop higher intellectual faculties. Social competition refers to a social setting in which the gain of one individual is the loss of another. An example is the competition for food: when one individual eats an animal in the habitat, other individuals cannot eat that specific animal anymore. ToM allows individuals to reason about the ideas and intentions of others, and thus it can help to 'outsmart' others in order to win the competition. This provides an evolutionary advantage

compared to others that do not use ToM.

The Vygotskian intelligence hypothesis indicates cooperative settings as the cause of the development of ToM in humans (Vygotsky and Cole, 1978, as cited in Moll and Tomasello, 2007). A cooperative social setting is a setting in which individuals share a common goal. To reach their goal, these individuals take either a reciprocal or complementary role, and they support each other in their roles if necessary. These characteristics of a cooperative setting were defined by Bratman (1992). An example of such a cooperative setting is hunting. The total reward increases when individuals work together to corner and kill prey. Even though cognition in animals may have emerged as a result of competitive situations, according to Vygotsky and Cole, the more advanced cognitive abilities of humans, such as higher-order ToM, are a result of cooperative settings. ToM could help to improve coordination during a cooperative effort to reach a common goal (De Weerd, Verbrugge, & Verheij, 2015).

According to the third hypothesis, the Mixed-motive interaction hypothesis, social settings that require both cooperation and competition caused humans to develop ToM (Verbrugge, 2009). In these mixed-motive settings, there exists no outcome that is optimal for all individuals, but the individuals can reach a mutually beneficial outcome. One example of such a mixed-motive setting is a negotiation: all participants want to reach a mutual agreement, but each of them also wants to maximize their own gain. A more concrete example is bargaining about exchanging a number of coins for a piece of bread. Both parties want to reach an agreement, but the buyer of the bread wants to minimize his spending, while the seller of the bread wants to maximize his earnings. ToM can allow individuals to predict the intentions and actions of others, which can help them find a beneficial agreement.

ABMs show that agents benefit from the use of ToM in each of the three types of settings: competitive, cooperative, and mixed-motive. These studies are discussed in the next section (Section 2.2.2). The results indicate that higher-order ToM, specifically ToM4, mainly benefits agents in mixed-motive settings.

The finding that ToM4 only offers a benefit to agents in mixed-motive settings provides a plausible explanation as to why humans may have developed higher-order ToM. A skill or ability evolves in a species over time when it provides the species with an environmental benefit (see Section 2.3). As mentioned in Section 2.1.1, primates that encountered cooperative and competitive situations did not show behavior consistent with the use of higher-order ToM. For that reason, De Weerd et al. (2014) state that the Mixed-motive interaction hypothesis provides a more likely explanation for the emergence of ToM in humans than the other two hypotheses.

More recent work (De Weerd et al., 2022) suggests that the benefit of ToM in social settings depends on the level of predictability of the environment. In dynamic environments, where observable variables change over time, agents without ToM struggle to predict the actions of others and, as a consequence, struggle to make desirable decisions. Agents capable of ToM1, ToM2, and ToM3 all outperform agents without ToM. Based on these results, De Weerd and colleagues conclude that higher-order ToM may have emerged in humans due to complex social

settings and unpredictable environments encountered during evolution. Additionally, a recent study by Lenaerts et al. (2024) explores ToM in the context of a sequential dilemma, i.e., the Incremental Centipede Game. The research implemented bounded rationality, considering that people might make mistakes in their rational reasoning. Their results, which show similarities to behavioral data, suggest that ToM could have evolved because it helps people navigate complex social situations by balancing cooperation and competition.

These conclusions are in line with the Social Brain Hypothesis, which designates our social lives as the cause of the development of higher-order reasoning (Gamble et al., 2014). This hypothesis is built on the ideas presented in Dunbar (1996), that language is the key to being part of a community. Dunbar states that the vast majority of language is used for social matters: gossip. This gossip makes the bonds between members of a group of people stronger. Gamble et al. (2014) further developed these ideas into the Social Brain Hypothesis. They advocate that, as group sizes grew for humans, grooming like other primates became too time-consuming. This may have led to the development of language as a form of vocal grooming, which allowed humans to connect with larger groups. The social emotions associated with members of these larger groups require ToM. In turn, language possibly also forms a requirement for reasoning with high-order ToM.

To summarize, considering the research done so far, it seems that humans have developed higher-order ToM due to our embodiment in dynamic environments, and our encounters with situations that require a combination of cooperation and competition. It is, however, important to note that these conclusions were drawn based mainly on research consisting of pairwise interactions (De Weerd, Verbrugge, & Verheij, 2015; De Weerd et al., 2013a, 2013b, 2014, 2017, 2022). This means that the research included cooperation, competition or negotiation between two agents that were not part of a population. Therefore, it remains an open question if the advantages of the use of ToM can also be found on a population level. Since humans are known to be social animals that generally live in groups (Gamble et al., 2014), this question is crucial for our understanding of the evolutionary development of the skill. The research by Lenaerts et al. (2024) did focus on the emergence of ToM in a population of agents, but here the population was modeled using evolutionary game theory, thereby disregarding individual characteristics of agents. The current research instead simulates an evolutionary process and gives each agent individual beliefs. This makes each agent a unique member of the population, whilst they still belong to a particular group (i.e., the order of ToM).

2.2 Agent-Based Modeling

Certain phenomena can be studied more easily than others. Whereas water evaporation can be studied simply by using a kettle, a stove and water, finding methods to study how panic spreads in a group of people is a lot more challenging. One possibility is to find video footage

of several incidents and study these, but there may not always be enough data available, and gathering new data can be unethical. The latter example therefore highlights the difficulty of capturing and studying the behavior of a group of individuals. This section presents Agent-Based Models (ABMs) as a research tool (Section 2.2.1), highlights previous research using ABMs (Section 2.2.2), and discusses agent communication in ABMs, focusing specifically on negotiations (Section 2.2.3).

2.2.1 Agent-Based Modeling as a Research Tool

ABMs can serve as a tool to study phenomena that are difficult to recreate (Axelrod, 1997, as cited in Bryson et al., 2007). This computational tool can be used to simulate environments with agents in it. The agents are autonomous entities that attempt to fulfill their objectives (Wooldridge, 2009). Agents can interact with each other and with the environment. From these interactions, real-life phenomena may emerge. This way, ABMs can be used to study real-life phenomena through simulations.

The term 'multi-agent system' is regularly used interchangeably with 'agent-based model' (Niazi & Hussain, 2011). Although both belong to the same field of agent-based computing, ABM as a term is used more widely than multi-agent systems (MAS), since the former is used in fields ranging from ecology to economics, whilst the latter is especially used in the context of technology, mainly Artificial Intelligence (AI). MAS focuses on the interaction of different agents and the social processes emerging due to those interactions (Wooldridge, 2009). It is therefore a useful tool to simulate populations and societies. The current research regards only ABMs in the context of AI, and with multiple agents. Therefore, in the context of this research, ABM refers to simulations that can also be considered MAS.

According to the principles of Hogeweg (1988) for 'good' simulations, there is no need for the simulation to be a replica of reality. Instead, it should capture the patterns of life that are of interest in the research. Even with a simple model, the interaction between multiple agents and between agents and their environment can lead to unexpected emergent effects (Hemelrijk, 1999). Since the agents are generally much simpler than the real entities that they represent (e.g., animals), their behavioral patterns are in a way exaggerated and therefore the resulting emergent patterns can be observed closely. ABMs are therefore bottom-up models (Klügl & Bazzan, 2012). When the same patterns are validated by empirical evidence, existing hypotheses can be tested and new hypotheses can be generated.

2.2.2 Previous Research using Agent-Based Modeling

Over the past 60 years, the use of ABMs for research about populations of animals has become more and more common. Schelling (1969) was one of the first scientists who used an ABM for his research about residential segregation. It was not until 1996 however, when Epstein

and Axtell published their book *Growing artificial societies: Social science from the bottom up* (Epstein and Axtell, 1996, as cited in Klügl and Bazzan, 2012), that ABM became a more commonly used research tool. In their book, Epstein and Axtell introduce Sugarscape, which is an ABM where agents harvest sugar to survive. Even though the model is simple, it managed to generate emergent behavior that resembled real social phenomena. This showed the significance of ABMs as a research tool.

Some of the most influential work using ABMs simulating animal populations comes from Hemelrijk. She used ABMs to study, among others, various types of primates, fish schools, and birds. Her models are simple, adhering to the principles of Hogeweg (1988), but allowed her to generate and test hypotheses for various phenomena in animal populations. Hemelrijk (1999) presents a model of a population of agents that 1) tend to group together, and 2) that can win or lose dominance interactions with other agents. The agents use their previous experience for new encounters to estimate if the risk is low enough to approach. The encounters shape the behavior of individuals and thus the social environment, which in turn changes the structure of the society. The model shows that despotic and egalitarian societies emerge as a result of the level of aggression of the individual agents. This research illustrates that what was previously assumed to be a result of natural selection, is an emergent effect of the environment that the individuals are situated in.

Other research by Hemelrijk builds upon this model. Gradually, more features are added to test increasingly complex hypotheses. An example is Hemelrijk (2002) proposing that the variance in intersexual dominance between various primate species may be the result of the difference in cohesiveness in the grouping of individuals. Furthermore, Hemelrijk et al. (2008) used a model to illustrate the winner-loser effect. Puga-Gonzalez et al. (2009) add an anxiety level to individuals in the model to motivate the grooming of other individuals; this is the model that was introduced in Section 2.1.1, where affiliative effects emerged which are also observed in real primates.

Van der Vaart and Hemelrijk (2014) suggest using ABMs to overcome the lack of consensus regarding ToM research in animals. They encourage creating models that simulate individuals as embodied and embedded in their environment, and use this to study the emerging patterns. The environment can constrain the individuals in such a way that behavior is observed that seems to result from the use of ToM, while in reality, the behavior is a direct consequence of the complex dynamic environment (Van der Vaart & Hemelrijk, 2014). These relations can be observed using ABMs. One example is the grooming model mentioned in the previous paragraph, by Puga-Gonzalez et al. (2009). Another example is research by Van der Vaart et al. (2012), who used an ABM to show that the re-caching behavior of corvids is not necessarily evidence of ToM, but can alternatively be explained by the stress levels of the birds.

ABMs are not just useful for studying animals and their ToM, they are also a useful tool for testing hypotheses regarding the evolutionary development of higher-order ToM in humans. ABMs were used to study the advantage of using ToM in various settings:

Competitive settings: De Weerd et al. (2013b) tested the Machiavellian intelligence hypothesis by simulating four competitive two-player games, played by agents capable of ToM0, ToM1, ToM2, ToM3, or ToM4. The results indicate that ToM1 and ToM2 benefit agents playing competitive games, whilst the additional advantage of ToM3 and ToM4 is limited. Furthermore, Devaine et al. (2014) use an ABM to investigate whether agents with various orders of ToM survive hide-and-seek. They conclude that higher-order reasoning does increase the survival chances of the agents, but the use of ToM2, ToM3, or ToM4 is not necessarily better than the use of ToM1.
Cooperative settings: De Weerd, Verbrugge, and Verheij (2015) tested the Vygotskian intelligence hypothesis by simulating a two-player cooperative communication game with an information Sender and a Receiver, again played by agents of varying orders of ToM. The results show that ToM1 and ToM2 allow agents to establish communication more quickly than ToM0. However, the benefit of higher-order ToM depends on the role of the agent in the game (i.e., Sender or Receiver). The results furthermore suggest that in some cases, lower orders of ToM allow for more effective cooperation.
Mixed-motive settings: De Weerd et al. (2013a), De Weerd et al. (2017) and De Weerd et al. (2014) each test the Mixed-motive interaction hypothesis by using a two-player negotiation framework called Colored Trails. Again, different agents used different orders of ToM. The results of De Weerd et al. (2013a) and De Weerd et al. (2017) indicate that ToM0 agents might perform well, depending on their partner. ToM1 agents understand the need for a mutual agreement and thus prevent the negotiation from halting. ToM1 agents therefore do not necessarily outperform ToM0 agents. ToM2 agents can find the best possible agreement for themselves by finding the best mutual agreement. The results of De Weerd et al. (2014) indicate that ToM1 and ToM2 agents benefit from their ToM, and their trading partners do too. Furthermore, the use of ToM4 allowed agents to increase their personal gain, contrasting with the results of De Weerd et al. (2013b) in competitive settings. However, Devaine et al. (2014) found that in the battle of the sexes setting, ToM1 and ToM2 agents performed better than ToM3 and ToM4 agents. A recent research by Lenaerts et al. (2024) explored the development of ToM in the Incremental Centipede Game using evolutionary game theory. ToM evolved in the agents, allowing them to cooperate with their partners, causing a higher reward for both of them.

ABMs thus showed that ToM can benefit individuals in competitive, cooperative, and mixed-motive settings. Cooperative and competitive tasks are both tasks that especially primates like apes face on a daily basis, but this apparently did not trigger the development of higher-order ToM in these species. The aforementioned research therefore suggested that the Mixed-motive interaction hypothesis seems most likely to explain the evolutionary development of higher-order ToM in humans, even though this is still a topic of debate. The existing models that were used to test the hypothesis include mainly pairwise single-shot interactions, whilst the answer to the question of how and why humans developed higher-order ToM may lie in the population-wise dynamics resulting from these interactions.

2.2.3 Agent Interactions

The aim of the use of ABMs is to see the effect of interactions between individual agents on the population as a whole. Agents interact through some form of *communication*. This communication serves as a means to reach a certain goal, like gaining/sharing information or finding an agreement in a negotiation (Wooldridge, 2009). Agent communication can take many forms, although within AI we generally focus on verbal communication through the use of a communication language. We thus disregard other means of communication, like the human ability to use facial expressions to convey opinions, since it is not straightforward to model nonverbal communication in an ABM (Wang & Ruiz, 2021).

Communication languages can be used to set up verbal communication between agents. Examples of such languages are KQML and FIPA ACL. Both are based on the principles of speech act theory (Austin, 1962; Searle, 1969): sending a message is seen as performing an action. According to the speech act theory, communication is divided into categories like request, inform, and promise. KQML (Finin et al., 1996; Patil et al., 1992, as cited in Wooldridge, 2009) provides a format for communicating messages, distinguishing between different categories. FIPA ACL (FIPA, 1997, as cited in Wooldridge, 2009) is similar to KQML in that it provides a format for messages, but its performatives, i.e., communication categories, are different. Furthermore, FIPA ACL provides more detailed formal semantics than KQML. Both communication languages facilitate agents to share or receive knowledge, which can alter their mental state (Wooldridge, 2009). In addition to these communication languages, messages can also be sent in the form of interactions, where agents signal their desires by performing a certain action. In this research, the latter type of communication is used. The mentioned communication languages thus merely function as background knowledge on communication types.

Communication is essential in several social situations, one of them being negotiations. In a negotiation, two or more agents try to reach an agreement regarding some exchange of items/services (Kraus, 2001). An example of such a negotiation is the bread example detailed in Section 2.1.3, where a seller wants to earn money, and a buyer wants to receive the bread while paying as little as possible. Agents in a negotiation setting like this example need to communicate to convey their proposals and to respond to the offers of the other party (Lewicki et al., 2020). Negotiations are a popular topic of research within AI and MAS due to the increasing importance of automated negotiation, which can be faster and may result in better agreements than human negotiation (Baarslag et al., 2017; Jennings et al., 2001).

Formally, negotiations consist of a negotiation set, a protocol, a collection of strategies, and an agreement deal (Wooldridge, 2009). The negotiation set consists of all the proposals that agents can make, and the protocol is the subset of allowed proposals. Each agent has a set of strategies that it can use to decide on its proposal. Finally, the agreement deal refers to the outcome of the negotiation. An agreement cannot always be reached immediately. Therefore, a negotiation usually comprises several rounds. An example of a negotiation protocol to structure

the negotiation is the alternating offers model (Osborne, 1990; Rubinstein, 1982), in which two agents negotiate by altering which agent makes the offer, and which agent accepts or rejects. The outcomes of negotiations between pairs of individuals impact their future encounters and thus the dynamics of the population. ABMs (see Section 2.2) are therefore a useful tool to model negotiations. The other way around, negotiations are also a useful tool to resolve conflicts of agents in ABMs (Wooldridge, 2009).

2.3 Evolution

Evolution is a concept referring to the theory that animal species are dynamic and that they have developed from ancient species, gradually changing over time (De Lamarck, 1809, as cited in Nowak, 2006). According to the theory, species are forced to adjust to their environments to survive, which is what catalyzes the gradual change in their internal and external characteristics. This is how complex cognitive skills, like ToM, can develop in a species. The principles of evolution contrast with the view that God created the Earth including all animal species, and that these species are static. This section gives a brief overview of the history of Darwinism (Section 2.3.1) and discusses the characteristics of an evolutionary process, which can be used to simulate evolution and study the evolutionary development of a skill (Section 2.3.2).

2.3.1 History of Darwinism

Before the 18th century, the generally accepted worldview was based on the Bible (Ray, 1691, as cited in Bowler, 2000). According to the Bible, God created the Earth around 4000 BC, including all of its flora and fauna. The animal species that were created by God were made to survive all conditions on Earth and remained the same over time. All species, including humans, are thus static according to this theory.

In the 19th century, Jean-Baptiste Lamarck was the first person to create a theory of biological evolution, which contrasted with the generally accepted idea at the time (De Lamarck, 1809, as cited in Nowak, 2006). He believed that species can initiate their own 'improvements', and additionally that their environment forces them to change. This theory was backed by evidence from research of the time: paleontologists found fossils that were older than 6000 years, and that must therefore have existed before the creation of the Earth (Bowler, 2000). These findings suggested that the Earth was much older than assumed at that time. Furthermore, the fossils showed that certain species had gone extinct, and others had changed: both conclusions contradicting the generally accepted worldview (Pojeta & Springer, 2001). This led Lamarck to develop his theory of dynamic species.

With his theory, Lamarck paved the way for Charles Darwin. Darwin's ideas were inspired by several researchers other than Lamarck, such as Condorcet, Linnæus, E. Darwin, Lyell,

and Malthus (Avery, 2003). In 1859, Darwin published a book: *On the Origin of Species* (Darwin, 1859), in which he formulated his beliefs. Darwin proposed that changes in species are unintentional and happen due to chance, a concept which is known as natural selection, as opposed to artificial selection (Darwin, 1859, as cited in Nowak, 2006). Alfred R. Wallace had also sent Darwin a paper, *On the Tendency of Varieties to Depart Indefinitely from the Original Type*, in which he discussed ideas similar to Darwin's beliefs (Wallace, 1858, as cited in Avery, 2003). Together they are seen as the pioneers of the evolution theory as we know it today: the Theory of Evolution by Natural Selection. This theory is sometimes also referred to as Darwinian Evolution or Darwinism.

2.3.2 Characteristics of an Evolutionary Process

Darwinian Evolution is characterized by the following key ideas: populations can reproduce, all individuals part of a population have descended from the species that existed before them, and genetic diversities in new generations emerge due to mutations (Bowler, 2000; Darwin, 1859). 'Fitter' individuals are the individuals that are more suited to their environment. A population/species with fit individuals therefore has bigger chances of surviving and as a consequence, its individuals also have bigger chances of reproducing. Therefore, favorable traits are developed over time, while unfavorable traits disappear. New traits like cognitive skills are thus developed only when the demand is high enough to outweigh the cognitive demand it takes to use such a skill (Aiello & Wheeler, 1995). If a species fails to adapt to the environment, it may become distinct.

Darwinism not only offers an explanation for the development of modern species, but it also sets forth a set of conditions for evolutionary processes. The three components of an evolutionary process are selection, replication, and mutation. The mathematical biology professor Nowak states that "Wherever information reproduces, there is evolution" (Nowak, 2006). The key features of evolutionary processes all have a mathematical nature. Therefore, according to Nowak, all evolutionary processes can be characterized by mathematical formulae to analyze evolutionary dynamics. This mathematical nature of evolution argues for the use of simulations to study the evolution of complex cognitive skills, such as ToM.

Chapter 3

Methods

This chapter describes the methods that were used to answer the research question. The simulation environment in which the experiments were conducted is described in Section 3.1. Then, the agents are detailed in Section 3.2. This includes a discussion of their movement, the logic of the negotiations used to trade resources, and their Theory of Mind (ToM). The evolutionary process established in the environment is described in Section 3.3. Finally, the experiments are described in Section 3.4, followed by a hyperparameter overview in Section 3.5, and the implementation in Section 3.6,

3.1 Environment

The environment that was created consists of a simple square arena (600×600 tiles), which is the area in which agents exist and can move. A screenshot of the arena can be found in Figure 3.1). Within the environment, there is a notion of time, which is managed using 'ticks': each tick represents one time-step in the environment. The environment is fully accessible to all agents, with no obstacles or objects located in it apart from other agents. The environment contains a population of n agents, where n can be varied by the user. In this research, this number was set to 60, which was chosen based on a qualitative study [1]. This number of agents allows them to move freely through the environment, whilst still leading to frequent negotiations for all agents.

More information about the agents can be found in Section 3.2.

[1] The values 30 and 90 were also tested. The results were different from those for 60 agents. This is therefore a design choice that influenced the results, and it is discussed in the discussion section.

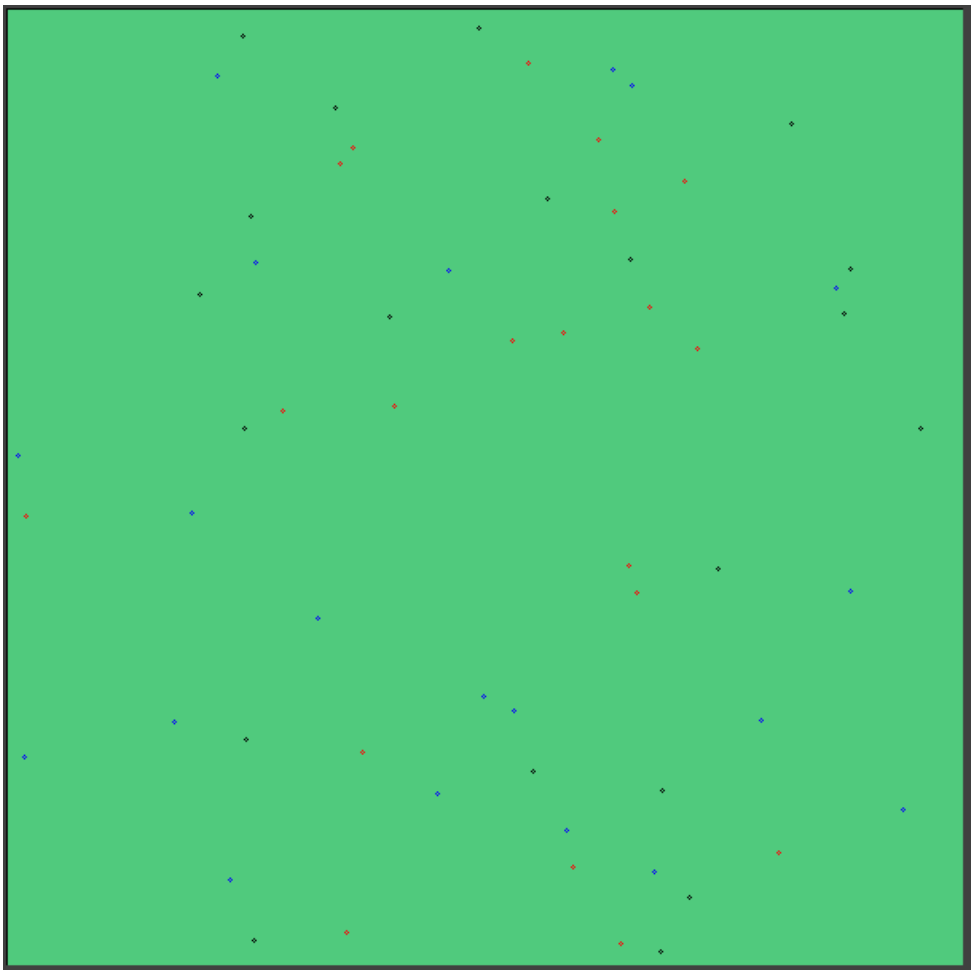

Figure 3.1: The arena (green square) in which the agent population is situated, where the black surrounding indicates the boundaries of the arena. The arena includes 60 agents, where the black dots represent the ToM0 agents, the red dots the ToM1 agents, and the blue dots the ToM2 agents.

3.2 Agents

Each agent can move through the environment (Section 3.2.1). The agents have resources and can use these to negotiate with other agents that they encounter (Section 3.2.2). There exist three types of agents, each reasoning with a different order of ToM (Section 3.2.3). No ToM orders higher than two were implemented, to limit the computational complexity, and to resemble the research by De Weerd et al. (2017) as closely as possible. After initialization, there are $n/3$ agents of each type. This distribution over the agent types may change over time due to the evolutionary process in the environment, whilst n always remains the same.

3.2.1 Movement

The initial position of each agent is random, with the only restrictions being that the location is within the bounds of the arena and that it is not yet occupied by another agent. The agents can move through the environment by taking one step every tick. Their initial direction is random. They continuously move in that direction unless they either encounter the edge of the arena, in which case their direction changes randomly, or they encounter an agent within negotiating distance, in which case they may start to negotiate (see Section 3.2.2). The negotiating distance is a constant that may be varied by the user. For this research, it was set to 10px. The radius of an agent is 2px, and the distance is measured from the center of the agents. Therefore, agents will interact when there is a maximum of 6px between the two agents. This value was chosen to resemble an average talking distance between two humans. If the agent that they encounter is already negotiating with someone else, the agent changes its direction randomly to avoid the negotiating pair.

3.2.2 Resources and Negotiations

The mixed-motive aspect of the environment lies in the negotiations between individuals. Previous research by De Weerd et al. (2017) used the game Colored Trails (CT) as a mixed-motive setting. In this game, two agents negotiate about a redistribution of chips. The agents need certain chips to buy a path toward their goal on the game board. The current research does not use the CT setting. Instead, agents in the environment have resources that they need to trade to survive. For these trades, negotiation is used.

The choice of using a trading system instead of CT games between pairs of agents was made to make the environment as realistic as possible, as barter existed long before currencies were introduced (Anbugeetha & Nandhini, 2021). This section first describes the logic of the negotiations used for trading, followed by a formal definition. The (formal) model of negotiation used for this research was largely inspired by that of De Weerd et al. (2017), which they implemented for the CT setting.

Each agent can possess four types of resources. Of each type, they can carry a maximum of four (see Table 3.1 in Section 3.5 for an overview of all parameters). Each agent has the ability to produce one of the four resources themselves. Which resource this is, is arbitrary, but consistent throughout the run. At the start of each negotiation, the agent will have four of its producing resource. More details about the production interval of this resource can be found in Section 3.3. The other resources can be acquired through negotiation with other agents. This negotiation is essential since agents need to reach a resource threshold of two for each resource type within 2500 ticks in order to survive (see Section 3.3). The number of items of the producing resource is virtually limitless, so the threshold for this resource is always met. For the other resource types, the agents start with one resource per type. They thus need to gather at least one additional item of each resource type in order to reach this threshold. If they do not manage to do that before the time limit, they die.

The described parameter values can be altered by the user, except for the number of resource types (see Section 3.5 for an overview of the hyper-parameters). The number of resource types was set to four as a balance between the computational complexity and the negotiation possibilities: it allows for a wide range of offers, whilst the complexity is still reasonable. The more resource types, the more offers need to be evaluated for each negotiation round and thus the higher the computation time. The same holds for the maximum number of resources per resource type. For this parameter, also a value of four was chosen. This, in combination with the resource threshold of two (see Section 3.3), gives agents the possibility of having spare resources that can be offered. Increasing this maximum capacity increases the computational complexity. It is important to note, however, that this limit of four resource types, four resources per type, and the resource threshold of two affect the results. If the limit had been higher, there would be more offer possibilities, which in turn increases the chances of a successful negotiation. For this research, the number of initial resources per type was set to one to pressure the agents to negotiate in order to survive. See Section 3.3 for further details about the evolutionary process.

Negotiation is thus the key element of the behavior of the agents. Whenever another agent is within the negotiation threshold of 10, the pair of agents can start a new negotiation. This new negotiation partner cannot be their most recent negotiation partner to avoid infinite negotiation loops. One of the two agents makes the initial offer. This means that this agent uses its ToM to make an offer or withdraw immediately. The initial agent is chosen randomly, because the type of ToM that is used for the first offer oftentimes impacts the course of the following offers. To avoid agents of a certain ToM order always being the initiator, this decision was randomized.

When an agent receives an offer, it can choose to accept the offer, propose a counteroffer, or withdraw from the negotiation. Which of these options the agent chooses depends on the state of the environment and the order of ToM that the agent uses, and this is described in Section 3.2.3. The negotiation then consists of a sequence of alternating offers between the two agents. It ends when one of the agents withdraws, accepts an offer, or when a maximum number of

rounds is reached. Each round lasts for one tick. The maximum number of rounds/ticks is implemented to avoid infinite loops, where agents keep going back and forth without reaching a consensus. For this research, this number was set to 50, as can be seen in Table 3.1. This value was chosen based on a qualitative observation of multiple test runs. The maximum was first set to an unrealistically high number (10000). This showed that the negotiations that did terminate, all terminated within 35 rounds. Therefore, 50 was taken as the maximum.

There are two restrictions on the offers that an agent can make. Firstly, the offer should be possible, meaning that the resources it asks from the trading partner are owned by the trading partner. Secondly, an agent can only offer excess resources, meaning that it cannot offer resources from which it has less than the resource threshold of two. This threshold needs to be reached within 2500 ticks in order to survive, and agents thus are prevented from taking an unnecessary risk by offering a non-spare resource. An exception to this second condition is the producing resource r, for which it does not matter if it decreases beneath the threshold.

We adapt the formal definition of negotiation that is given in De Weerd et al. (2017). The negotiation is a tuple $\langle \mathcal{N}, \mathcal{D}, \mathcal{R}, \pi_i, \pi_j, D_0 \rangle$, where:

- $\mathcal{N}$ = the set of agents participating in the negotiation: $\{i, j\}$;
- $\mathcal{D}$ = the set of possible distributions of resources, based on the current resources of i and j;
- $\mathcal{R}$ = the set of possible producing resources;
- π_i, $\pi_j : \mathcal{R} \times \mathcal{D} \rightarrow \mathbb{R}$ = the score functions such that $\pi_i(r, D)$ denotes the score of agent i when its producing resource is $r \in \mathcal{R}$ and the resources are distributed according to distribution $D \in \mathcal{D}$. The score function is defined in the next paragraph;
- $D_0 \in \mathcal{D}$ = the initial distribution of resources over i and j.

The alternating offers can be represented as a sequence of offers $\{O_0, O_1, O_2, ...\}$, where O_0 is the initial offer. An agent can thus receive an offer O_t, which it may counter with counteroffer O_{t+1}. The negotiation ends when either of the agents accepts a received offer O_t, either of the agents withdraws, or when the negotiation limit of 50 rounds is reached.

A notational difference with the negotiation model of De Weerd and colleagues is that in our research $\mathcal{R}$ is used instead of L. L represented the set of possible goal locations, whereas the $\mathcal{R}$ in this research represents the set of possible producing resources. The utility of a state in this environment thus depends only on what resource the agents produce themselves $r_i \in \mathcal{R}$ and not on a goal location l_i. Apart from this (mainly notational) difference, the main difference between the model of De Weerd and colleagues and the one presented here is the score function π. The function was redesigned to fit the new environment. In this research, there is no time penalty in the utility function like there is in the model of De Weerd and colleagues. This

was removed due to the nature of the setting of this research: agents need to gather enough resources of each type before they are 'checked': the environment thus already reflects the time pressure on the negotiations without specifically adding this to the utility function. Another important difference is that in the current research, a negotiation may not have the possibility of a successful outcome, whilst those modeled by De Weerd and colleagues were always at least potentially successful. This is a consequence of the varying resources that two agents bring to a negotiation.

The score function π should capture the need for reaching the threshold for each of the four resource types. This is more important than having a high total number of resources. Therefore, a shortage weighs heavier than a surplus. The π function is defined as follows: for each shortage the agent receives -2 points, and for each surplus it receives +1. For the producing resource r, the agent always receives 0 points. This definition leads to $\pi \rightarrow [-12,6]$. This is mapped to $[0,18]$ by adding 12 points to remove negative numbers, which are inconvenient in the beliefs computations that are described in Section 3.2.3. This can be compared to the CT settings by considering a game board where each step towards the goal location is worth two points, and each additional unused chip is worth one point. Furthermore, each player has one token that has no value to them. Example 1 provides an example of the use of the function.

From this definition of the utility function, it follows that agents are indifferent to owning their producing resource. They do not receive a penalty for having little of this resource, nor do they benefit from having four of this resource. Within the negotiation, this may result in the agents accepting offers where they only give a number of items from their producing resource, without receiving anything in return. This is a design choice that was made to stimulate free trade, where agents own at least one resource with which they may make an acceptable offer. How various agents reason about an offer like this is detailed in the corresponding ToM Section 3.2.3.

Example 1: Let's take an example to illustrate the use of the π function. Assume we want to know the utility of a state of agent i after accepting offer O_t. Assume that O_t leads to the distribution D $[1,4,0,4,2,3,1,3]$. This means that after accepting the offer O_t, i will have the following resources: $[1,4,0,4]$, i.e., one of resource 1, four of resource 2, zero of resource 3 and four of resource 4. See Figure 3.2 for a visualization of the state of the agent. Assume that the producing resource of i, r_i, is resource type 2. Then the utility $\pi_i(2,[1,4,0,4,2,3,1,3]) = -2+0-4+2 = -4$. Mapping this to [0,18] gives $\pi_i(r,D) = 8$.

Figure 3.2: A visualization of the state of an agent, contrasting the current resources with the threshold to stay alive.

Following De Weerd et al. (2017), each agent knows its own producing resource r, but not that of other agents. Additionally, they do know the resources that others have. In other words: we make the assumption that the set of possible offers is fully observable, but agents do not have complete information about the preferences of other agents. The decision that agents can observe the resources of others was made for two reasons. Firstly, it is less computationally expensive than examining all offers, including the ones that are not actually possible because the trading partner does not have the required resources. Secondly, the focus of this research is on the adaptation of negotiations with ToM from pairwise settings to a population setting. It is therefore important to imitate previous implementations of negotiations as closely as possible. In the CT setting, the number of chips in the game is constant and thus the distribution of them is observable to both agents.

3.2.3 Theory of Mind

The agents use Theory of Mind (ToM) to determine their move in the negotiations. Which order of ToM they use is determined during their initialization. Agents can have a ToM order of zero, one, or two. The model for ToM that was used in this research was based on De Weerd et al. (2017). Due to the difference in environment, there are some key differences between the ToM model in this research and that of De Weerd and colleagues. Whenever this is the

case, it is explicitly stated. Contrary to De Weerd and colleagues, this paper does not always omit the notation of variables when they can be inferred from the context. Instead, all variables are included, or, in the case that this leads to a lack of space, they are mentioned underneath the formula. Despite those differences, it is still recommended to read the paper by De Weerd and colleagues for relevant background information. Furthermore, De Weerd and colleagues contains relevant examples of the behavior of agents with various orders of ToM.

Zero-order ToM

The zero-order agents, ToM0 agents, are unable to reason about the mental content of others. They therefore do not consider the intentions or desires of others, and make decisions and offers based solely on their personal gain. However, even though they do not model what their trading partner wants to achieve, ToM0 agents can use the response of the trading partner as an indication of how likely it is that other offers will be accepted. This is known as their zero-order beliefs.

There are two types of zero-order beliefs: the beliefs that are updated across the negotiations (*initialBeliefs*), and beliefs that are updated within the negotiation ($b^{(0)} : D \rightarrow [0,1]$). Contrary to De Weerd and colleagues, this paper indicates these two with different names to highlight the difference between the two. The former is more general, as it contains the beliefs that an offer that proposes an x number of resources for an y number of resources will be accepted. The latter is specific to the current negotiation, as it contains the beliefs that an offer that proposes an x number of a specific resource type for an y number of a specific resource type will be accepted by this trading partner. Initially, each negotiation starts with a $b^{(0)}$ that is equal to *initalBeliefs*. Then, $b^{(0)}(O)$ is the belief that offer $O \in \mathcal{D}$ will be accepted by the trading partner. More details about these beliefs can be found in Section 3.2.3.

Which offer a ToM0 agent will make is decided based on the expected value of that offer, which is the utility that the agent thinks it will receive when making that offer $O \in \mathcal{D}$. For this computation, the zero-order beliefs are used, meaning that the EV of an offer decreases when the zero-order belief of this offer decreases. This expected value, EV, is computed using this formula:

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

where $b^{(0)}$ are the zero-order beliefs of the agent, and π, r_i and D_0 are as described in Section 3.2.2. Note that this formula differs from the formula provided by De Weerd et al. (2017) in two ways: the use of r_i instead of l_i, and the omission of the time index, both of which are described in Section 3.2.2.

Offers with a high EV are more beneficial for the agent than offers with a low EV score, and thus if the agent wants to make an offer it will choose (one of the) offers with the maximum

EV. In other words, the agent assesses all offers $O \in \mathcal{D}$ and chooses offer O_t^* where

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

However, there are situations in which it is not beneficial for the agent to make this (counter-) offer O_t^*. There is one case in which the agent has two available actions, which is during the initial round: in this round, the agent can make the initial offer or withdraw. In the other rounds, there are three available actions to the agent: it can make a counteroffer, but it can also accept the offer from the partner O_{t-1}, or withdraw from the negotiation. The agent will accept the offer from the trading partner if it is at least as good as O_t^*, and it will withdraw from the negotiation if there exists no offer that is expected to result in a higher score. Which of the actions the agent takes can be computed using the following function:

$$ToM0_i(O_{t-1}, r_i, b^{(0)}) = \begin{cases} O_t^* & \text{if} \quad EV_i^{(0)}(O_t^*, r_i, b^{(0)}) > \pi_i(r_i, D_0) \text{ and} \\ & \quad EV_i^{(0)}(O_t^*, r_i, b^{(0)}) > \pi_i(r_i, O_{t-1}) \\ \text{accept} & \text{if} \quad \pi_i(r_i, O_{t-1}) > \pi_i(r_i, D_0) \text{ and} \\ & \quad \pi_i(r_i, O_{t-1}) \geq EV_i^{(0)}(O_t^*, r_i, b^{(0)}) \\ \text{withdraw} & \text{otherwise.} \end{cases} \tag{3.3}$$

In this function, $EV^{(0)}$ refers to Equation 3.1.

Equation 3.3 allows the ToM0 agent to take part in negotiations without having the ability to reason about the intentions of its trading partner. Due to the zero-order beliefs, the ToM0 agent can propose offers that are beneficial to the trading partner as well, even though the agent does not explicitly model that this is the case. An example of the behavior of a ToM0 agent can be found in Example 2. A ToM0 agent may accept an offer in which it needs to give a number of items from its producing resource r without receiving anything in return. This can be interpreted as the agent not caring about this resource in any way, and thus not minding it gone. A comparison is when a Dutch person has some foreign coins left, from a country it will not visit in the near future. The coins hold no value to the owner, but they do hold value for another person. However, since a ToM0 agent cannot reason about how other value resources, it is a rational decision to accept an offer where it gives away its producing resource. Note that this can only happen when there are no better offers available to the agent, so in practice, it rarely happens.

Example 2: ToM0 agents do not reason about the mental content of trading partners. Let us consider a situation where two agents are negotiating. We look at the negotiation from the point of view of a ToM0 agent, i, who is choosing an initial offer to make to its trading partner j. Figure 3.3 shows an example situation, where i has a total of ten resources: two

of type 1, three of type 2, one of type 3 and four of type 4 (represented here are food, water, tools, and materials respectively). Its producing resource is type 4, as indicated by the underlining.

If i would not have initialized zero-order beliefs *initialBeliefs* at the start of the negotiation, it will simply ask for everything it needs and can get from j, as that would maximize π_i. Ideally, i would like to have four items of every resource type. However, when looking at the resources of j it sees that that is impossible to achieve. Therefore, i asks for every resource that j has that i still needs to reach four of every resource type. Note that agents will never ask for their producing resource since they always have four at the start of a negotiation, meaning that their maximum capacity for the resource is reached. The production logic of the resources can be found in Section 3.3.

This example highlights the importance of zero-order beliefs of ToM0 agents. Without those, their offers are very unlikely to be accepted, since they will ask for a lot, but won't be offering to give anything themselves. Since this is an initial offer, $b^{(0)}$ is not updated yet and reflects the *initialBeliefs*. When ToM0 agents use their *initialBeliefs* they know from experience that offers like this, giving zero and receiving five (i.e., 0:5), are not likely to be accepted. More information about the offer types can be found in Section 3.2.3, in Example 4.

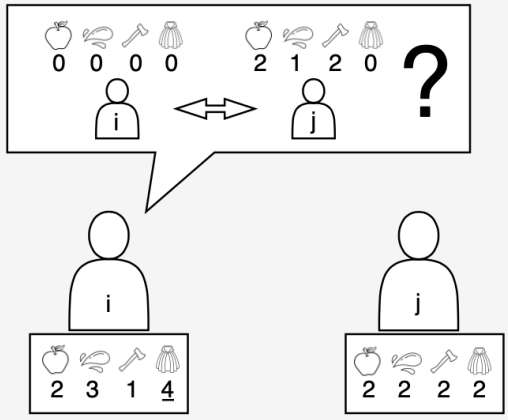

Figure 3.3: Visualization of an offer by a ToM0 agent (i) without (initialized) zero-order beliefs. The producing resource of i is resource type 4: the materials. In this case, agent i asks to be given five resources, while offering nothing in return.

Figure 3.4 shows the same situation, except now the ToM0 agent has zero-order beliefs. It is still the initial offer, so the beliefs that are used are the general *initialBeliefs*. i comes

up with (an example of) a different offer. This time, it asks for one resource of type 3 (tools) and gives one resource of type 4 (materials). It offers this because i knows that generally, an offer of the type 1:1 is likely to be accepted. In the case that this offer is rejected, i updates its $b^{(0)}$: apparently j is not interested in giving resource type 3.

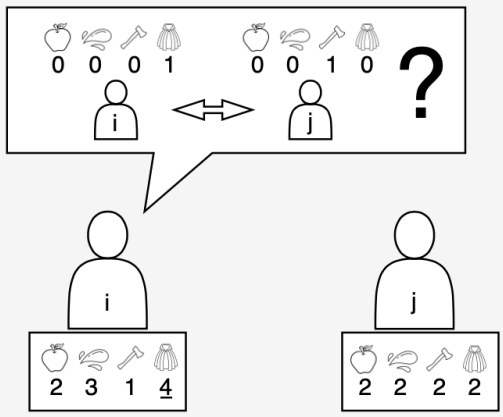

Figure 3.4: Visualization of an offer by a ToM0 agent (i) with (initialized) zero-order beliefs. The producing resource of i is resource type 4: the materials. In this case, agent i offers one of its producing resource for one resource that it is missing.

Note that the situation that was discussed here serves only as an example of the behavior of a ToM0 agent, but that in reality, the trading partner j would always have four items of at least one of the resource types (since it has a producing resource).

First-order ToM

The first-order agents, ToM1 agents, do have the ability to model the mental content of other agents. This allows them to reason about the desires of the trading partner. The ToM1 agent therefore considers what benefit the trading partner receives from the offers as well, instead of focusing solely on their own gain. To do so, the agent looks at the negotiation from the perspective of the trading partner. It then reasons what action the ToM1 agent itself would take if it were in the position of its trading partner. It then uses this information to determine which offer O it should make that is beneficial to itself, but at the same time is likely to be accepted because the trading partner benefits from the offer as well.

Three factors make it difficult for the agent to predict the action of the trading partner: it doesn't know the zero-order beliefs, the producing resource, and the order of ToM of the trading partner. For this reason, ToM1 agents do not only have zero-order-beliefs ($b^{(0)}$ and

initialBeliefs), but also first-order beliefs: $b^{(1)} : D \rightarrow [0,1]$. These beliefs represent what the ToM1 thinks the $b^{(0)}$ of the trading partner are. $b^{(1)}(O)$ thus represents what the ToM1 agent thinks the beliefs of its trading partner are about the ToM1 agent accepting offer O. It can then use these first-order beliefs for the computation of what action it would take in the position of the trading partner. The ToM1 agent computes the first-order beliefs by reflecting on its own actions, and reasoning how it would update its own $b^{(0)}$ if it were the trading partner. More information about these belief updates can be found in Section 3.2.3.

Furthermore, the agent also models the expected producing resource of the trading partner, which is updated based on the offers it receives from this trading partner. This is modeled as a probability for each of the four resources: $p^{(1)} : R \rightarrow [0,1]$. $p^{(1)}(r)$ then indicates how likely the agent finds it that r is the producing resource of the trading partner. Additionally, it uses a confidence score that represents its confidence in the first-order ToM: $c_1 \in [0,1]$. If the predictions made by the ToM1 agent are incorrect, this confidence will decrease, and the agent will simply use ToM0. More about the model of the producing resource of the trading partner and the confidence score can be found in Section 3.2.3.

The predicted action of the trading partner is used by the ToM1 agent to compute the expected value (EV) of making offer O. This $EV_i^{(1)}$ is computed using the following function:

$$EV_i^{(1)}(r_j, O, r_i, b^{(1)}) = \begin{cases} \pi_i(r_i, D_0) & \text{if } ToM0_j(O, r_j, b^{(1)}) = \text{ withdraw,} \\ \pi_i(r_i, O) & \text{if } ToM0_j(O, r_j, b^{(1)}) = \text{ accept,} \\ \max\left\{\pi_i(r_i, \hat{O}_t^{(1)}), \pi_i(r_i, D_0)\right\} & \text{otherwise,} \end{cases} \tag{3.4}$$

where r_j is the (expected) producing resource of the trading partner and

$$\hat{O}_t^{(1)} = ToM0_j(O, r_j, b^{(1)}) \tag{3.5}$$

is the counter-offer that the ToM1 agent thinks the trading partner will make. Furthermore, note that ToM0 refers to Equation 3.3.

Equation 3.4 shows that the prediction of the action of the trading partner influences the EV of the offer O that the ToM1 agent evaluates. If the trading partner is expected to withdraw, the resulting EV is simply the EV score of the current state D_0. If the trading partner is expected to accept offer O, the EV is the utility of the state after the trade. If the trading partner is expected to counteroffer, this counteroffer is used to compute the EV. This function shows that the ToM1 agent looks ahead one step in the negotiation. It considers the response of the trading partner in the next timestep $t+1$ and uses this to choose an offer in the current timestep t.

As mentioned above, the ToM1 agent may not be 100% sure that its first-order predictions of the actions of the trading partner are correct. Therefore, the agent may use its zero-order beliefs instead. This is reflected in the following formula, which is used to compute the EV of

an offer O.

$$EV_i^{(1)}(O, r_i, b^{(0)}, b^{(1)}, p^{(1)}, c_1) = (1 - c_1) \cdot EV_i^{(0)}(O, r_i, b^{(0)}) + c_1 \cdot \sum_{r \in R} p^{(1)}(r) \cdot EV_i^{(1)}(r, O, r_i, b^{(1)}) \quad (3.6)$$

In this formula, $EV^{(0)}(O, r_i, b^{(0)})$ refers to Equation 3.1 and $EV_i^{(1)}(r, O, r_i, b^{(1)})$ refers to Equation 3.4. Furthermore, this formula uses the described producing resource beliefs $p^{(1)}$ and the confidence score c_1.

Equation 3.6 has a key difference compared to the computation of $EV_i^{(1)}$ by De Weerd et al. (2017). Instead of computing $EV^{(1)}$ with Formula 3.4 using $U(b^{(1)}, O)$ like De Weerd and colleagues, we simply use $b^{(1)}$. $U(b^{(1)}, O)$ refers to an update of the first-order beliefs that is made when considering an offer: the beliefs are updated to reflect how the ToM1 agent thinks the trading partner updates its zero-order beliefs. This belief-update for considered offers is omitted in this research to decrease the computational complexity of the program. Whereas De Weerd and colleagues only had two agents, this research contains 60 agents, making it infeasible to compute all these belief updates. Instead, the first-order beliefs are only updated based on offers that were made, not those that were only considered. This allows the program to compute expected counter offers only once. Note that this does remove some of the randomness that may occur when an agent chooses (randomly) between multiple offers with the same EV score. Instead, when an agent computes an expected counteroffer, it uses this same expectation for all of its computations.

Similarly to the ToM0 agent, the ToM1 agent chooses (one of) the offers with the maximum EV. That is, it picks an offer O_t^* where

$$O_t^* := \arg\max_{O \in \mathcal{D}} EV_i^{(1)}(O, r_i, b^{(0)}, b^{(1)}, p^{(1)}, c_1). \quad (3.7)$$

The agent does not always make offer O_t^*: it may also withdraw or accept the previous offer O_{t-1}. Note that the latter is only possible after the initial round of the negotiation. The function that determines which action the ToM1 agent will take is similar to that of the ToM0 agent, except for the use of $EV_i^{(1)}$ instead of $EV_i^{(0)}$:

$$ToM1_i(O_{t-1}, r_i, b^{(0)}, p^{(1)}, b^{(1)}, c_1) = \begin{cases} O_t^* & \text{if} \quad EV_i^{(1)}(O_t^*) > \pi_i(r_i, D_0) \text{ and } EV_i^{(1)}(O_t^*) > \pi_i(r_i, O_{t-1}) \\ \text{accept} & \text{if} \quad \pi_i(r_i, O_{t-1}) > \pi_i(r_i, D_0) \text{ and } \pi_i(r_i, O_{t-1}) \geq EV_i^{(1)}(O_t^*) \\ \text{withdraw} & \text{otherwise,} \end{cases} \quad (3.8)$$

where $EV_i^{(1)}(O_t^*)$ follows the notation of De Weerd and colleagues by omitting variables that can be derived from the text. $EV_i^{(1)}(O_t^*)$ is thus a shorter version of $EV_i^{(1)}(O_t^*, r_i, b^{(0)}, b^{(1)}, p^{(0)}, c_1)$.

A ToM1 agent may accept an offer in which it needs to give a number of items from its producing resource r without receiving anything in return. This happens rarely, since ToM1 agents can try to exploit the value others assign to the (for them) worthless resource. If however there are no other offers available, the ToM1 agent will accept.

Example 3: Assume we have a ToM1 agent i. When i has to make a move, either in the initial round or later on, it will consider all offers that it is allowed to make. For each offer, i considers what action it would take if i were a ToM0 agent and received that offer. It takes this predicted response of j into account when choosing which (counter)-offer to make, or whether to accept or withdraw from the negotiation.

Figure 3.5 provides a visualization of i considering making an offer O where it asks for three items of resource type 1. It predicts that j will reject O and therefore will not make this offer. Note that this differs from $b^{(0)}$ in that $b^{(0)}$ are only used in the computation of an EV score of an offer, whereas ToM1 agents really determine what their move would be if they were in the position of their trading partner. If they were to reject, the offer would receive an EV of $\pi_i(r_i, D_0)$. Due to a lack of benefit from i, O will not be chosen.

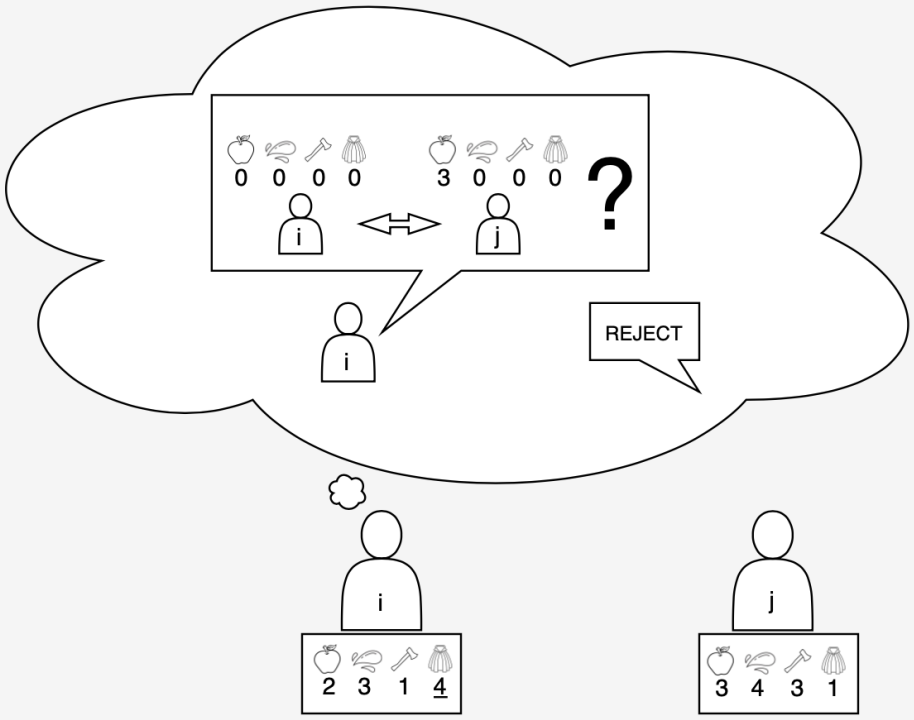

Figure 3.5: ToM2 agent *i* is considering an offer and predicts that trading partner *j* will reject this offer.

Second-order ToM

The second-order agents, ToM2 agents, also have the ability to model the mental content of other agents. Additionally, a ToM2 agent is aware that others may reason about its own intentions and goals. The ToM2 agent is thus capable of looking not just one, but two steps ahead in the negotiation. This allows the ToM2 agent to choose an action that it knows will influence the producing resource beliefs of the trading partner. It may use this for tactical moves, where it manipulates the beliefs of the trading partner so that it may receive more beneficial offers from this trading partner. In other settings, ToM2 can thus be used to manipulate the trading partner by sending 'false' information about which resource it produces, by asking for a resource that it produces. This is not possible in this setting, since all agents start a negotiation with the maximum number of items from their producing resource. This design choice allows for offers with more resources, but as a consequence, the described type of manipulation is not possible. However, ToM2 agents can use their second-order reasoning to make offers that signal what their producing resource is by asking for everything except their producing resource. A lower-

order ToM agent is unlikely to do this because it expects that the trading partner will reject the offer. ToM2 agents can reason that signaling their preferences may result in better offers from the trading partner, similar to how a human may start a negotiation by directly asking what it wants.

Using second-order ToM, the agent computes the expected value (EV) for an offer O. It does so by using Equation 3.9, which is similar to Equation 3.4, except for the use of ToM1, $b^{(2)}$ and $\hat{O}_t^{(2)}$ instead of ToM0, $b^{(2)}$ and $\hat{O}_t^{(1)}$ respectively.

$$EV_i^{(2)}(r_j, O, r_i, b^{(1)}, b^{(2)}, p^{(1)}) = \begin{cases} \pi_i(r_i, D_0) & \text{if } ToM1_j(O, r_j, b^{(2)}) = \text{ withdraw,} \\ \pi_i(r_i, O) & \text{if } ToM1_j(O, r_j, b^{(2)}) = \text{ accept,} \\ \max\left\{\pi_i(r_i, \hat{O}_t^{(2)}), \pi_i(r_i, D_0)\right\} & \text{otherwise,} \end{cases} \tag{3.9}$$

where

$$\hat{O}_t^{(2)} = ToM1_j(O, r_j, b^{(2)}). \tag{3.10}$$

This function shows the recursive nature of ToM, since $ToM1_j(O, r_j, b^{(2)})$ (short for $ToM1_j(O, r_j, b^{(1)}, p^{(1)}, b^{(2)}, 1)$ refers to Equation 3.8, which in turn calls Equation 3.3.

Similarly to the ToM1 agent, the ToM2 agent is also not certain about its (second-order) predictions of the behavior of the trading partner. The agent has a certain confidence, $c_2 \in [0,1]$ that its second-order predictions are correct. If it observes behavior from the trading partner that contrasts with the predicted ToM1 behavior, c_2 decreases. Note that the agent does not model the confidence of the trading partner, and that thus c_2 should not mistakenly be interpreted as the expected confidence of the trading partner. More details about the confidence updates can be found in Section 3.2.3. The ToM2 agent may choose to disregard its second-order beliefs, and instead behave as a ToM1 agent. This is reflected in the following formula, which describes how the expected value of an offer O can be computed:

$$\begin{aligned} &EV_i^{(2)}(O, r_i, b^{(0)}, b^{(1)}, b^{(2)}, p^{(1)}, p^{(2)}, c_1, c_2) = \\ &\quad (1-c_2) \cdot EV^{(1)}(O, r_i, b^{(0)}, b^{(1)}, p^{(1)}, c_1) + c_2 \cdot \sum_{r \in R} p^{(1)}(r) \cdot EV_i^{(2)}(r, O, r_i, b^{(1)}, b^{(2)}, p^{(2)}, 1)\,. \end{aligned} \tag{3.11}$$

In this formula, $EV^{(1)}(O, r_i, b^{(0)}, b^{(1)}, p^{(1)}, c_1)$ refers to Equation 3.4 and $EV_i^{(2)}(r, O, r_i, b^{(1)}, b^{(2)}, p^{(1)}, 1)$ refers to Equation 3.9, where c_1 is substituted by a 1. The ToM2 agent does not attribute a confidence score to the trading partner, instead it assumes that this trading partner always uses ToM1. This is a design choice that was taken from De Weerd and colleagues. It prevents the expected ToM1 model of the ToM2 agent from converging to a ToM0

model. In other words: if the ToM2 agent models the trading partner as a ToM1 agent, this should not be replaced by modeling it as a ToM0 agent. Equation 3.11 omits the belief update for considered offers described De Weerd et al. (2017), similar to Equation 3.6. Furthermore, it uses the confidence in the second-order reasoning c_2 and $p^{(2)}$. The latter is similar to $p^{(1)}$ in the sense that it models beliefs about the producing resource of an agent. However, $p^{(2)}$ is not the beliefs about the producing resource of the trading partner, but what the ToM2 agent believes that the trading partner's producing resource beliefs $p^{(1)}$ are about the ToM2 agent. How these beliefs are formed is described in Section 3.2.3.

The $EV_i^{(2)}$ function is used to compute the expected value of each offer $O \in \mathcal{D}$. The ToM2 agent then chooses (one of the) offers O_t^* with the highest EV:

$$O_t^* := \arg\max_{O \in \mathcal{D}} EV_i^{(2)}(O, r_i, b^{(0)}, b^{(1)}, b^{(2)}, p^{(1)}, p^{(2)}, c_1, c_2)\,. \tag{3.12}$$

Apart from the computation of the expected value, the behavior of the ToM2 is comparable to that of the other agent types. Again, there are two alternative actions apart from making a counter-offer: withdrawing, or accepting the previous offer O_{t-1}. Note that the latter is only possible after the initial round. The function that is used by the ToM2 agent to determine its action is:

$$\begin{aligned} &ToM2_i(O_{t-1}, r_i, b^{(0)}, b^{(1)}, b^{(2)}, p^{(1)}, p^{(2)}, c_1, c_2) \\ &\qquad = \begin{cases} O_t^* & \text{if} \quad EV_i^{(2)}(O_t^*) > \pi_i(r_i, D_0) \text{ and} \\ & \qquad\quad EV_i^{(2)}(O_t^*) > \pi_i(r_i, O_{t-1}) \\ \text{accept} & \text{if} \quad \pi_i(r_i, O_{t-1}) > \pi_i(r_i, D_0) \text{ and} \\ & \qquad\quad \pi_i(r_i, O_{t-1}) \geq EV_i^{(2)}(O_t^*) \\ \text{withdraw} & \text{otherwise,} \end{cases} \end{aligned} \tag{3.13}$$

where $EV_i^{(2)}(O_t^*)$ denotes $EV_i^{(2)}(O_t^*, r_i, b^{(0)}, b^{(1)}, b^{(2)}, p^{(1)}, p^{(2)}, c_1, c_2)$.

A ToM2 agent may accept an offer in which it needs to give a number of items from its producing resource r without receiving anything in return. Again, this offer type is only accepted by the ToM2 agent if there are no better offers available. Furthermore, ToM2 agents also model how they expect others to think about them. Although it is not explicitly modeled in Experiment 1 (Section 3.4), ToM2 agents might thus care about how other agents perceive them, since it may impact future negotiations. The acceptance of such offers thus relates to the interplay between ToM and reputation.

Learning Within Negotiations

Within a negotiation, an agent (indirectly) learns about the desires of the trading partner, even without modeling their mental content. If a specific offer is rejected, an agent will be more

likely to try another offer in the next round, instead of trying the same offer again. This is reflected in the negotiation-specific beliefs: $b^{(0)}, b^{(1)}, b^{(2)}$.

How quickly an agent learns from the actions of its trading partner depends on its learning-speed parameter $\lambda \in [0,1]$. Although in reality, different people have different learning speeds, in this research this variable was chosen to be constant for every agent. Furthermore, the focus of this research lies on the dynamics of different orders of ToM within a population. The results might be impacted by the difference in learning speed between agents. The learning speed was thus chosen to be constant to make the results of the research more interpretable and reproducible. A qualitative study of multiple test runs showed that the results were impacted by the value of the learning rate. Therefore, three variants of the experiment were created, each using a different value for λ: 0.8, 0.5, or 0.2. These values were chosen to test a slow learning process, an average learning process, and a fast learning process.

Let's first consider $b^{(0)}$: zero-order beliefs. When an agent receives an offer O_{t-1}, it can use this to update his beliefs about which offers the trading partner might accept, and which not. That is, if an agent i receives an offer from trading partner j and this offer gives an x amount of resources of a specific resource to i, it decreases its $b^{(0)}(O)$ for offers $O \in \mathcal{D}$ that ask for more resources of that type from j than O_{t-1} gave i. This in turn decreases the likelihood of the agent picking O for O_t, i.e., the offer it will make this round. An example can be found in Example 3.

Example 3: Suppose an agent i received an offer from the trading partner j that gives three of resource type 1 to i. In later steps of this negotiation, i will be less likely to make offers where it asks for more than three of resource type 1. This is reflected by the zero-order beliefs: $b^{(0)}(O)$ decreases for all offers $O \in \mathcal{D}$ that ask more than three of resource type 1. This example is visualized in Figure 3.6. The image should be observed from right to left, as i is influenced by the offer that was made by j. It shows how i decreases its zero-order beliefs about an offer where it asks for four items of resource type 1 (food), which is more than j gave in its offer. Note that the same update happens for offers where i asks for five items of resource 1 (food).

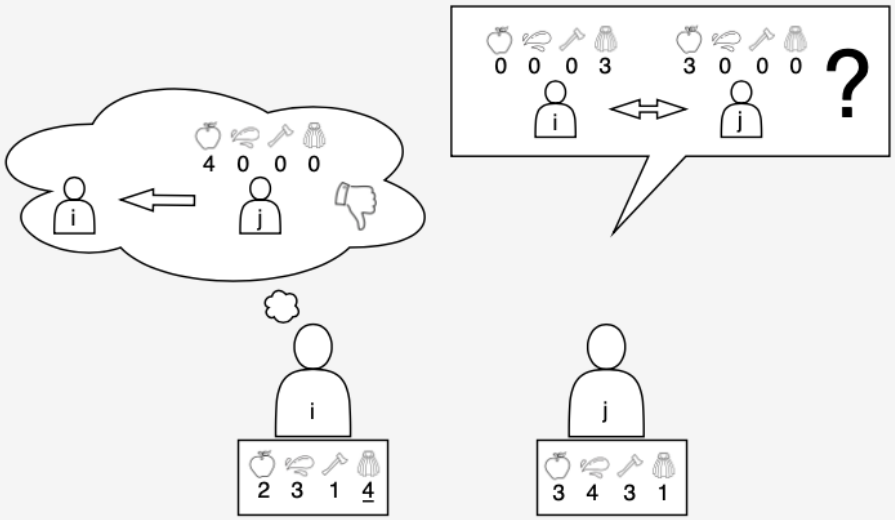

Figure 3.6: A visualization of an example where agent i decreases its belief about an offer when it has received an offer from j.

The update rule for the zero-order belief about an offer $O \in \mathcal{D}$ based on receiving offer O_{t-1} from the trading partner is:

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

where m indicates the number of resource types for which the offer O asks more from the trading partner than that trading partner gives of that resource type in offer O_{t-1}. The range of m is thus $[0,4]$. This definition of m differs from the one given by De Weerd et al. (2017) due to the difference in environment. In the CT setting by De Weerd and colleagues, the number of chips in the game is constant. Therefore, in the CT setting the number of items you ask for of a type/color is equal to giving the total number of items of that type/color minus what you ask for. In this research, the number of resources per type differs per negotiation, and there is not a complete redistribution of the resources. As a result, an agent could ask for a number of resources of a type, but this would not say anything about what this agent gives. The definition of m was chosen to closely resemble that of De Weerd and colleagues to be able to compare the results with those obtained by them.

Another situation that causes an agent to update its zero-order beliefs is when it proposes an offer O_t, and that offer is rejected by the trading partner. This gives the agent information about what the trading partner does not want. If an agent i made an offer O_t that asks an x amount of resources of a specific resource from trading partner j, and this offer was rejected, it decreases

its $b^{(0)}(O)$ for offers $O \in \mathcal{D}$ that ask for least as many resources of that type from j than it did in O_t gave i. This in turn decreases the likelihood of the agent picking O later on. The idea is that if the trading partner does not want to give this, it probably does not want to give more. An example can be found in Example 4.

Example 4: Suppose an agent i made an offer O_t that asks for three of resource type 1 from trading partner j. This offer is rejected by j. In later steps of this negotiation, i will be less likely to make offers where it asks for three or more of resource type 1. This is reflected by the zero-order beliefs: $b^{(0)}(O)$ decreases for all offers $O \in \mathcal{D}$ that ask three or more of resource type 1. This example is visualized in Figure 3.7. It shows how i decreases its zero-order beliefs about an offer where it asks for three items of resource type 1 (food), which is the same amount that was asked in its previous offer that was rejected. Note that the same update happens for offers where i asks for four or five items of resource 1 (food).

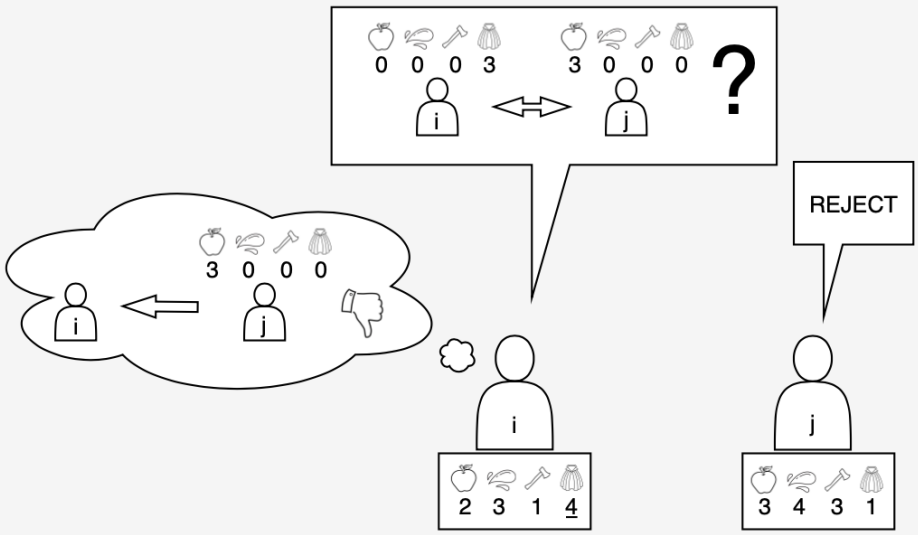

Figure 3.7: A visualization of an example where agent i decreases its belief about an offer when its previous offer was rejected by j.

The update rule for the zero-order belief about an offer $O \in \mathcal{D}$ based on offer O_t being rejected by is:

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

where m' indicates the number of resource types for which the offer O asks as least as many from the trading partner as the rejected offer O_t did. The range of m' is thus $[0,4]$. Similar to the

definition of m, also the definition of m' differs from the one given by De Weerd et al. (2017) due to the difference in environment. Contrary to the CT setting, in this research, the number of resources per type differs per negotiation, and there is not a complete redistribution of the resources. The definition of m' was chosen to closely resemble that of De Weerd and colleagues to be able to compare the results with those obtained by De Weerd and colleagues.

The first-order beliefs of ToM1 and ToM2 agents are updated similarly. $b^{(1)}$ is updated when a ToM1 knows the trading partner is updating its zero-order beliefs. The ToM1 agent then uses the described update rules to update its own first-order beliefs as if it were the zero-order beliefs of trading partner. Since this research takes the learning speed λ to be constant, the $b^{(1)}$ of the ToM1 agent and the $b^{(0)}$ of the trading partner are updated the same way. The only possible difference between the two is caused by a difference in *initialBeliefs*, since the history of the two agents is likely to differ.

The second-order beliefs $b^{(2)}$ of ToM2 agents represent what this agent thinks the trading partner thinks that the ToM2 agent has as zero-order beliefs $b^{(0)}$. To make this a bit more concrete, consider a ToM2 agent i. This agent has zero-order, first-order and second-order beliefs that it uses to choose an action during a negotiation with a trading partner j. i is aware that j might reason about its mental content, i.e., that of i. Therefore i models what it thinks j thinks are the zero-order beliefs $b^{(0)}$ of i. In this research, the second-order beliefs are identical to the zero-order beliefs. That is, $b^{(0)} = b^{(2)}$. This is caused by the resources of the trading partner being fully observable, and by the constant λ. When i models the second-order beliefs, it knows that j can see what resources i has and how these are updated after the trade. It furthermore assumes that the *initialBeliefs* of j are the same and thus that j assigns the same zero-order beliefs to i as i actually has.

Apart from the acceptance beliefs, ToM1 and ToM2 agents also model what they think the producing resource of their trading partner is. Receiving an offer from a trading partner gives insight into what that agent is and is not interested in, which relates to the producing resource of this agent. Agents know that all agents play rationally, and that they will thus only make an offer if it benefits them, else they would withdraw from the negotiation. An agent will therefore not ask for a resource that it produces itself. The producing resource beliefs are reflected by a probability p over each of the four resources. ToM1 agents use $p^{(1)}$ to decide their action, and ToM2 agents use $p^{(1)}$ and $p^{(2)}$ The former is what the agent thinks the producing resource is of the trading partner, and the latter is what the agent (i) thinks that the trading partner (j) thinks that its (i's) producing resource is.

A producing resource probability $p^{(k)}(r)$, where $k \in \{1,2\}$ and $r \in R$, is set to zero if the trading partner would not benefit from making the offer O_{t-1} that it made. Furthermore, it is increased when O_{t-1} made by the trading partner has an EV close to the EV of the best offer O^* for this agent if it would choose an action while having r as producing resource. It is decreased when the EV that the ToM agent thinks its trading partner receives for O_{t-1} is less than what the ToM agent would achieve if it had r as producing resource and chose its best offer O^*. The

update rule for $p^{(k)}$ for a resource $r \in R$ is:

$$p^{(k)}(r) := \begin{cases} 0 & \text{if } \pi_j(r, O_{t-1}) \leq \pi_j(r, D_0) \\ \beta \cdot p^{(k)}(r) \cdot \frac{1+EV_i^{(k-1)}(O_{t-1})}{1+\max_{O \in \mathcal{D}} EV_i^{(k-1)}(O)} & \text{otherwise,} \end{cases} \tag{3.16}$$

where $EV_i^{(k-1)}(O)$ refers to $EV_i^{(0)}(O, r_i, b^{(0)})$ (Equation 3.1) or $EV_i^{(1)}(O, r_i, b^{(0)}, b^{(1)}, p^{(2)}, c_1)$ (Equation 3.6) for $k = 1$ and $k = 2$ respectively. Furthermore, β is a normalizing constant that ensures that the four probabilities always add to one. This update rule is called whenever an agent receives an offer ($p^{(1)}$), or when it gives an offer and thus knows that the trading partner receives and offer ($p^{(2)}$). If the p-value for each of the four resources is set to zero, the agent has made a misjudgment and thus $p^{(k)}$ is reset to 0.25, the initial value, for each of the four resources. This may happen when a trading partner with ToM2 has made a certain offer to mislead this agent by making an offer that does not benefit the agent.

Finally, the confidence score c_1 needs to be updated for ToM1 and ToM2 agents, and c_2 for ToM2 agents. These scores reflect how much confidence the ToMk agent has in the k-th order, where $k \in \{1, 2\}$. c_k is updated whenever an agent receives an offer. The ToM agent compares this offer O_{t-1} with the offer it predicted based on the $k-$th order. If O_{t-1} gives a high $EV^{(k)}$ to the agent, then the confidence in this order k increases, and vice versa. The update rule for c_k is:

$$c_k := (1-\lambda) \cdot c_k + \lambda \cdot \sum_{r \in R} p^{(k)}(r) \cdot \frac{1 + EV_i^{(k)}(O_{t-1})}{1 + \max_{O \in \mathcal{D}} EV_i^{(k)}(O)}, \tag{3.17}$$

where $EV_i^{(k)}$ refers to $EV_i^{(1)}(O, r_i, b^{(0)}, b^{(1)}, p^{(2)}, c_1)$ (Equation 3.6) or $EV_i^{(2)}(O, r_i, b^{(0)}, b^{(1)}, b^{(2)}, p^{(1)}, p^{(2)}, c_1, c_2)$ (Equation 3.11) for $k = 1$ and $k = 2$ respectively. An example of the intuition behind these updates of c_k is given in Example 6.

Example 6: A ToMk agent i, where $k \in \{1, 2\}$, models the behavior of its trading partner j as if it were an ToM$(k-1)$ agent. When the trading partner behaves differently than expected, i decreases its confidence c_k. This is visualized in Figure 3.8.

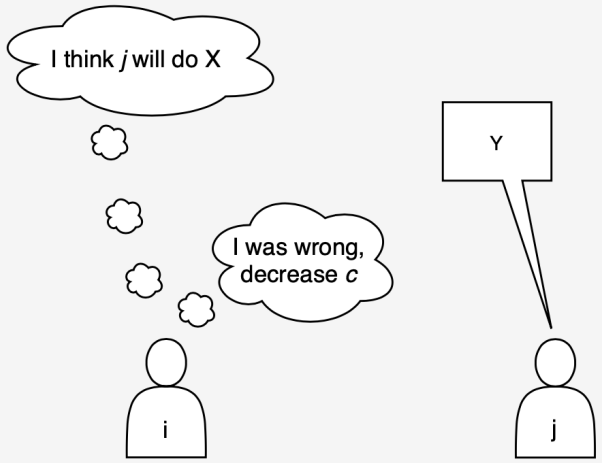

Figure 3.8: A visualization of the idea behind c updates, where agent i decreases its confidence after trading partner j behaves differently than predicted.

The confidence scores start at 1, representing that initially, ToMk agents model their trading partner as a ToM$(k-1)$ agent. Gradually, this confidence decreases if the used ToM model seems inaccurate. This implementation was used to resemble the methods of De Weerd et al. (2017) as closely as possible.

Learning Across Negotiations

Agents learn from previous encounters. Regardless of the order of ToM of the agent, agents realize that other agents differ from each other, and thus that an offer O may be accepted by one agent, but rejected by another. For that reason, they keep track of past experiences, and use the general information from these encounters as a baseline for new negotiations. Note that this type of learning requires no ToM, since the agent does not reason about the desires of the trading partner, it only requires looking at previous actions from trading partners.

Practically, this means that agents keep track of the number of offers they made per offer type, and how many of those were accepted. An offer type in this case refers to a ratio of the number of resources that were proposed to be given, and the number of resources that were proposed to be received. This ratio is general to all negotiations, since it does not consider which resource type is traded for which. There exist a total of 400 offer types. Example 5 gives some further explanation about how the offer type is determined and why there are 400 types. For each offer type, the agent computes the probability of this offer type being accepted by the

trading partner as follows:

$$initialBeliefs(\text{offer\_type}) = \frac{\#\text{offers accepted} + 5}{\#\text{offers made} + 5}, \tag{3.18}$$

where a constant of five is added to the numerator and denominator to avoid the probability of reaching a value of zero, and thus a zero percent chance of accepting a certain offer. This can be seen as giving each agent five positive experiences for each offer type.

Example 5: Suppose an agent proposes the offer that was described in Figure 3.3, i.e., the agent asks for two of resource type 1, one of resource type 2, two of resource type 3, and zero of resource type 4, and gives none of the four resource types. In this case, the number of resources given is zero, and the total number of resources asked for is five. The ratio and thus the offer type is therefore 5:0.

For each of the four resources, there are five possible options for giving: giving $[0,4]$. There are also five options for receiving for each of the four resources. Thus, there exist a total of $4 \cdot 5 \times 4 \cdot 5 = 20 \times 20 = 400$ offer types.

These initial beliefs are used throughout the life of an agent, and the agent updates them after every offer it makes. At the start of each negotiation, they are used as a baseline for the negotiation, by initializing the negotiation-specific beliefs (i.e., $b^{(0)}, b^{(1)}, b^{(2)}$) with these probabilities. Note that it takes a while before the initial beliefs are properly initialized since each of the offer types needs to be encountered several times.

3.3 Evolutionary Process

Within the arena, an evolutionary process is established. The three key aspects of evolution as described by Nowak (2006), selection, replication, and mutation, were implemented in the environment. This section describes how agents can survive (Section 3.3.1), how the process continues by replicating agents when another agent dies (Section 3.3.2), and how mutation is introduced in the population (Section 3.3.3).

3.3.1 Selection

Agents can only survive when they have enough resources to continue living, that is, if they reach the resource threshold for each of the four resource types. The threshold was set to two, since this leaves enough room for negotiations, even though agents are only able to offer their spare resources. Initially, the agents have one resource per type. Agents thus do not survive if they are inactive. Instead, they have to negotiate with others in order to trade resources. This also justifies the choice of the threshold of two, since the agents need to gather at least one

resource from every resource type, except for their producing resource type. Each agent has one producing resource type *r*, of which it always has four at the start of a negotiation to allow for free trade. From the perspective of an agent, this *r* is worthless: they can use it to make offers but do not wish to receive it. They would also not mind giving the trading partner all their resources of this type, although they do not offer this themselves. The negotiation process is explained in Section 3.2.2, and the behavior of the agents during a negotiation is explained in Section 3.2.3.

Agents need some time to be able to reach the resource threshold of two. Therefore, their resources get checked in intervals, once every CHECK_INTERVAL checks, which is 2500 in this research. This value was chosen based on a qualitative study. The interval was tested in three scenarios: with $\lambda = 0.2, \lambda = 0.5$ and $\lambda = 0.8$. This resulted in an average (over 30 runs) number of 41, 43, and 43 agents surviving the check, respectively. Thus, the check interval of 2500 causes approximately two-thirds of the agents to reach their threshold. All agents have an age, which is used for the evolutionary process. Initially, their age is initialized to a random integer between zero and a chosen CHECK_INTERVAL, to ensure that not all agents in the population have the same age, which would be counter-intuitive [2]. As a result, the ticks at which agents get an evolutionary check vary. When the resources of an agent are checked and the agent has enough of each type, it continues to live. One resource is deducted from each resource type. This resembles the usage of those resources. It also stimulates further negotiation, since the agent again needs to reach the threshold of two for each resource type. In the case that the agent does not manage to reach the threshold for each of the resource types, it dies. This agent thus dies through natural selection: it was not fit enough to survive.

In this experiment, it is possible that an agent is checked whilst it is taking part in a negotiation. If the agent did not reach the threshold, is continues this negotiation, but dies afterwards, regardless of the outcome of the negotiation. The same holds for the case where the agent did reach the threshold: it's resources are reduced after the negotiation. This design choice was made to ensure that the negotiation process of the trading partner is not affected by the revolutionary process.

3.3.2 Replication

When an agent fails its evolutionary check, it dies. This means that the agent is removed from the population. Then, a surviving agent is randomly selected. This individual serves as the parent of a new agent; some of its characteristics are replicated. The new agent inherits the order of ToM of its parent, as well as the *initialBeliefs*. This ensures that this new agent has a baseline and thus a fair chance at surviving. The probability of a new agent being of a certain ToM order is proportional to the quantities of these different types in the population. The age of this new agent is zero and its producing resource is random. The new agent then behaves as

[2] For the interval tests, the initial age of all agents was zero

any other agent. That is, there is no special communication or behavior between the parent and the child. The initial location of the new agent is arbitrary and is therefore not influenced by that of its parent.

3.3.3 Mutation

When an agent dies, there is a small probability that the new agent will be a mutation. In other words, there is a chance that the new agent does not copy both characteristics (i.e.,, the ToM order and the *initialBeliefs*) from a parent. Instead, the mutation causes the order of ToM k to be chosen randomly such that $k \in [0,2]$. In practice, this means that all orders of ToM have an equal chance of being chosen for this mutated agent. Therefore, also extinct orders can return to the population. The beliefs were chosen to be copied from a random living agent. This gives these mutated agents a baseline and thus still a shot at surviving like the other agents. There is a 2% chance that a new agent is mutated. This value was chosen to be low enough to keep stability in the population, while still allowing for occasional mutations to introduce diversity.

3.4 Experiments

Two types of experiments were conducted. In the first experiment, previous negotiations do not influence the movement of agents through the arena. This experiment is described in Section 3.4.1. Section 3.4.2 describes the second experiment, in which agents rejected negotiations with incompatible partners and remembered what they learned from previous negotiations with this partner. Finally, the procedure of the experiments is discussed in Subsection 3.4.3.

3.4.1 Experiment 1

At the start of the experiment, a baseline is created for the behavior of the agents by initializing the *initialBeliefs* of the agent. This is achieved by allowing a population of 120 ToM0 agents to negotiate and move through the environment as if the experiment had started, but without the possibility of dying. This allows the agent to gain experience that is used to update their zero-order beliefs. As soon as an agent gains four items of each resource, their resources are reset so that they can continue learning. Although ideally, this initialization period would be infinite to give each of the ToM orders a fair chance in the experiment by ruling out the lack of experience, this is not feasible. The initialization period was therefore set to 1,000,000 ticks. Qualitative research showed that by that time, all agents had encountered each type of offer multiple times and thus updated the corresponding *initialBeliefs*.

When a baseline has been created, this needs to be implemented into the agents that participate in the experiment. This is done by randomly sampling (with replacement) one agent from

the initial population of ToM0 agents, and copying these *initialBeliefs* into those of a newly created agent. This is repeated for each of the 60 new agents of the experiment. Thus, all agents start with a baseline, but this baseline differs per agent. Another option would have been to reuse the same *initialBeliefs* all agents. The latter option was disregarded because the influence of one abnormality on the outcome of the experiment would be large. As an example, consider a ToM0 agent of the initial population, that by chance repeatedly encountered similar situations and did not develop an informed initial belief about another situation. If this agent were to be taken as the prototype of all agents of the experiment, the results may not be representative of the experiment. To rule this out, the randomized prototype-picking tactic was chosen. For similar reasons, the initialization process was repeated for every run of the experiment, instead of saving the baseline and using this for all the runs of the experiment.

When the population of agents is created, the experiment starts. This means that the agents are free to move throughout the arena and negotiate with each other as described in Section 3.2. The evolutionary data, i.e., the number of agents per agent type per time step, is tracked. The experiment ends when one of the following two stopping conditions is met:

1. One type of agent remains,

2. The maximum number of ticks of 3,750,000 is reached.

The maximum number of ticks, 3,750,000, was chosen based on a qualitative study of multiple test runs. Since the check interval of the resource thresholds is 2500, this maximum represents 1500 generations. In most of the test runs, the experiment terminated earlier due to two orders going extinct. In a few cases, this was not the case. However, a maximum of 1500 generations was also picked because of the memory complexity of the program. Several types of data (described in the next paragraph) were collected and saved, and 1500 generations were approximately the maximum number of generations that were feasible.

Note that due to the nature of the evolutionary process and the possibility of mutations, extinct orders of ToM can come back due to a mutation. Therefore, if one type of agent remains, this does not necessarily mean that the other types are gone forever. However, in this research, the experiment does stop when one type remains.

3.4.2 Experiment 2

Experiment 2 builds upon experiment 1: the description of experiment 1 thus still holds. There are two differences between the two experiments. Firstly, in experiment 2, agents save information about their trading partners. In practice, this means that all agents save their producing resource beliefs about the trading partner. If they encounter an agent again, they do not have to restart forming this belief. ToM1 agents additionally save their c_1, so the confidence in the ToM order of the trading partner. ToM2 agents also save this c_1, as well as c_2. Agents thus recall

their previous beliefs about the agent. Note that the other beliefs, so $b^{(0)}, b^{(1)}$, and $b^{(2)}$, are not saved. These beliefs are negotiation-specific and may be misleading in a new negotiation.

Another difference compared to experiment 1 is that now, negative encounters cause agents to avoid the agents that they had these negative encounters with. If the outcome of a negotiation was that either of the agents withdrew, or that the time limit was reached, an agent saves the index of this trading partner to a queue. Agents avoid the trading partners that are in the last five elements of this queue. This stimulates them to negotiate with new agents. If an agent encounters a trading partner that is in its queue but not in the last five elements, it gives this trading partner a new chance. It can then either be removed from its queue, if the outcome of the negotiation is positive, or be placed at the end again.

3.4.3 Procedure

Several types of data were collected. Firstly, the final distribution of the agents was saved: so the number of agents per type at the moment of the termination of the experiment. This data gives insight into which orders of ToM were useful to survive the evolutionary checks. When an agent dies, its age is saved. So, for each of the ToM orders, the ages of all the agents of that type are collected as well. Additionally, the history of these distributions per type step can also be saved. This can be determined by the user. The dominance frequency for each ToM order was saved as well. It is the fraction of the ticks that this order was (one of) the most occurring order of ToM in the population at that time step. The frequency does thus not always add to 1, since multiple orders can be dominant. For example, after the initialization, there are 20 agents per type, and thus each of the three ToM orders is dominant in the first time step.

Furthermore, information about the negotiations was collected. For each negotiation in an experiment, the type of negotiation (so the orders of ToM of both agents), the negotiation length, and the gain of both agents after the negotiation were saved. The agent that made the initial offer was determined as well, as this may influence the negotiation outcome.

The two experiments were both run for each of the three values for $\lambda \in [0.2, 0.5, 0.8]$. Each sub-experiment was repeated 120 times, resulting in 360 runs per experiment. This value was chosen to balance between robust results and computational feasibility.

3.5 Parameters

Table 3.1 gives an overview of all hyper-parameters of this research, and their corresponding values. The variables in caps-lock can be adjusted by the user, whereas the others are implicitly implemented in the code. The table is structured as follows: in the top section, all agent parameters are includes, followed by the hyper-parameters of the evolutionary process and the experiments.

Parameter	Value	Description
n_resources	4	number of resource types that exist
N_AGENTS	20	initial #agents per ToM-type
AGENT_RADIUS_THRESHOLD	10	negotiation radius
N_START_RESOURCES	1	initial #resources per resource type
MAX_N_RESOURCES	4	max #resources per resource type
LEARNING_SPEED	[0.2, 0.5, 0.8]	how quickly the agent learns from experiences
MAX_N_ROUNDS	50	max #rounds per negotiation
age (initial)	[0,CHECK_INTERVAL]	initial age that determines when the agent gets evolutionary checks
CHECK_INTERVAL	2500	#timesteps between evolutionary checks
RESOURCE_THRESHOLD	2	min # resources per type to survive
MUTATION_CHANCE	2	probability of new agents being mutated (out of 100)
INIT_TIME	1,000,000	#ticks of initialBelief forming
EXP_LENGTH	3,750,000	#ticks before experiment stops

Table 3.1: Hyper-parameters of the agents, the evolutionary process, and the experiments.

3.6 Implementation

The experiments were implemented in the programming language Java[3]. Along with the experiments, a simple graphical user interface (GUI) was created to visualize the agents and their environment, serving as a way to obtain qualitative results. The interface displays the arena in which the 60 agents are positioned, some buttons to manage the experiment, and an information pane. A screenshot of the interface during an experiment can be found in Figure 3.9. The green square is the arena, which includes black, red, and blue dots, which are the ToM0, ToM1, and ToM2 agents respectively. The screenshot furthermore shows two buttons, the 'New' button and the 'Pause' button. The information pane on the right of the interface shows the time, i.e., the number of ticks that have passed, the current agent counts per ToM order, and a graph showing the most recent evolutionary activity.

The interface can be used to run the experiments. A new file can be created by clicking the 'New' button in the top left corner. This generates a new arena with new, uninitialized agents, and it cleans the information pane. Furthermore, the interface can be used to initialize the *initialBeliefs* of the agents by clicking the button 'Initialize', which is located at the position of the 'Pause' button in Figure 3.9 when a new file is created. The 'Pause' button can be clicked to pause the initialization or the experiment. A 'Start' button appears that can be used to start the initialization or the experiment again after pausing it.

The program requires three parameters, one that specifies whether or not to enable the GUI (true or false), one that specifies which experiment to run (1 or 2), and one that specifies the name of the file to which the results are written.

When an experiment ends, the data are saved to the computer automatically, as a CSV file with the specified name. This file is stored in the 'EvolutionToM' folder. The GUI can thus be used to start a new experiment without the risk of losing the data of the previous experiment. Note that the previous results file will be replaced if the file name is not altered.

For this project, the data are structured as follows: all the results are placed in the 'finalResults/quantitative' folder, which contains a folder for experiment 1 and experiment 2. Within both folders, the results are divided based on their corresponding λ values. The folders furthermore contain a file that can be used to visualize the data 'processData.py', and 'processDataExp2.py' for experiments 1 and 2 respectively. This program requires an integer input argument, which specifies how many experiments were done per lambda value. It can then be used to create graphs that show the collected data.

A detailed explanation of how to run the simulation and the other programs can be found in the README of the GitHub repository (see footnote 3).

[3]For the code, see `https://github.com/SanneBerends/EvolutionToM`

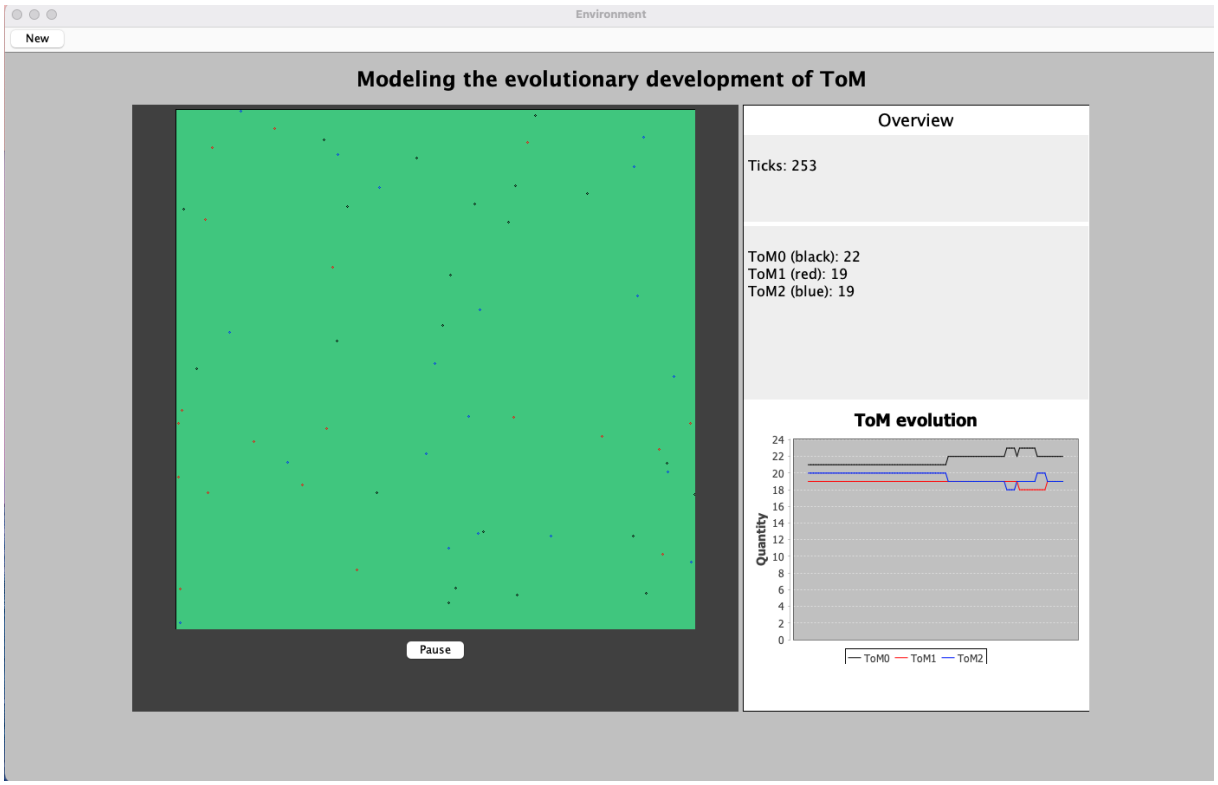

Figure 3.9: A screenshot of the GUI, including the arena (green), two buttons, and an information pane. The information pane shows the number of ticks so far, the agent count per ToM order, and a graph with the most recent progression of these agent counts.

Chapter 4

Results

This chapter presents the results of experiment one and two, as described in Section 3.4.

4.1 Experiment 1

The results of Experiment 1 are discussed in this section. First, some qualitative results are presented, followed by the quantitative results.

4.1.1 Qualitative Results

Although the focus of this research lies mainly on the quantitative analysis, a qualitative analysis was done to check if the agents exhibited any unexpected or unwanted behavior. Furthermore, it was used for parameter tuning, as mentioned in the Methods section (Section 3).

The first observation is that agents with second-order Theory of Mind (ToM2) sometimes make an offer that seems to be disadvantageous for them. An example is an agent with initial resources [4, 4, 4, 2], with $r = 1$. This agent makes offer where it gives [0, 0, 1, 0] and asks for [0, 0, 0, 0], even though he would receive nothing in return for the (non-producing) resource that it gives. This would yield the following resources: [4, 4, 3, 2]. The reason behind these types of offers is that the ToM2 expects a good offer in return. In this particular case, the agent was certain about the producing resource of the trading partner and therefore expected the following counteroffer: receiving [0, 0, 0, 2] and giving [1, 0, 1, 0]. Accepting this offer would yield the following resources: [3, 4, 3, 4], which is better than the current state of the agent: recall that the agent does not receive points for its producing resource. This example highlights the ability of ToM2 agents to signal their preferences. By making the disadvantageous offer, the agent shows that it is willing to give an item of resource type 3.

Another observation is that an agent often has higher initial beliefs for offers where it asks for m resources of a type than offers where it asks for n resources, where $m > n$. This is

caused by the way that the belief updates work. Initial beliefs can only go down, so agents who encounter a certain situation many times are likely to develop a low belief for this situation. There are fewer scenarios where an agent can consider asking for many resources of a type than scenarios where it can ask for little resources of a type, since this depends on the inventory of the trading partner. As a result, the beliefs for those offers where it asks for little may get a lower belief than those where it asks for a lot. This sometimes causes an agent to make an offer where it asks for *m* of a resource, even though an offer where it asked for *n* of this same resource was already rejected. In practice, this can increase the negotiation length.

Furthermore, a (ToM0) agent sometimes withdraws even though there appears to option for a deal with its trading partner. This has to do with the way the negotiation starts. If a ToM0 agent receives a 'high' offer like one where it has to give [1, 1, 1, 0] in exchange for [0, 1, 0, 1], this agent will decrease the beliefs for offers where it gives less than asked for in the received offer. As a result, the agent will not make the offer of giving [0, 0, 0, 1] in exchange for [0, 0, 1, 0], even though it would benefit both agents. This behavior is a consequence of the implementation of zero-order ToM.

Another important observation is that a relatively large amount of negotiations do not terminate before the negotiation limit. It occurs mostly when the learning speed of the agents is 0.2, in which case it can occur in each ToM type. For the other learning speeds, it only happens when no ToM0 agent participates. An inspection of the negotiations where the limit was reached, showed that the offers that were made were repeated every turn. This behavior is the result of the implementation of the ToM, especially the confidence scores used to determine which ToM order is projected on the trading partner. Each ToM1 and ToM2 agent starts a negotiation with a confidence score of 1 in its order of ToM. This confidence can be decreased when different behavior is observed from the trading partner than expected. When an agent made an offer and received a counteroffer, the zero order belief for that offer decreases, which ideally would decrease the chance of that offer being chosen again the next round. However, these beliefs are only used when the confidence score is lower than 1.0. This pattern would be resolved as soon as the confidence score decreases, but this happens only very slightly. The confidence updates depend on the difference between the expected action of the trading partner and what this agent would do, but this difference is often very small since the agent is good at estimating the behavior of the trading partner. This especially leads to repeating offers if the agent has no way of knowing the producing resource of the trading partner. This explains why it is more frequent in negotiations with a ToM1 agent: these agents need the producing resource of the trading partner to find a good offer. If there is no way of finding out what this producing resource is, there will be no improvement in offers. The problem occurs less often for ToM2 agents, since these agents might instead start signaling their own producing resource. If this is impossible, they too will likely start repeating offers. If the learning speed is low, the confidence scores reduce even slower, explaining why it is a more-often occurring problem in these experiments. These findings are reflected in the qualitative results of the negotiation

termination reasons (see Section 4.1.2).

Finally, apart from the experiments with 60 agents, also two alternatives were tested: 30 agents and 90 agents. Initially, 180 agents were tested, but this resulted in a memory issue. These variations were tested by running them three times for each learning value. The number of experiments was thus too small to draw any conclusions, but a qualitative inspection of the results indicated that for the experiments with 30 agents, the ToM1 and ToM2 agents had a higher survival rate than for the experiments with 60 agents. The results of the experiments with 90 agents seem similar to those found for the experiments with 60 agents.

4.1.2 Quantitative Results

The quantitative results of the first experiment include all the data that were collected in the 120 runs per learning speed value, as described in Section 3.4.3.

Experiment Lengths

The maximal length of the experiment was set to 3.750.000 ticks, corresponding to 1500 generations. An experiment terminated either when this maximum was reached, or when only one type of agent remained. The experiment lengths of all experiments are visualized in a violin plot, shown in Figure 4.1. The plot displays the spread of the experiment lengths per learning speed. Two horizontal lines were added to indicate the minimum and maximum length of the experiments. Note that these limits were not exceeded. The plot reveals that for each learning speed, the majority of experiments ended before reaching the maximum length. In other words, the majority of the experiments terminated due to one species surviving. The plot furthermore shows that the median of the experiment lengths was highest for the learning speed of 0.8, then for 0.2, and finally 0.5. The latter also shows the widest shape, indicating that the variability in experiment lengths is the smallest for this value. The shape of the violin for the learning speed of 0.8 is the narrowest, and the inter-quartile range is the largest, indicating that the experiment length varied more than for the other learning speeds.

Final Agent Distribution

The final distribution of agent types was saved for each experiment. The 360 experiments were divided into two categories: the experiments where only one agent type remained, and those where multiple types remained. This was done to make the graphs more readable since the majority of the experiments ended with only one species surviving. The average final distribution of agents, when one type survived, is plotted in Figure 4.2. The figure also shows the corresponding error bars, which reflect one standard error. As can be concluded from the figure, if the experiment terminated because one Theory of Mind (ToM) order survived, this surviving order was always ToM0. This was the case for each learning speed.

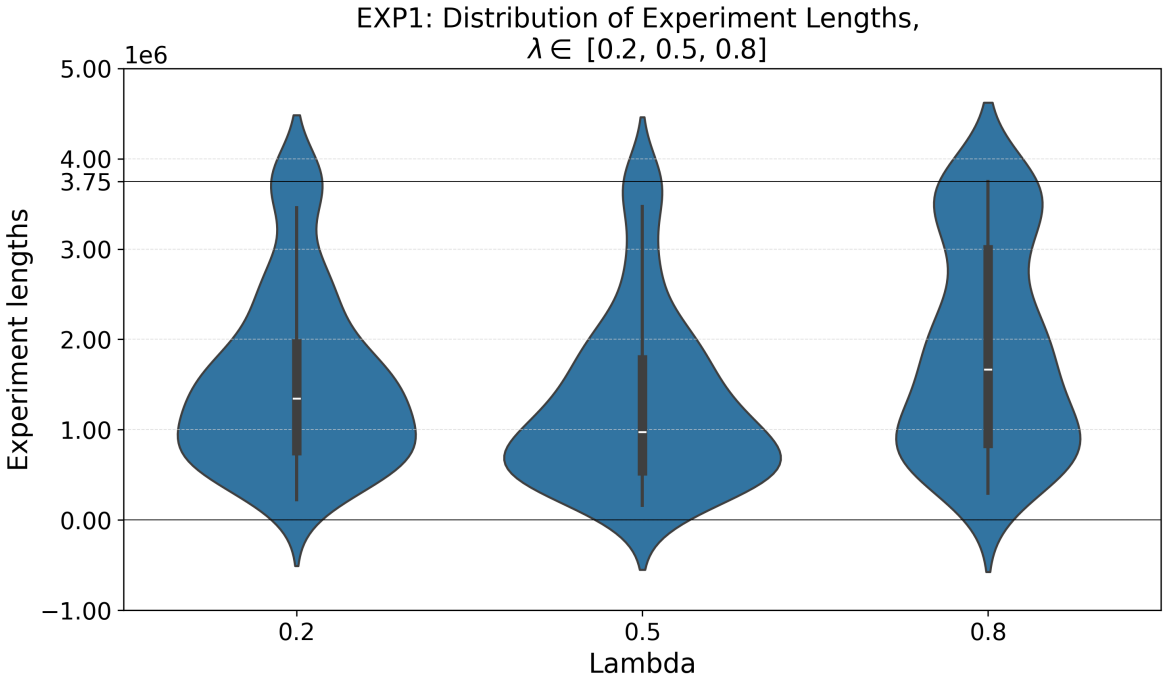

Figure 4.1: A violin plot showing the distribution of experiment lengths of experiment 1, plotted for the experiments with a learning speed of 0.2, 0.5, and 0.8.

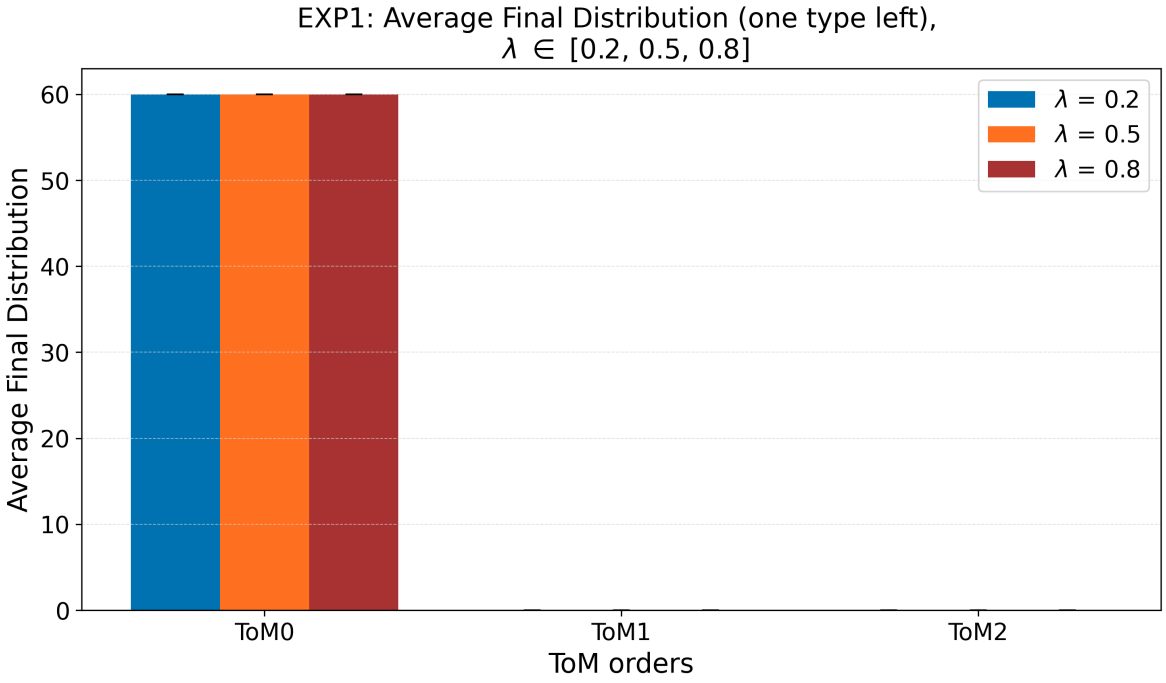

Figure 4.2: The average final distribution of agents for experiment 1 when only one ToM order survived, plotted for each learning speed. The error bars indicate one standard error.

In the case that multiple species survived until the end of the experiment, the number of surviving species was always two. This can be concluded from Figure 4.3, which displays the average final distribution when multiple ToM orders survived. Again, the error bars indicate one standard error. The figure shows that on average, ToM0 was still the most dominant species at the end of the experiment, but that ToM1 also had an average of 3.8, 5.9, and 6.0 agents left for learning speeds of 0.2, 0.5, and 0.8 respectively. In other words, whilst ToM2 always went extinct before the end of the experiment, this was not the case for ToM1. A closer look at the (standard) error bars reveals that the average final number of agents was not significantly different for the experiments with learning speeds of 0.2 and 0.8. For the experiments with a learning speed of 0.5, the average number of surviving ToM1 agents was lower than for the other learning speeds, and, consequently, the average number of surviving ToM0 agents was higher.

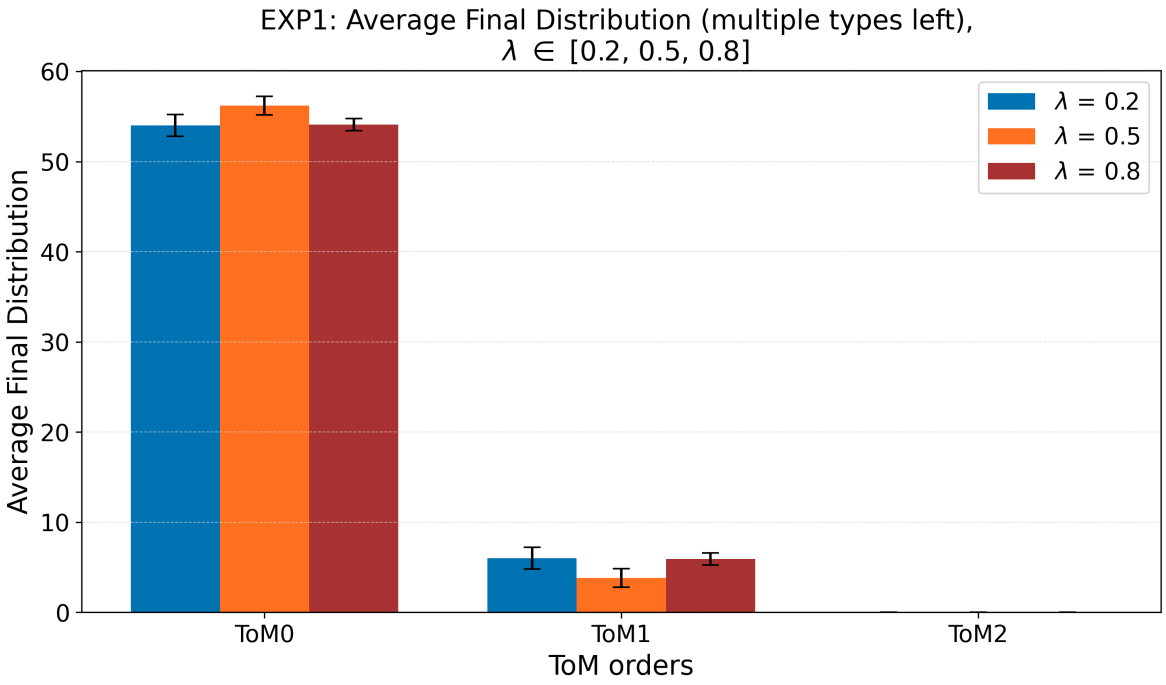

Figure 4.3: The average final distribution of agents for experiment 1 when multiple ToM orders survived, plotted for each learning speed. The error bars indicate one standard error.

Dominant Species

Besides the final distribution of agent types, the dominance frequencies of each ToM order were also collected. Recall that the dominance frequency of a ToM order is the fraction of the ticks that this order was (one of) the most occurring order of ToM in the population at that time step. Note that multiple species can be dominant simultaneously and that the summed frequencies can thus exceed 1.0. The average dominance frequencies can be found in Figure 4.4, where the error bars indicate one standard error. The figure shows that, for each learning speed, ToM0 had a significantly higher average dominance frequency than ToM1 and ToM2. The average dominance frequency of the former is close to 1.0, whilst that of the latter two are hard to distinguish from the graph due to the large difference compared to ToM0. Therefore, Figure 4.5 shows a zoomed-in version of a section of Figure 4.4. From this figure, it follows that the average dominance frequency of ToM2 was higher than that of ToM1, even though both were substantially smaller than that of ToM0. The (small) overlap of the error bars suggests that there was no significant difference between the results of the different learning speeds.

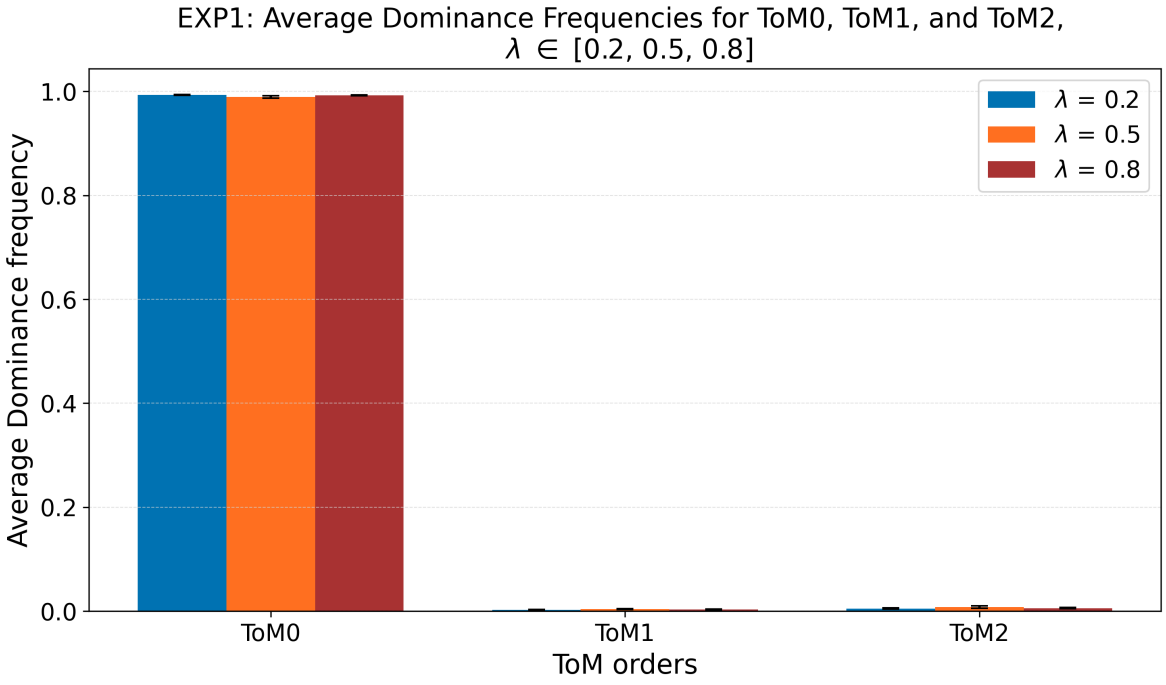

Figure 4.4: The average dominance frequency for experiment 1 per ToM order, plotted for each learning speed. Frequencies do not add to one because multiple types can be dominant simultaneously. The error bars indicate one standard error.

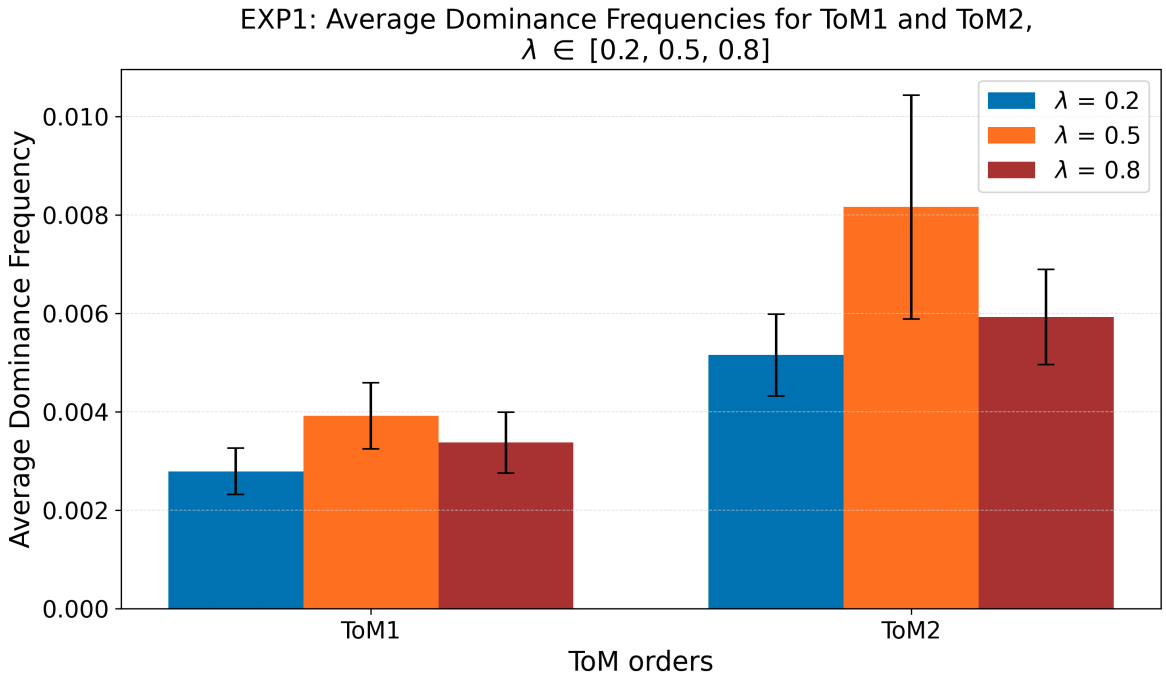

Figure 4.5: The average dominance frequency for ToM1 and ToM2: an enlarged version of a section of Figure 4.4. The error bars indicate one standard error.

Typical Run

The progression of the distribution of the agent types over time was collected for a selection of the experiments. An example of a typical run for the experiments with a learning speed of 0.2 can be found in Figure 4.6. The number of agents per ToM order was collected once every five ticks. The graph shows that in the first 50,000 ticks, the blue (dotted) line has the upper hand, indicating that ToM2 was dominant here. This matches the found dominance frequencies, in the small portion of the time-lapse where black (solid), i.e., ToM0, is not dominant, ToM2 is. The graph furthermore shows that the number ToM1 agents (red: dashed) quickly beats the number of ToM2 agents (blue: dotted). ToM1 went extinct around 430,000 ticks but came back due to a mutation. The experiment was terminated due to both ToM1 and ToM2 being extinct at the same time. This happened after approximately 1,250,000 ticks. Note that the red (dashed) line corresponding to the ToM1 agents does not reach zero; this is a consequence of only plotting one out of five ticks.

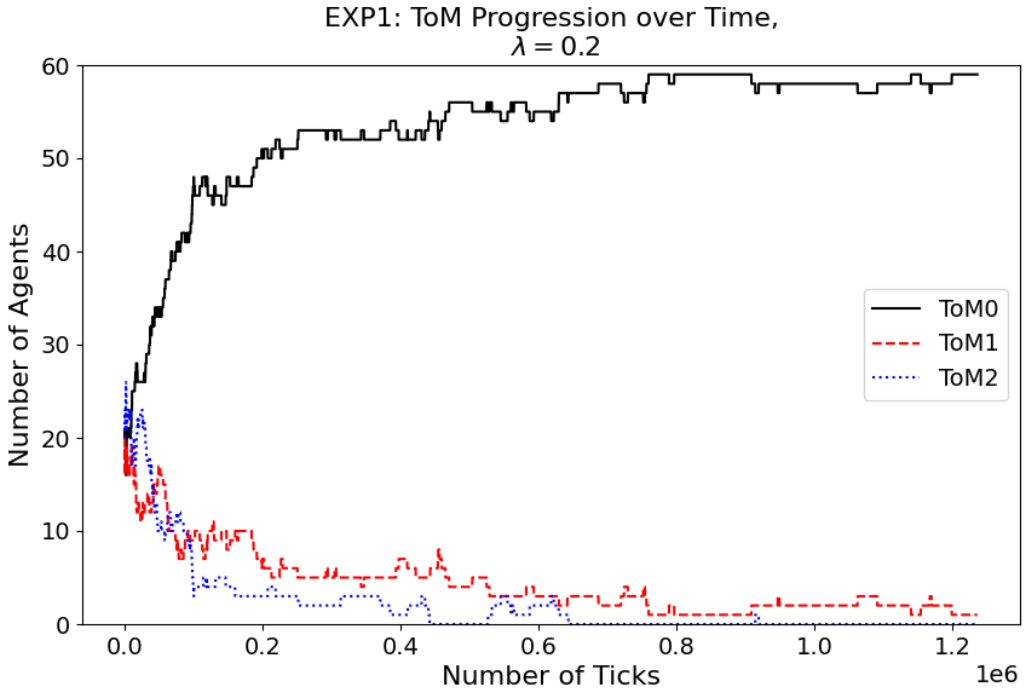

Figure 4.6: An example of a time-lapse of the distribution of agent types for the experiments of type 1 with a learning speed of 0.2. This example is representative of the other runs.

Figure 4.7 shows an example of the evolution of ToM over time for the experiments with a learning speed of 0.5. The progress was very similar to that discussed above. Again, ToM2

became the dominant species at the very start of the experiment. In this particular example, the period of its dominance was shorter than in the example run of $\lambda = 0.2$, but note that there was no significant difference overall. Again, the ToM0 agents were most dominant for the vast majority of the experiment. The ToM2 agents were the first to go extinct, whilst the ToM1 agents lasted longer in the environment, until also going extinct after approximately 1,080,000 ticks.

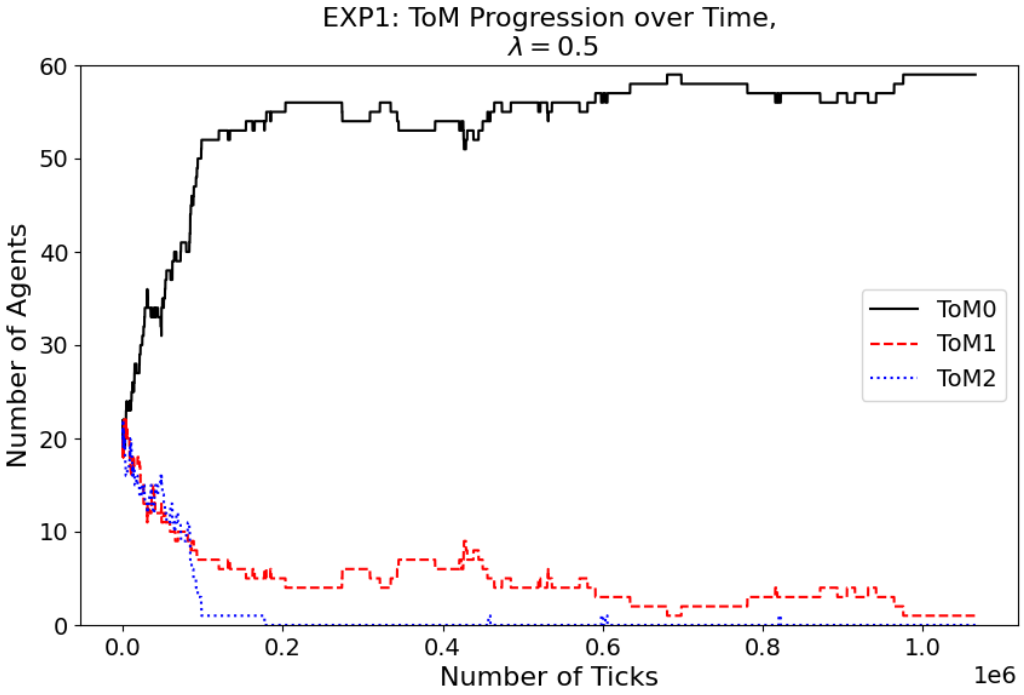

Figure 4.7: An example of a time-lapse of the distribution of agent types for the experiments of type 1 with a learning speed of 0.5. This example is representative of the other runs.

Finally, Figure 4.8 shows the evolution of ToM in a typical run of an experiment with a learning speed of 0.8. The pattern that can be observed is very similar to that of the other learning speeds. The main difference is that in this experiment, ToM1 agents lasted until the end of the experiment, so until the maximum experiment time was reached. In this situation, there were thus two agent types left at the end: ToM0 agents and ToM1 agents. This is in accordance with what can be seen in Figure 4.3.

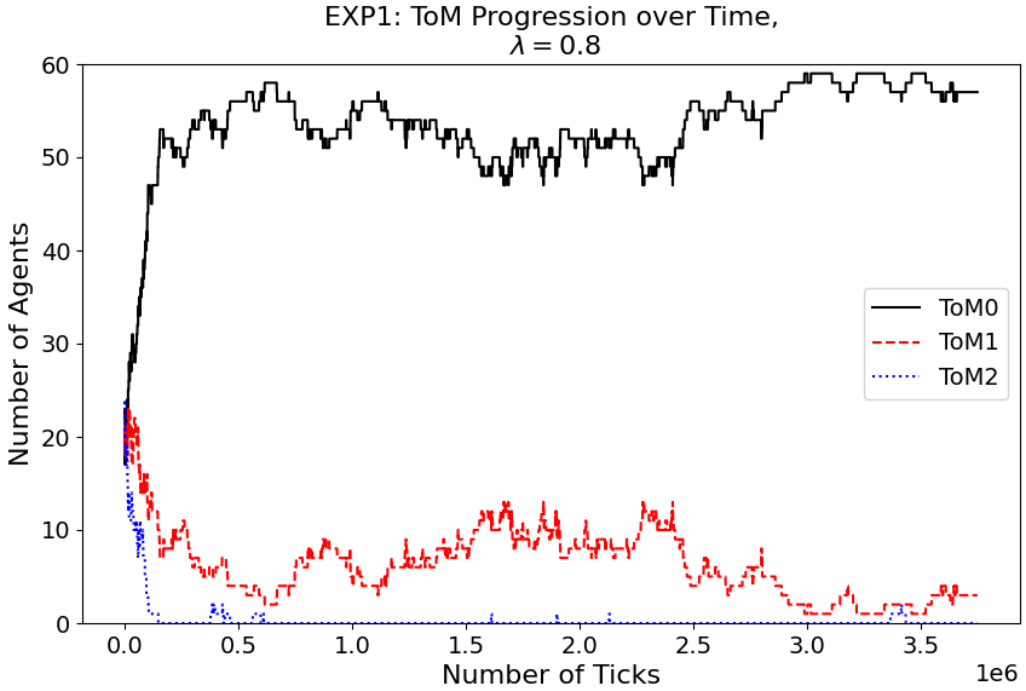

Figure 4.8: An example of a time-lapse of the distribution of agent types for the experiments of type 1 with a learning speed of 0.8. This example is representative of the other runs.

Agent Ages

The ages of the agents were collected for every experiment. The ages are plotted per learning speed, per order of ToM. Since the distribution of ages per ToM was very similar for the different learning speeds, the plots for $\lambda = 0.2$ and $\lambda = 0.8$ can be found in Appendix A.

Figure 4.9 shows the spread of the ages per ToM order for a learning speed of 0.5. Note that the ages did not exceed the lower bound of zero, even though it may appear so in the figure as a result of the wide spread of ages. The figure shows that ToM0 agents had the highest variety in ages. The wide bottom of the violin shape indicates that most agents died relatively young, but the long upper tail shows that some agents reached very high ages, up to approximately 2.4 million ticks. The distribution of ages of the ToM1 agents shows a similar pattern, except that the bottom of the violin is wider and thus that more agents died young compared to the ToM1 agents. The violin plot of ToM2 has a much shorter upper tail, indicating that agents of this type generally did not survive as long as the other types. The positions of the medians, shown by white lines in each violin, confirm that most agents across all ToM orders had relatively short lifespans.

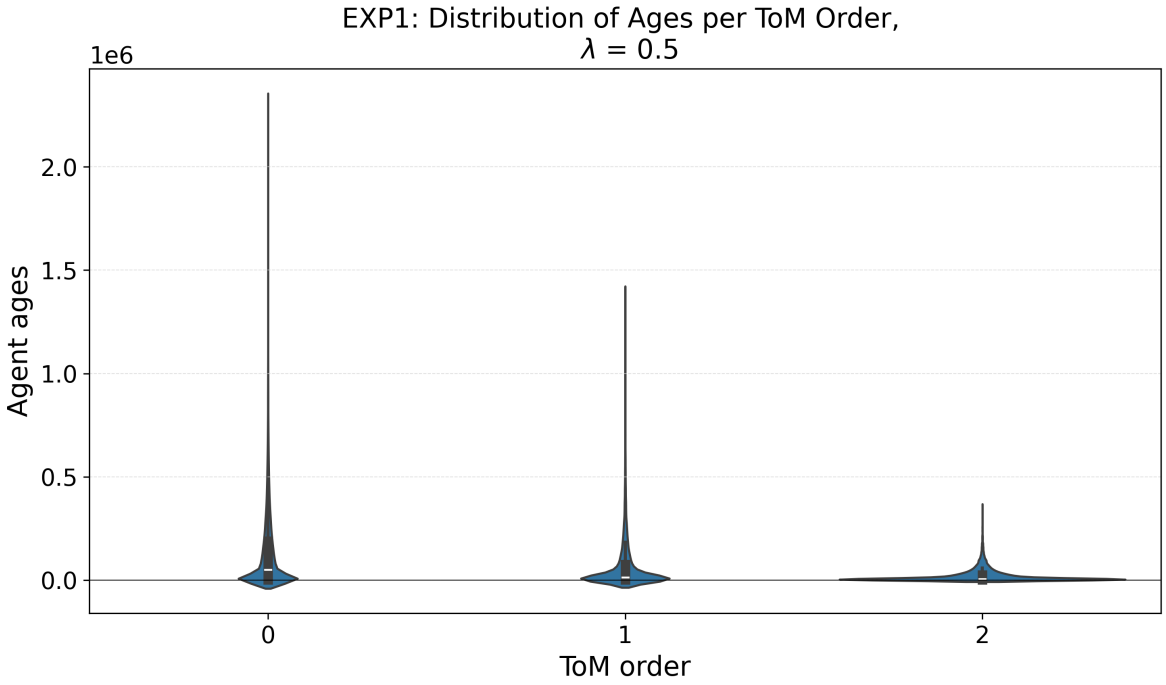

Figure 4.9: A violin plot of the spread of agent ages of experiment 1 per ToM order, for the experiments with a learning speed of 0.5.

Negotiation Lengths

Apart from the agent and experiment-related data, negotiation-specific data were also collected. Figure 4.10 shows the average negotiation length per negotiation type, for each learning speed. The negotiation type specifies the order of ToM of the agent that initiates the negotiation, and the order of its trading partner. The figure thus shows nine times three bars, along with their corresponding standard errors. A substantial amount of information can be deducted from the graph. Firstly, the results differ significantly between the three learning speeds. For each negotiation type, the average negotiation length was highest for a learning speed of 0.2, then 0.5, and finally 0.8. How large the difference in average negotiation length was between the experiments with the different learning speeds varied per negotiation type. For the ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2 negotiations, the difference was the smallest, although still significant (the error bars do not overlap). The difference in negotiation length was the largest for ToM0-ToM0 negotiations, with the average lengths being 16.4, 7.6, and 4.5 for λ = 0,2, 0.5, and 0.8 respectively.

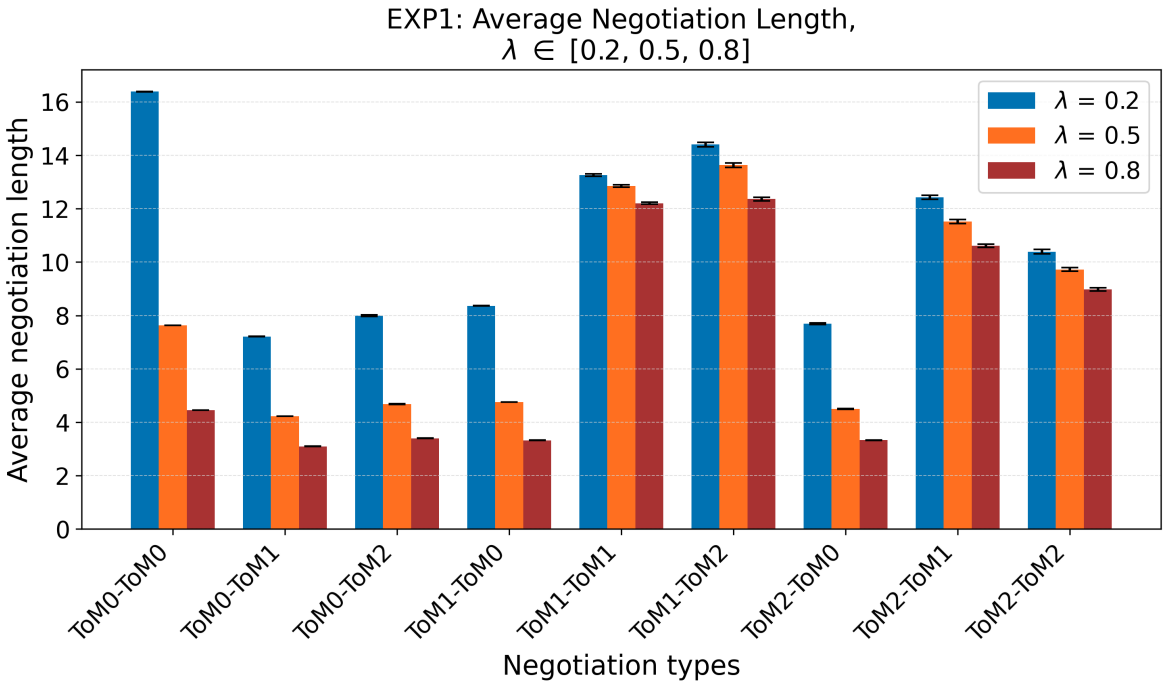

Figure 4.10: The average negotiation length for experiment 1 per negotiation type, plotted for learning speeds 0.2, 0.5, and 0.8. The negotiation type specifies the order of the initiating agent and its trading partner. The error bars indicate one standard error.

Due to the variations in the results of the different experiments, the rest of the observations are discussed per learning speed. For a learning speed of 0.2, the average negotiation length of type ToM0-ToM0 was significantly higher than for the other negotiation types. In fact, all the observed differences between the average lengths of the negotiation types for this learning speed were significant. This type is followed by ToM1-ToM2 and ToM1-ToM1, with an average negotiation length of 14.4 and 13.3 respectively. After that, negotiations of type ToM2-ToM1 and ToM2-ToM2 had the highest average length, Followed by ToM1-ToM0, ToM0-ToM2, ToM2-ToM0, and finally ToM0-ToM1.

For a learning speed of 0.5, again all differences in average negotiation lengths were significant. The negotiation type with the highest average negotiation length was ToM1-ToM2 with an average length of 13.6, followed by ToM1-ToM1 with an average length of 12.9. Next, ToM2-ToM1, ToM2-ToM2, and ToM0-ToM0 on average resulted in the longest negotiations. On average, the shortest negotiations were those of type ToM1-ToM0, ToM0-ToM2, ToM2-ToM0, and ToM0-ToM1 respectively. The results thus indicate that all negotiation types that include a ToM0 agent had a lower length on average than those without a ToM0 agent.

For a learning speed of 0.8, the negotiation types with the longest average length were also ToM1-ToM2 and ToM1-ToM1 respectively, followed by ToM2-ToM1 and ToM2-ToM2. After these negotiation types, ToM0-ToM0 and ToM0-ToM2 had the highest average length respectively. The differences between these six negotiation types were significant. There was no significant difference between the types ToM1-ToM0 and ToM2-ToM0, but both had a significantly higher average length than ToM0-ToM1. Again, all negotiation types that include a ToM0 agent, had a lower length on average than those without a ToM0 agent.

Reasons for Negotiation Termination

Apart from the lengths of the negotiations, the reason for the termination of these negotiations was also collected. A negotiation terminates when either of the agents withdraws or accepts, or when the negotiation limit (of 50) is reached. The frequency data are plotted in three graphs, one per learning speed. Figure 4.11 shows the results for a learning speed of 0.2. The data are visualized as a stacked bar plot per negotiation type. What is notable is that the majority of the negotiation did not end successfully, since for each negotiation type, the summed frequency of either of the agents accepting was lower than 30%. Instead, most negotiations were terminated due to one of the agents withdrawing from the negotiation or the negotiation limit being reached. The success rate was the lowest for negotiation types ToM1-ToM1 and ToM1-ToM2. The highest success rates were found for ToM0-ToM0, ToM1-ToM0, and ToM2-ToM0 negotiations. Thus, for the negotiations where the non-initiating agent was a ToM0 agent.

The graph shows that for all negotiation types except for ToM0-ToM1 and ToM0-ToM2, the most occurring reason for the termination of the negotiation was that the initiator withdrew. For the two exceptions, the most occurring reason was that the trading partner withdrew. This

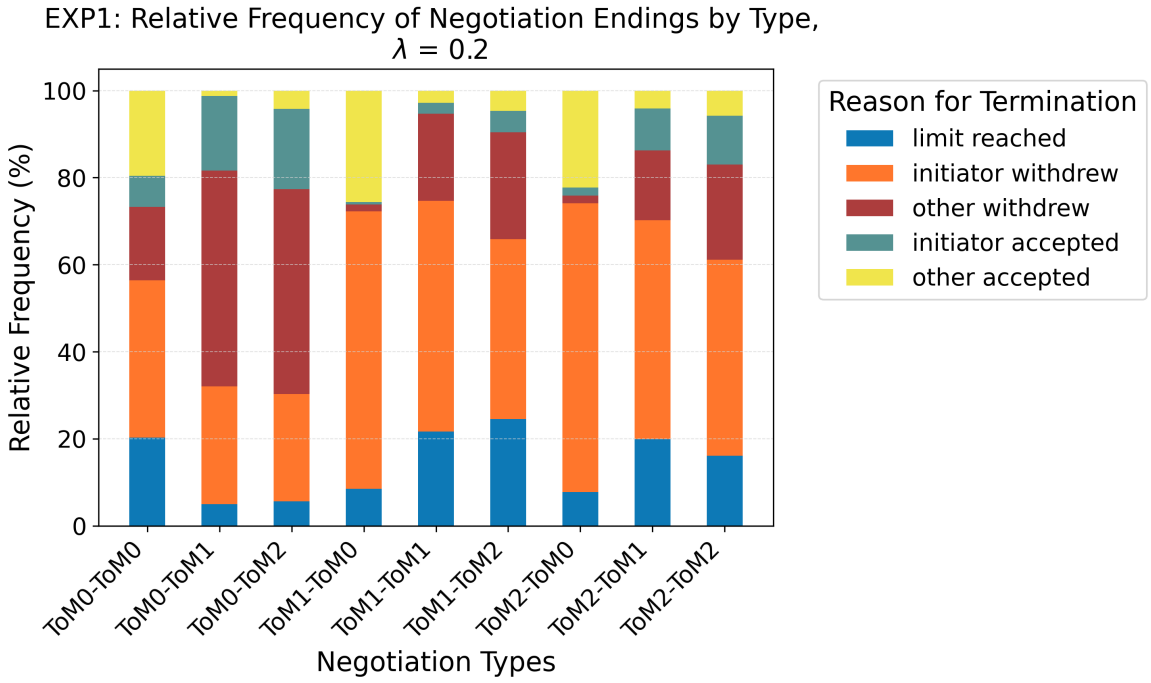

Figure 4.11: The reason for negotiation termination for experiment 1 per negotiation type, for the experiments with a learning speed of 0.2.

reason also has a relatively high frequency for the negotiation types ToM0-ToM0, ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2. The graph furthermore shows that for most negotiation types, a small frequency of the negotiations was terminated due to the negotiation limit. This frequency was relatively high (above 15%) for ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2: so in all negotiation types without a ToM0 agent. The graph furthermore indicates that for ToM0-ToM1, ToM0-ToM2, ToM2-ToM1, and ToM2-ToM2, the frequency of the initiator accepting was higher than that of the trading partner accepting. For ToM0-ToM0, ToM1-ToM0, ToM1-ToM1, and ToM2-ToM0, this was the other way around. For ToM1-ToM2 these frequencies were approximately the same.

Figure 4.12 shows the relative frequency of the reason for negotiation termination per negotiation type, for the experiments with a learning speed of 0.5. The pattern that can be observed is very similar to that for the experiments with a learning speed of 0.2. The main difference is that now, the time limit of the negotiation was only reached for negotiations of types ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2, and never for the other types. Still, for each negotiation type, over 70% of the negotiations did not end in a successful trade. Again, the negotiations where the non-initiating agent was a ToM0 agent had the highest success rate, whilst the negotiations of type ToM1-ToM1 and ToM1-ToM2 had the lowest success rate.

The relative frequencies of the reasons for negotiation termination per negotiation type, for the experiments with a learning speed of 0.8, were almost identical to those with a learning speed of 0.5. Therefore, the corresponding figure can be found in Appendix A.

Negotiation Gains

How useful a negotiation was to an agent can be measured in terms of the increase in its π-score. This represents the value of their resources. This increase is known as the negotiation gain for this agent. The gains were collected for each agent in each negotiation. The results are plotted for each learning speed, negotiation type, and the role of the agent (initiator or its trading partner).

Figure 4.13 shows the results for a learning speed of 0.2, i.e., the average gains for both agents, for each negotiation type. Note that only the gains of successful negotiations were included in the averages. The error bars indicate one standard error. The graph shows that there are negotiation types for which the average gain of the initiator was higher (ToM1-ToM0 and ToM2-ToM0), and types where that of the trading partner was higher (all the other types). In each of the negotiation types, the difference between the gain of the agents was significant, although some of the differences were much larger than others. For example, the difference in average gain was approximately 0.7 for negotiation type ToM0-ToM1, but only 0.1 for ToM1-ToM2.

Notable is that none of the average gains exceed 2.4. The highest average gain was approximately 2.3, which was achieved by the non-initiating agent in a ToM2-ToM2 negotiation.

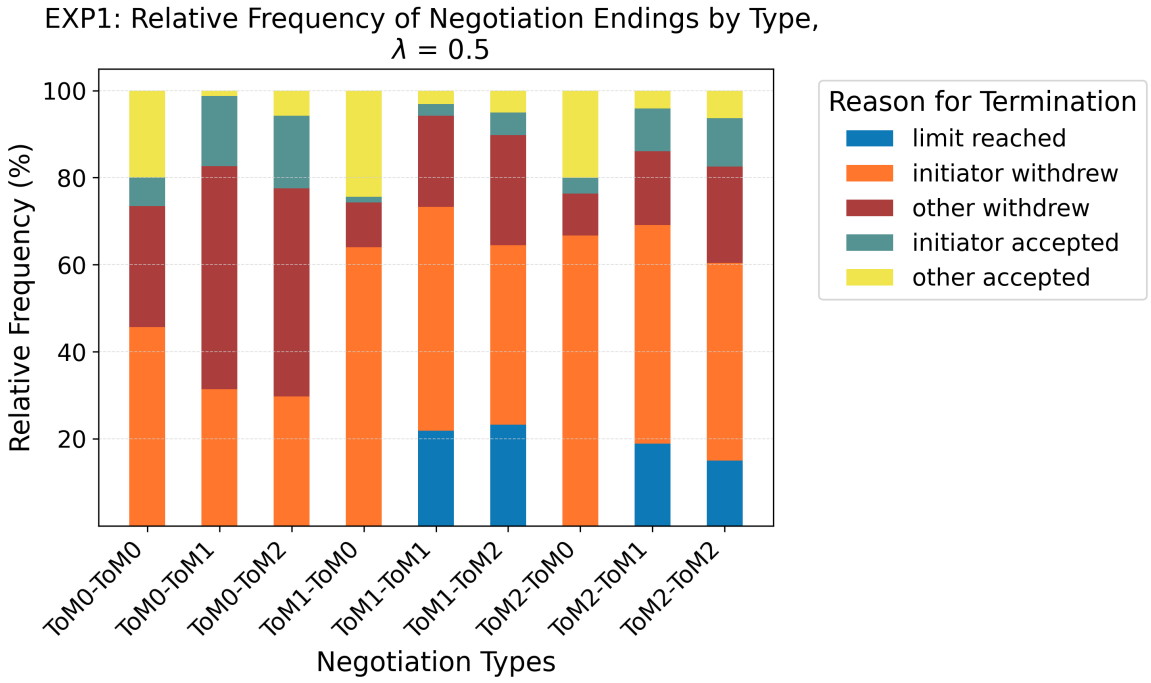

Figure 4.12: The reason for negotiation termination for experiment 1 per negotiation type, for the experiments with a learning speed of 0.5.

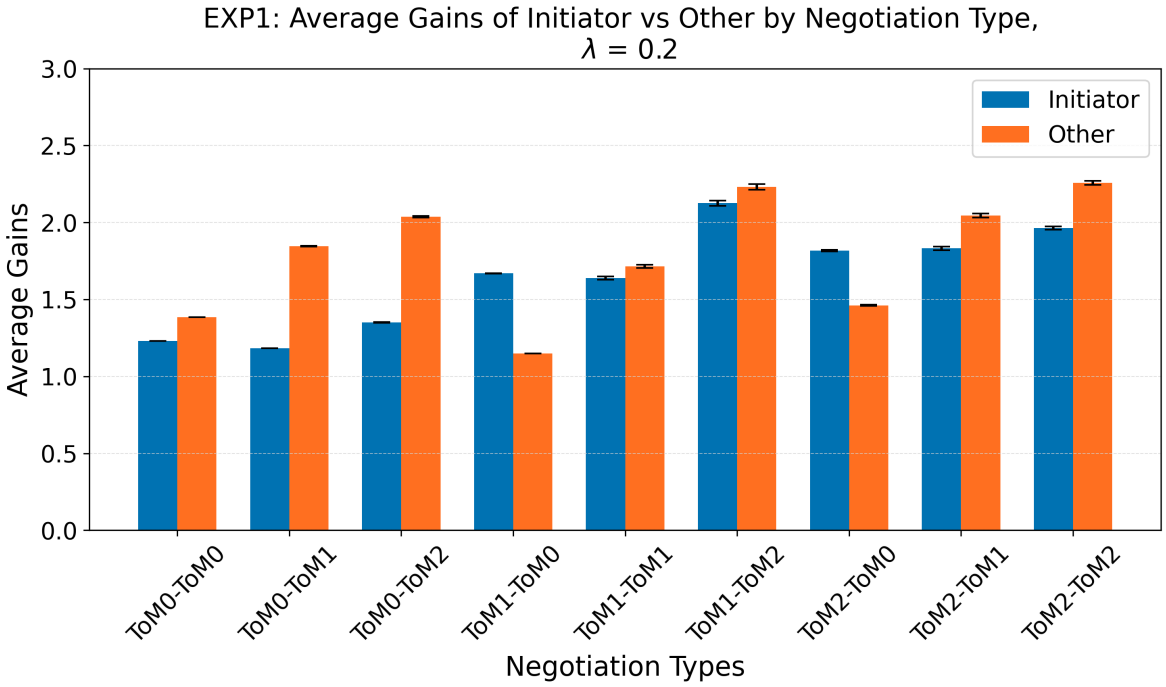

Figure 4.13: The average gains of the initiating agent and its trading partner for each negotiation type of experiment 1, plotted for the experiments with a learning speed of 0.2. The error bars indicate one standard error.

This average gain was not significantly higher of that obtained by the non-initiating agent of a ToM1-ToM2 negotiation. This was followed by both the initiating agent of this ToM1-ToM2 negotiation type. After that, the non-initiating agents of the ToM0-ToM2 and ToM2-ToM1 negotiations benefited the most from their negotiations, but there is no significant difference between these two negotiation types. The lowest gains (on average) were obtained by the initiating agent in the ToM0-ToM1 negotiation, and the non-initiating agent in the ToM1-ToM0 negotiation.

The graph indicates that the ToM1 agent received the highest average gain in most of the negotiation types that it was part of. The only exception is the ToM1-ToM2 negotiation type where the ToM2 agent received a higher gain on average. The ToM0 agent received the lowest average gain in each of the negotiation types that it was part of. The figure also shows how the average gain of an agent type depends on the order of its trading partner. A ToM2 agent, for example, received an average gain of 2.0 when being the non-initiating agent in a negotiation with a ToM0 agent, an average gain of 2.2 when being the non-initiating agent in a negotiation with a ToM1 agent, and an average gain of 2.3 when being the non-initiating agent in a negotiation with a ToM2 agent. The role of the agent within the negotiation also played a role: a ToM0 agent received a slightly lower gain on average when initiating a negotiation with a ToM2 agent than when a ToM2 agent initiated this negotiation. This can furthermore be seen in the ToM1-ToM2 and ToM2-ToM1 negotiation types. In both types, the agents had the same trading partner, but the average gain obtain depended on the role of the agent.

The average gains for the experiments with a learning speed of 0.5 were quite similar to those with a learning speed of 0.2. They can be found in Figure 4.14, where the error bars indicate one standard error. One of the differences is that in this scenario, the ToM1 agent received a slightly higher gain when initiating an experiment with another ToM1 than with a ToM0 agent, which was the other way around for the lower learning speed. Furthermore, for this learning speed of 0.5, there was no significant difference between the gains of the initiator and its trading partner in a ToM1-ToM2 negotiation. Also the difference in average gain between the initiator and its trading partner of a ToM1-ToM1 negotiation was no longer significant for this learning speed.

The average gains for the experiments with a learning speed of 0.8 can be found in Figure 4.15. The error bars indicate one standard error. The findings were again quite similar to those with a learning speed of 0.2. The main difference is that now, the initiating agent of the ToM1-ToM2 type received the highest average gain together with the non-initiating agent of the ToM2-ToM2 type. Furthermore, for the ToM1-ToM1 negotiation type, the initiator now obtained a higher average gain than its trading partner.

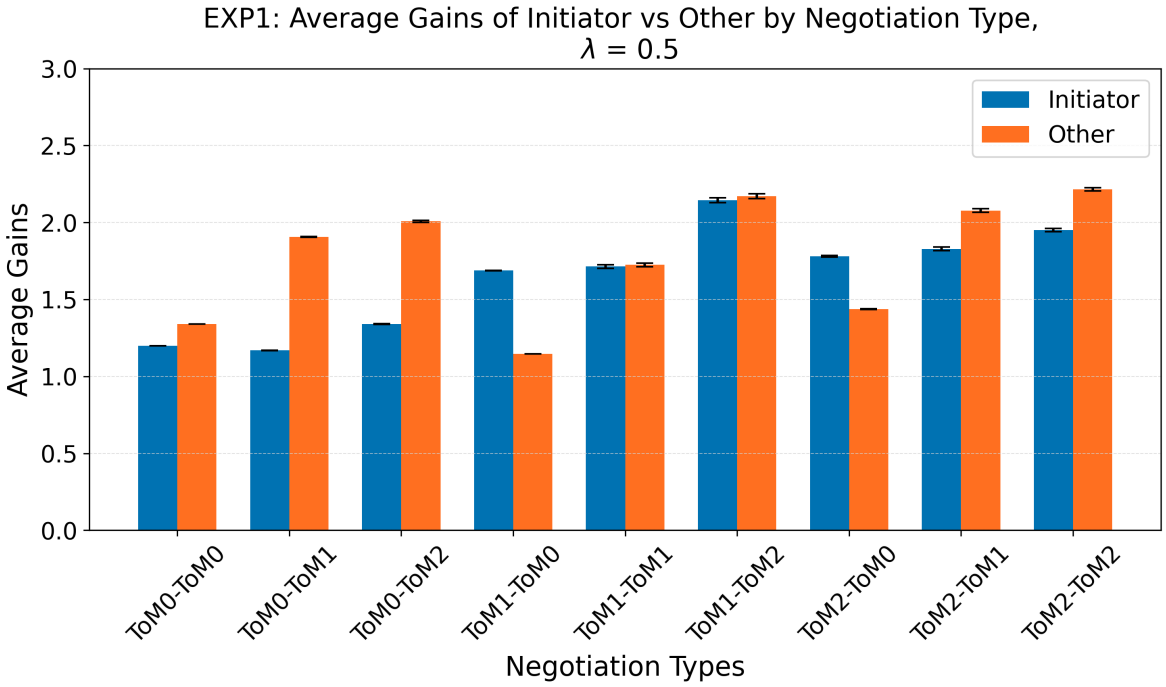

Figure 4.14: The average gains of the initiating agent and its trading partner for each negotiation type of experiment 1, plotted for the experiments with a learning speed of 0.5. The error bars indicate one standard error.

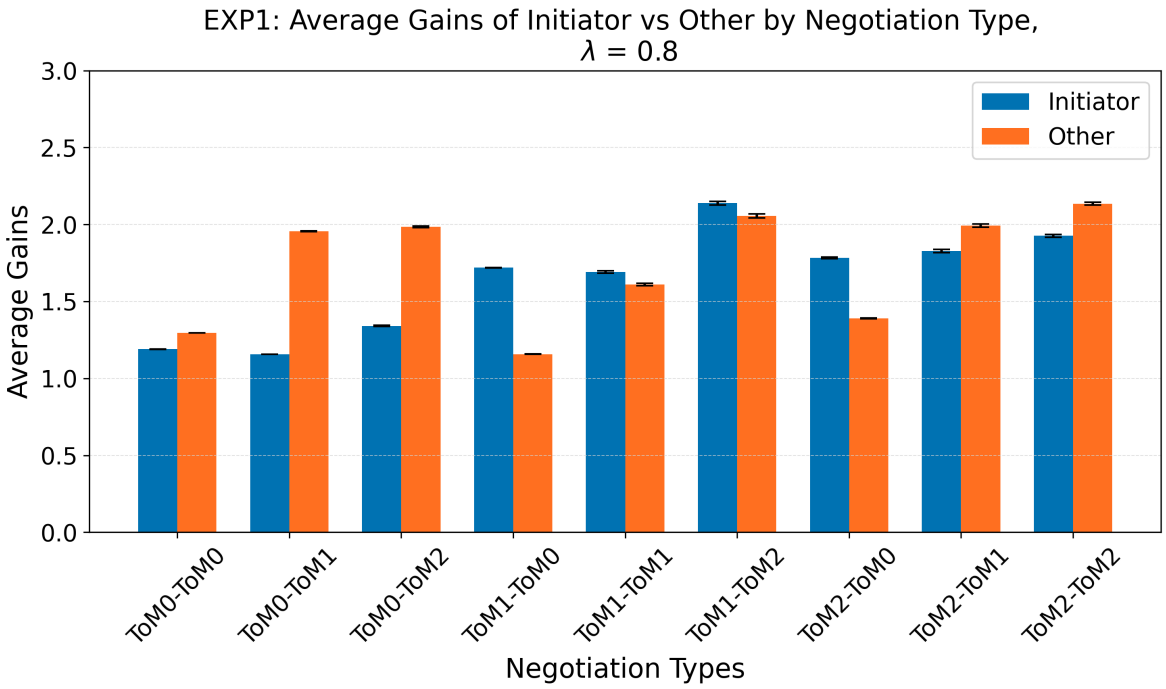

Figure 4.15: The average gains of the initiating agent and its trading partner for each negotiation type of experiment 1, plotted for the experiments with a learning speed of 0.8. The error bars indicate one standard error.

4.2 Experiment 2

The results of Experiment 2 are discussed in this section. First, some qualitative observations are presented, followed by the quantitative results.

4.2.1 Qualitative Results

The majority of the simulation was the same for both experiments. The logic of the ToM was not changed, and therefore the observations that were made for experiment one also hold for this experiment (see Section 4.1.1).

This experiment made the agents behave more realistically, since they remembered previous encounters with others. In real life, humans also use previous knowledge to make informed decisions. The model may thus be more accurate to answer the research questions. In the simulation, this meant that the agents now saved agent-specific data, so their confidence scores and producing resource beliefs. Agents could also reject agents based on previous experiments. Visually, the main difference compared to experiment 1 is that agents change their direction more often.

4.2.2 Quantitative Results

The quantitative results of the second experiment include all the data that were collected in the 120 runs per learning speed value, as described in Section 3.4.3.

Experiment Lengths

The distributions of the experiment lengths (per learning value) are visualized in the violin plot that can be found in Figure 4.16. The figure shows two horizontal lines: one that indicates the line where the number of ticks is zero, and one that indicates the maximal experiment length of 3.750.000 ticks. None of the experiments exceeded this length, even though this may seem the case for $\lambda = 0.2$. For this learning speed, the majority of the experiments ended due to the time limit being reached. This can be seen by the wide upper part of the violin. The white line represents the median, which lies around 3.25 million ticks. The long (relatively narrow) tail of the violin shows that some experiments ended before the time limit.

In contrast, the distribution of the experiment lengths for the learning speed of 0.5 shows that only a few experiments reached the limit, whilst the majority terminated around 1 million ticks, as indicated by the wide shape around this number. The narrower shape overall, compared to the left-most violin, indicates that the variation in experiment lengths was larger. The median of the experiment lengths with $\lambda = 0.5$ was lower than that of the others.

Finally, the right-most violin shows that for the experiments with a learning speed of 0.8, some experiments reached the maximum experiment length, as indicated by the wider shape at

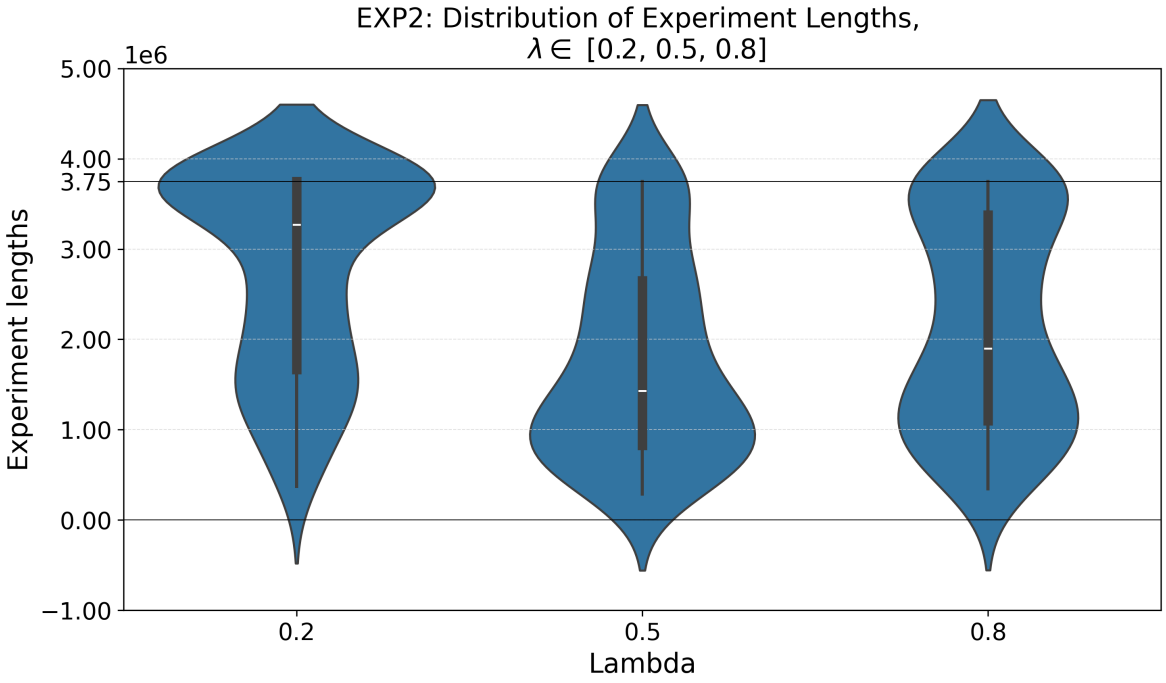

Figure 4.16: A violin plot showing the distribution of experiment lengths of experiment 2, plotted for the experiments with a learning speed of 0.2, 0.5, and 0.8.

the top. Furthermore, it shows a concentration of experiments that ended around 1 million ticks, similar to the violin of $\lambda = 0.5$. The median for this learning speed of 0.8 is lower than for 0.2 but higher than for 0.5.

Final Agent Distribution

For each experiment, the final distribution of agent types, and thus ToM orders, was saved and plotted. The 360 experiments were divided into those where only one type remained, and those where multiple types remained. The average final distribution of the former can be found in Figure 4.17. The error bars indicate one standard error. The figure shows that when the experiment ended because one ToM order survived, this order was always ToM0, regardless of the learning speed used in the experiment.

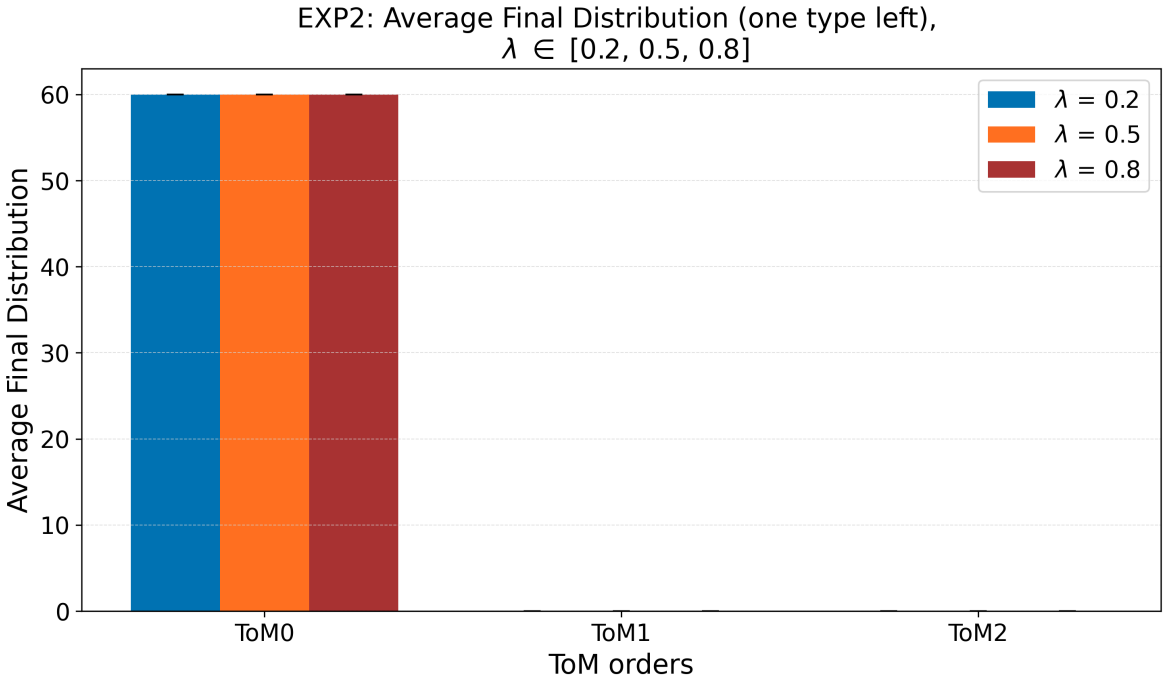

Figure 4.17: The average final distribution of agents for experiment 2 when only one ToM order survived, plotted for each learning speed. The error bars indicate one standard error.

In the case that multiple species remained at the end of the experiment, ToM0 agents still had the largest population at the end. This can be deduced from Figure 4.18, where the error bars again indicate one standard error. However, ToM1 agents also had an average population of 6.5, 4.4, and 5.5 left, for $\lambda = 0.2, 0.5$, and 0.8 respectively, although the difference between

these averages of learning speeds was not significant. Furthermore, at the end of the experiment, there were on average also still approximately 0.1 ToM2 agents left. In other words, there were cases in which at least one ToM2 agent survived until the end of the experiment. There was again no significant difference between the results of the different learning speeds.

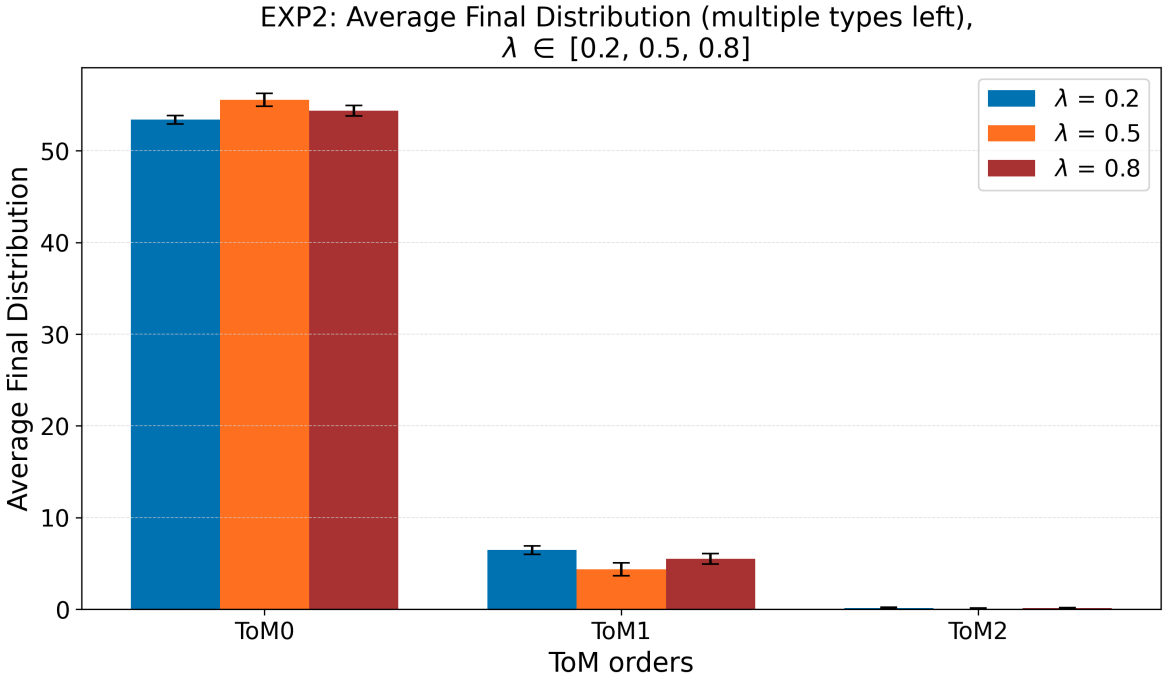

Figure 4.18: The average final distribution of agents for experiment 1 when multiple ToM orders survived, plotted for each learning speed. The error bars indicate one standard error.

Dominant Species

Besides the final distribution of agent types, the dominance frequencies of each ToM order were also collected. The average dominance frequencies can be found in Figure 4.19. The error bars indicate one standard error. Since more than one species can be dominant at a time, the summed frequencies can exceed 1.0. The figure shows that the dominance frequency of the ToM0 agents lies close to 1.0. There is a small effect of the learning speed: the dominance frequency of the ToM0 agents was slightly, but significantly, for $\lambda = 0.2$ than for the other two values. The average dominance frequencies of ToM1 and ToM2 can be inspected in Figure 4.20, which displays an enlarged version of a section of Figure 4.5. From this figure, it follows that there was no significant difference in the average dominance frequency of ToM1 and ToM2, but

that both are significantly smaller than that of ToM0. Furthermore, it shows that the frequency of both ToM1 agents and ToM2 agents was significantly lower for $\lambda = 0.2$ than for the other learning speeds.

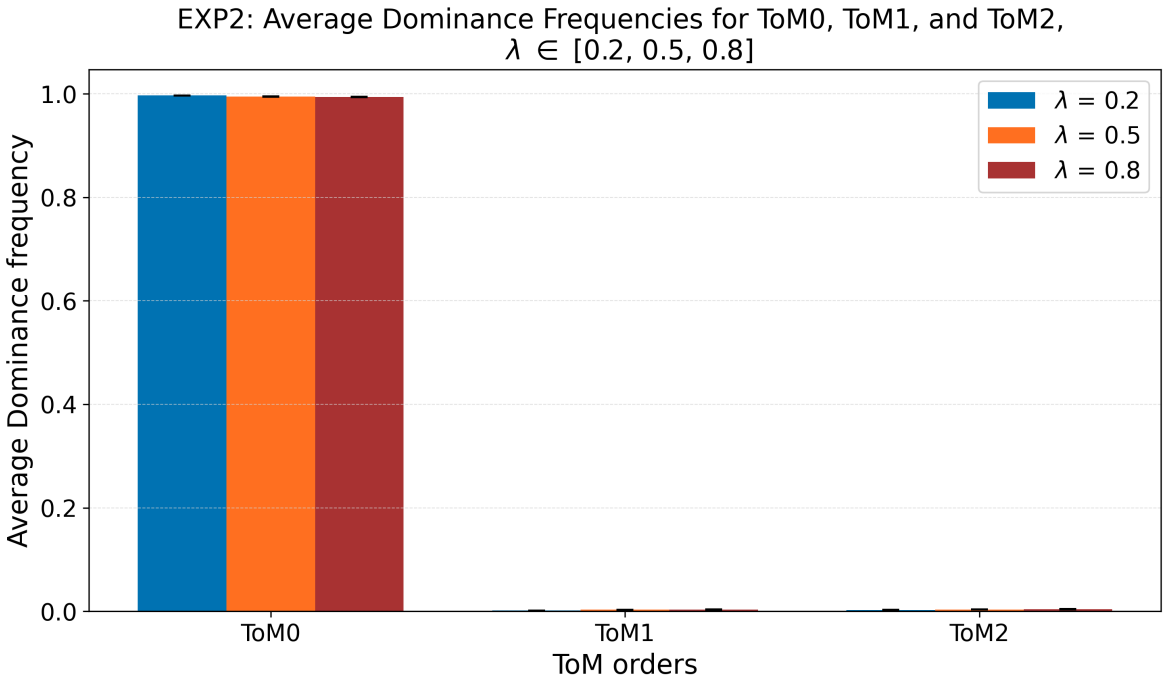

Figure 4.19: The average dominance frequency for experiment 2 per ToM order, plotted for each learning speed. Frequencies do not add to one because multiple types can be dominant simultaneously. The error bars indicate one standard error.

Typical Run

The evolution of the ToM orders over time was collected for a selection of the experiments. Figure 4.21 shows a typical run of the experiments with a learning speed of 0.2. ToM0 became the dominant order almost straight away. The graph shows that ToM2 went extinct within the first 100,000 ticks. Mutations allowed an agent of ToM2 to enter the environment again a few times, but these quickly died. ToM1 on the other hand, survived until the end of the experiment, which was terminated due to the limit being reached. The experiment thus ended with two species in the environment: ToM0 and ToM1.

A typical run for the experiments with a learning speed of 0.5 can be found in Figure 4.22. The graph shows that ToM1 (red: dashed) was the dominant species at the start of the exper-

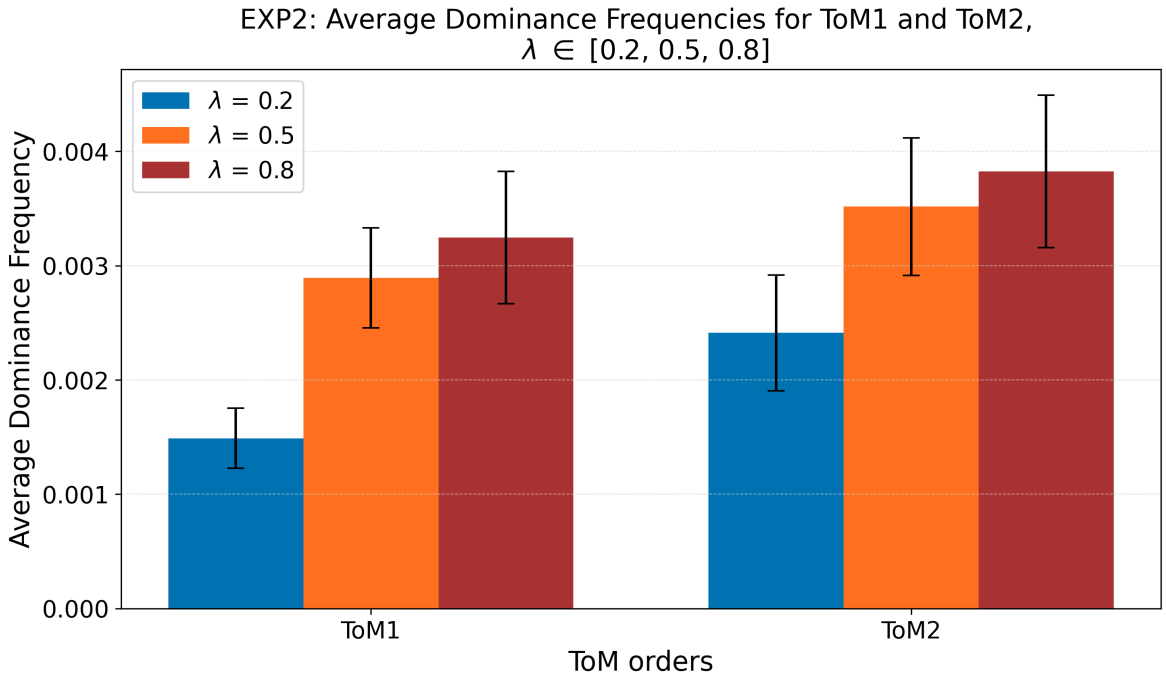

Figure 4.20: The average dominance frequency for ToM1 and ToM2: an enlarged version of a section of Figure 4.19. The error bars indicate one standard error.

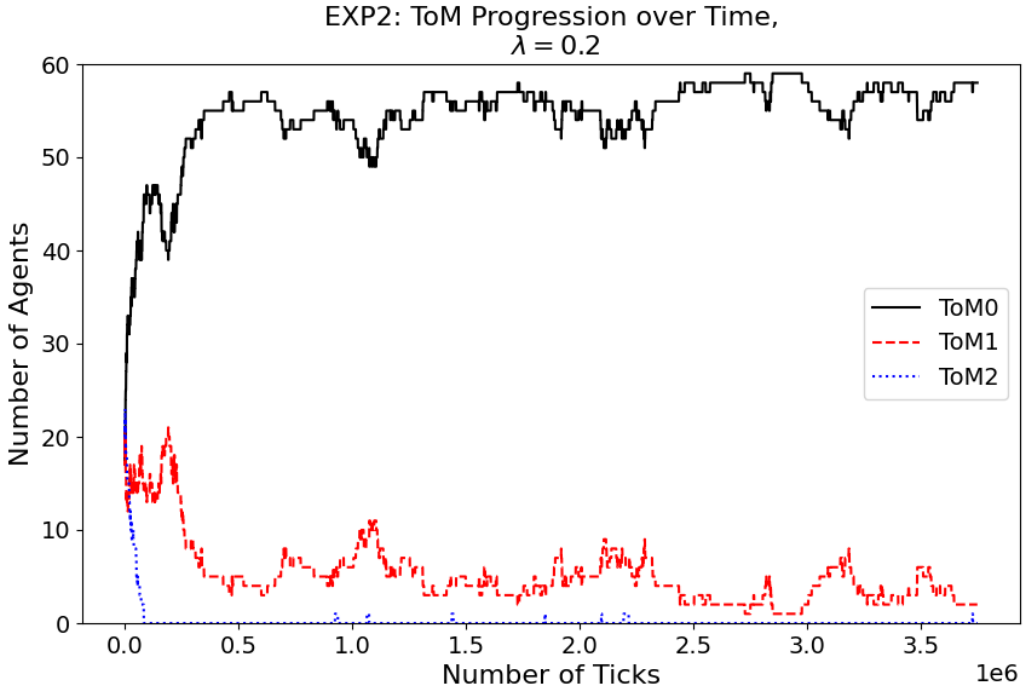

Figure 4.21: An example of a time-lapse of the distribution of agent types for the experiments of type 2 with a learning speed of 0.2. This example is representative of the other runs.

iment. Apart from that, the observed pattern was very similar to that of the experiments with $\lambda = 0.2$. This experiment ended due to both ToM1 and ToM2 going extinct.

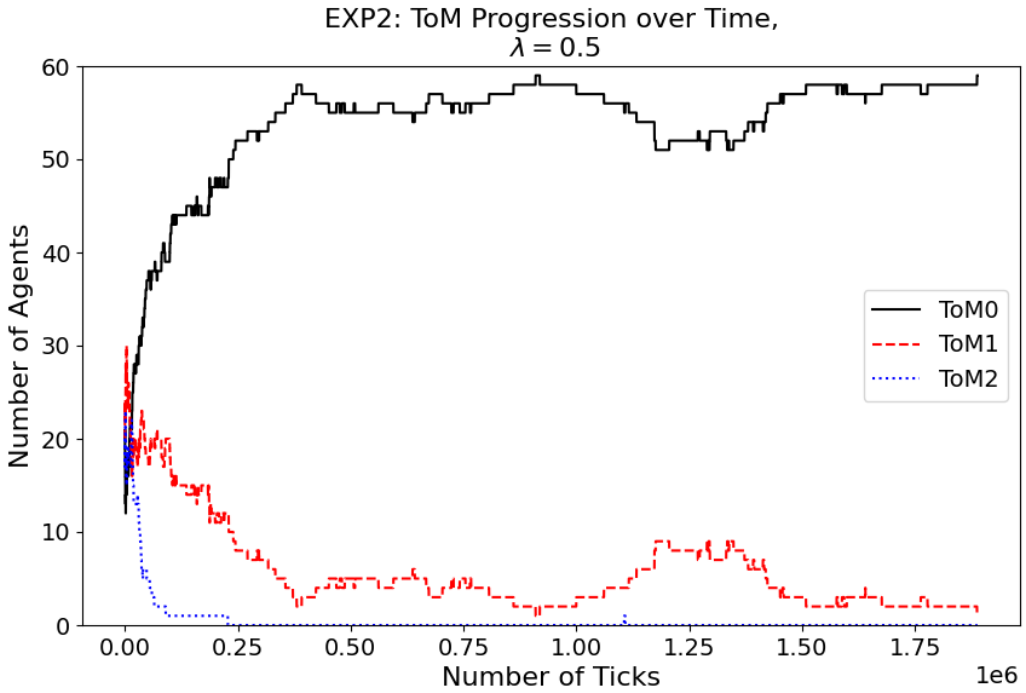

Figure 4.22: An example of a time-lapse of the distribution of agent types for the experiments of type 2 with a learning speed of 0.5. This example is representative of the other runs.

Finally, Figure 4.23 shows the evolution of ToM in a typical run of an experiment with a learning speed of 0.8. The pattern that can be observed is very similar to that of the experiments with a learning speed of 0.5. ToM1 was dominant at the start of the experiment, although very briefly. The experiment terminated because ToM0 was the only surviving species.

Agent Ages

The ages of all agents were collected and plotted per order of ToM. The results corresponding to a learning speed of 0.2, 0.5, and 0.8 were all very similar, so only those of $\lambda = 0.5$ are discussed, and the others can be found in Appendix B.

Figure 4.24 shows the distribution of the ages of the agents in experiments with a learning speed of 0.5. The ages did not exceed the lower bound of zero, even though it may appear so in the figure. This is a result of the wide spread of ages. The figure shows that ToM0 agents had the highest variety in ages. The majority of the ToM0 agents died relatively young, which can

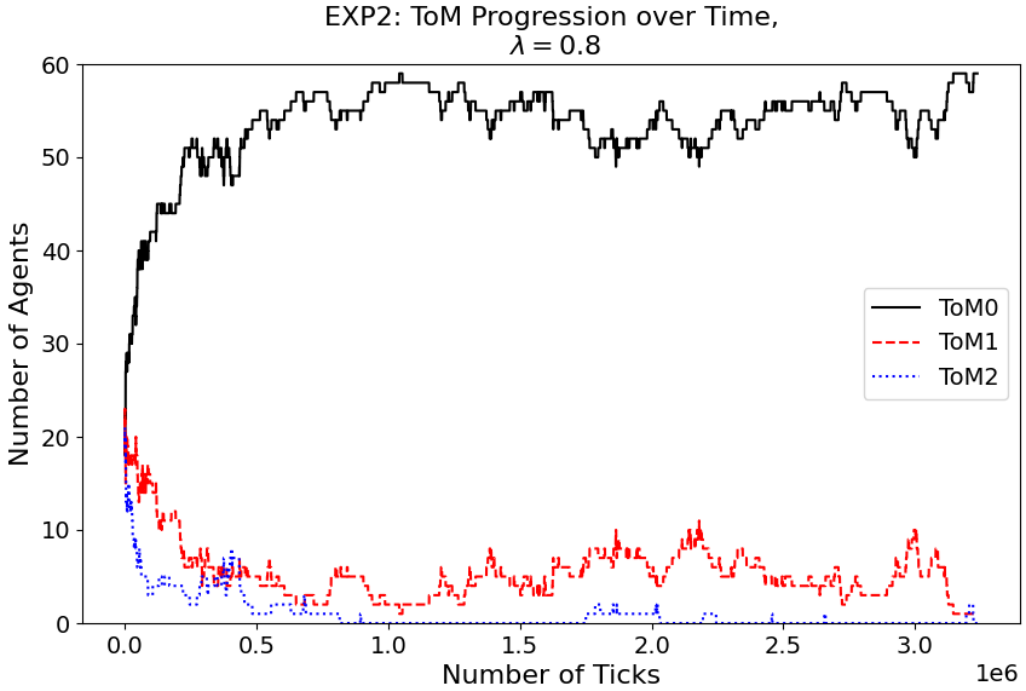

Figure 4.23: An example of a time-lapse of the distribution of agent types for the experiments of type 2 with a learning speed of 0.8. This example is representative of the other runs.

be concluded from the wide bottom of the violin shape. The length of the upper tail shows that some agents reached very high ages, up to approximately 2.0 million ticks. The distribution of ages of the ToM1 agents follows a similar pattern, except that the bottom of the violin is slightly wider and the upper tail is shorter. This indicates that more agents died young compared to the ToM1 agents. The violin plot of ToM2 has a shorter upper tail, indicating that agents of this type generally did not survive as long as the other types. The positions of the medians, shown by white lines in each violin, confirm that most agents across all ToM orders had relatively short lifespans. Additionally, the median age of the agent decreased as the ToM order increased.

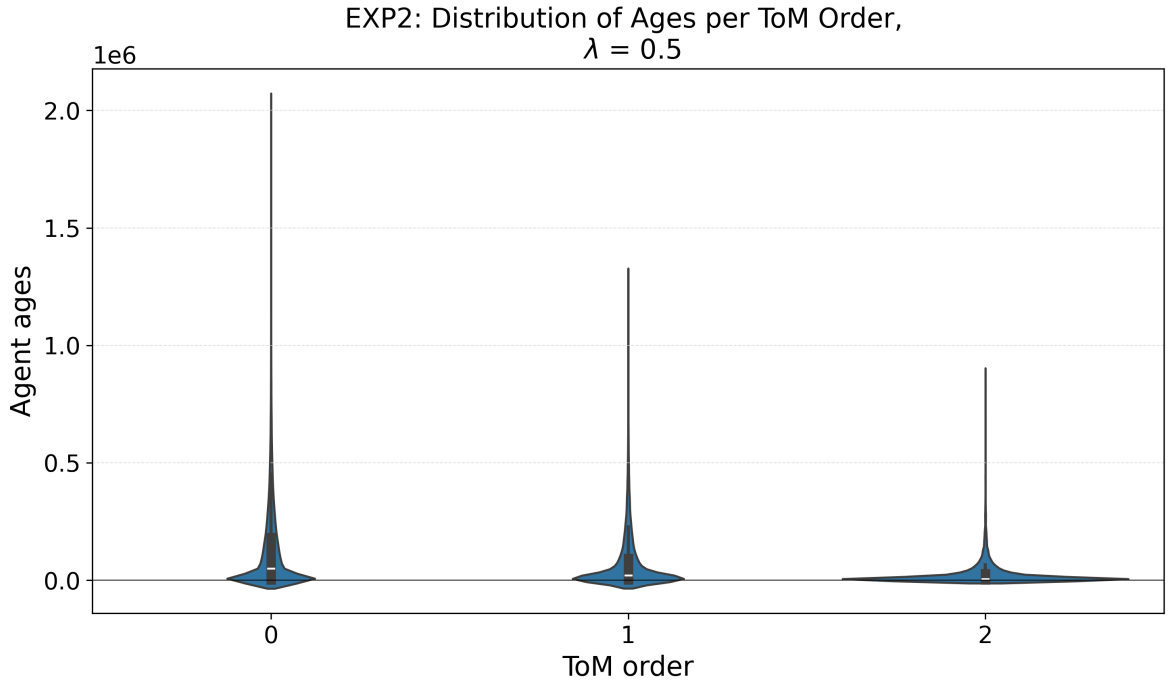

Figure 4.24: A violin plot of the spread of agent ages of experiment 2 per ToM order, for the experiments with a learning speed of 0.5.

Negotiation Lengths

Multiple types of data were collected from the negotiations. Firstly, the negotiation lengths were collected and averaged over all the negotiations per learning speed and negotiation type. The results can be found in Figure 4.25, which includes the corresponding error bars as well, which indicate one standard error. The figure shows that except for the ToM1-ToM1 negotiation type, for each type the average negotiation length was highest for a learning speed of 0.2, then

0.5, and then 0.8. For the ToM1-ToM1 type, the length was highest for $\lambda = 0.5$, followed by 0.2 and then 0.8. Each of these differences was significant, as can be concluded by the error bars (they do not overlap). The impact of the learning speed was highest for negotiations of type ToM0-ToM0, with average lengths of 16.6, 7.7, and 4.5 for $\lambda = 0.2$, 0.5, and 0.8 respectively.

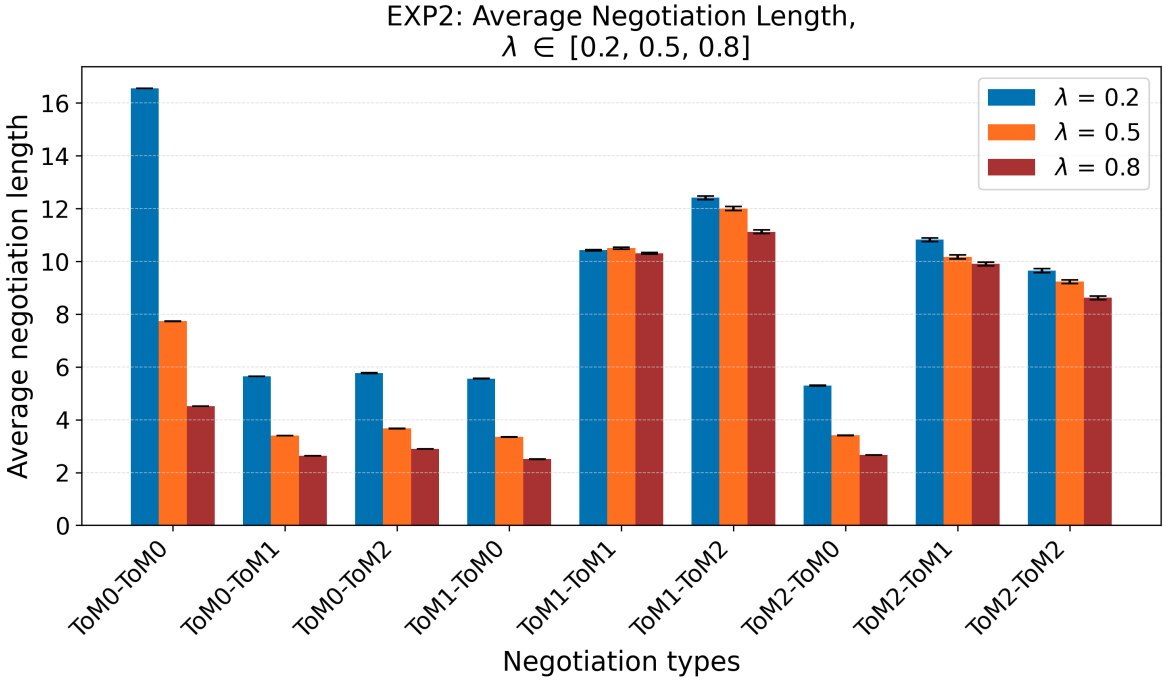

Figure 4.25: The average negotiation length for experiment 2 per negotiation type, plotted for learning speeds 0.2, 0.5, and 0.8. The negotiation type specifies the order of the initiating agent and its trading partner. The error bars indicate one standard error.

Since the graph contains a substantial amount of information, its contents are discussed per learning speed. For a learning speed of 0.2, the average negotiation length was highest (16.6) for negotiation type ToM0-ToM0. Each of the differences in average length between the different negotiation types for this learning speed was significant. After ToM0-ToM0, the next negotiation types with the highest average length were ToM1-ToM2, ToM2-ToM1, ToM1-ToM1, and ToM2-ToM2. The shortest negotiations were of the types ToM1-ToM0, ToM0-ToM2, ToM2-ToM0, and ToM0-ToM1 respectively, terminating after 5-6 rounds on average. This shows that all negotiation types with exactly one ToM0 agent had the shortest average length.

For a learning speed of 0.5, the longest average negotiation length belonged to the negotiation type ToM1-ToM2 with an average length of 12.0. After this, ToM1-ToM1, ToM2-ToM1,

and ToM2-ToM2 were the negotiation types that took the most time, respectively. These differences in average negotiation length were all significant. The shortest negotiations (on average) were those where at least one ToM0 agent participated: ToM0-ToM2 was followed by ToM0-ToM1 and ToM2-ToM0. The latter two did not have a significantly different average negotiation length: both took 3.4 rounds on average. The shortest negotiations were those of type ToM1-ToM0.

For a learning speed of 0.8, a similar pattern can be observed: the negotiation types with the highest average length were ToM1-ToM2 and ToM1-ToM1. These types were followed by ToM2-ToM1 and ToM2-ToM2. Again, the negotiation types with a ToM0 agent had the lowest average length. Each of the differences in average negotiation lengths was significant. The mentioned negotiation types were followed by ToM0-ToM0, ToM0-ToM2, ToM2-ToM0, ToM0-ToM1, and ToM1-ToM0 respectively, with the latter having an average experiment length of only 2.5.

Reasons for Negotiation Termination

The reasons for the termination were collected for each negotiation. There are five options: either of the agents withdrew, accepted, or the negotiation limit (of 50) was reached. The collected data were plotted per learning speed. Each graph is a stacked bar plot where each bar corresponds to one of the nine negotiation types. Figure 4.26 shows the results for a learning speed of 0.2. The plot shows that the majority of the negotiations did not result in a successful trade: for none of the negotiation types, the summed frequency of either of the agents accepting was higher than 30%. Most of the negotiations terminated due to a withdrawal of one of the agents or due to the negotiation limit being reached. The negotiation types with the highest success rates were ToM0-ToM0, ToM1-ToM0, and ToM2-ToM0. These are all the negotiations where the non-initiating agent was a ToM0 agent. The negotiation types with the lowest success rate were ToM1-ToM1 and ToM1-ToM2.

The plot shows that the most occurring reason for the termination of the negotiation was a withdrawal of the initiator, with an exception for the ToM0-ToM1 and ToM0-ToM2 negotiations, where the trading partner withdrew more frequently. The latter is also relatively frequent for negotiation types ToM0-ToM0, ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2. For each of the negotiation types, a percentage of them ended due to the negotiation limit being reached. This percentage was relatively high (above 15%) for the types ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2. These are all the types where no ToM0 agent participates. Furthermore, the results reveal that for the negotiations of type ToM0-ToM1, ToM0-ToM2, ToM2-ToM1, and ToM2-ToM2, the frequency of the initiator accepting was higher than that of the trading partner accepting. Additionally, the other types (ToM0-ToM0, ToM1-ToM0, ToM1-ToM1, ToM1-ToM2, and ToM2-ToM0) had a higher frequency of the trading partner accepting.

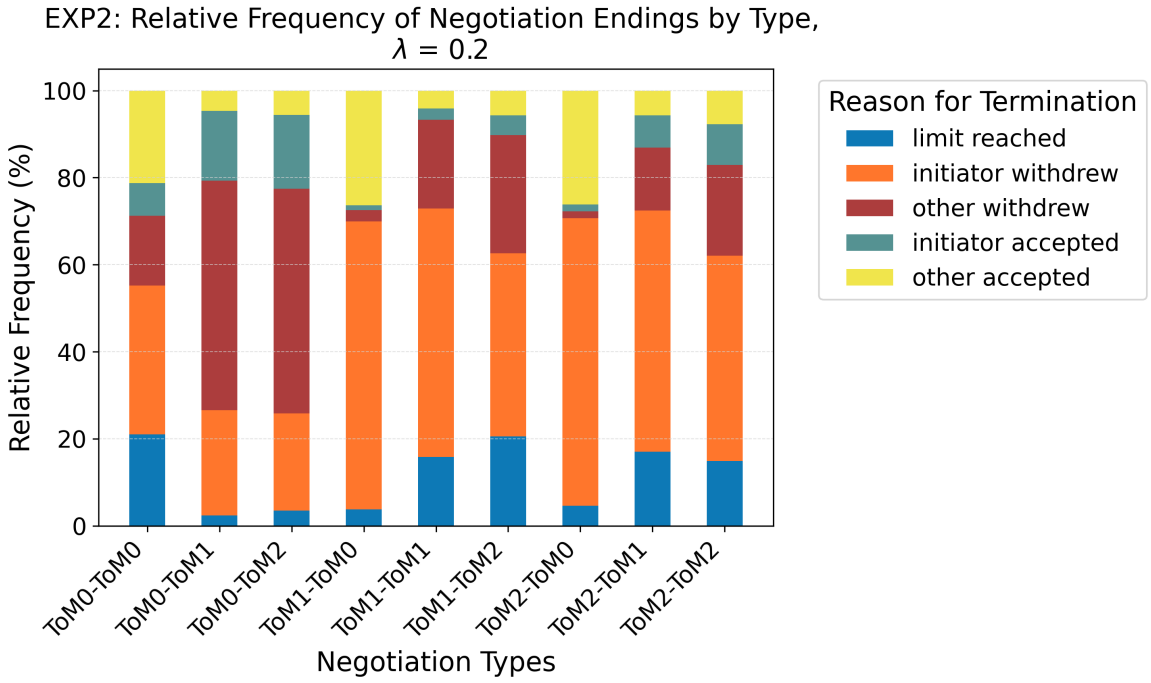

Figure 4.26: The reason for negotiation termination for experiment 1 per negotiation type, for the experiments with a learning speed of 0.2.

The relative frequencies of the negotiation termination reasons for the experiments with a learning speed of 0.5 can be found in Figure 4.27. The results are quite similar to those for $\lambda = 0.2$. There is one main difference. The negotiation limit was only reached for negotiations of type ToM1-ToM1, ToM1-ToM2, ToM2-ToM1, and ToM2-ToM2. However, still 70% of the negotiations did not end in a successful trade.

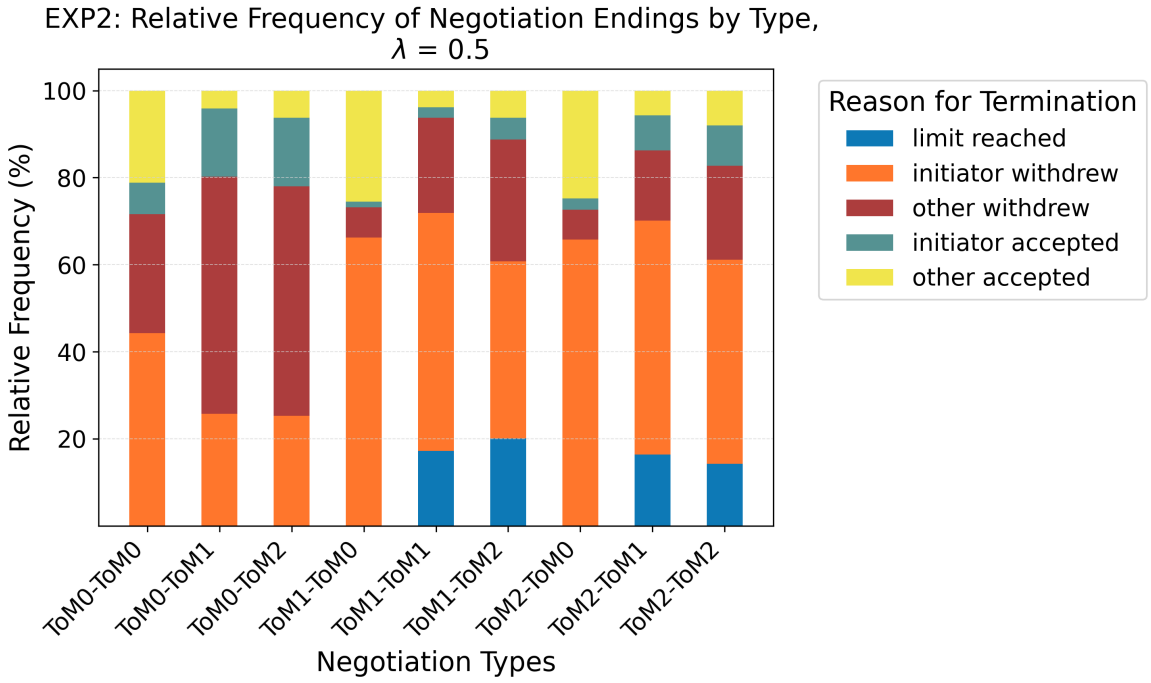

Figure 4.27: The reason for negotiation termination for experiment 2 per negotiation type, for the experiments with a learning speed of 0.5.

The results for the experiments with a learning speed of 0.8 show no notable differences from those for the experiments with a learning speed of 0.5. The graph can therefore be found in Appendix B.

Negotiation Gains

The usefulness of a negotiation is determined based on the gain that the participating agents obtain. The gains were collected for the initiator and its trading partner in each negotiation. The average gains are displayed per learning rate, for each negotiation type.

Figure 4.28 shows the average gains for the negotiations in the experiments with a learning speed of 0.2, along with error bars that indicate one standard error. The graph shows that there

were two negotiation types where the initiating agent received a significantly higher average gain than its trading partner. This was the case for negotiations of type ToM1-ToM0 and ToM2-ToM0. For the other negotiation types the trading partner received a significantly higher gain than the initiator. The only exception is the negotiation type ToM1-ToM2, where no significant difference was found. The largest difference in average gain was found for negotiation type ToM0-ToM1, where the difference was 0.7.

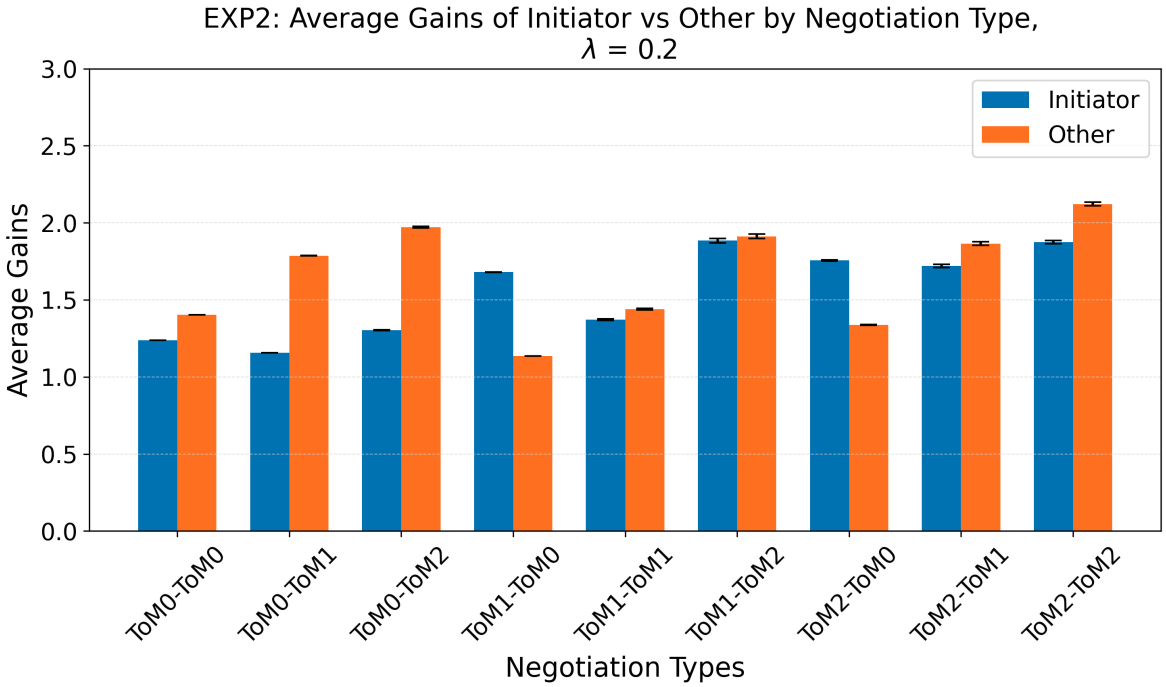

Figure 4.28: The average gains of the initiating agent and its trading partner for each negotiation type of experiment 2, plotted for the experiments with a learning speed of 0.2. The error bars indicate one standard error.

None of the averages exceeded a gain of 2.2 since the highest was 2.1, corresponding to the non-initiating agent in a negotiation of type ToM2-ToM2. The non-initiating agent of the ToM0-ToM2 negotiation type received the second-highest gain on average. This was followed by both the initiating agent and its trading partner in a ToM1-ToM2 negotiation. After that, the non-initiating agent of the ToM2-ToM1 negotiation type and the initiating agent of the ToM2-ToM2 type both received the highest gain on average, with no significant difference between their average gains. The lowest gains (on average) were obtained by the initiating agent in the ToM0-ToM1 negotiation, and the non-initiating agent in the ToM1-ToM0 negotiation.

The results thus show that the ToM1 agent received the highest average gain in most of the

negotiation types that it was part of. On the contrary, ToM0 received the lowest average gain in each of the negotiation types that it was part of. The figure furthermore shows that the average gain of an agent type is dependent on the order of ToM of its trading partner. An example is that a ToM1 obtained a higher average gain when negotiating with a ToM2 agent than when negotiating with another ToM1 agent. The average gain also depends on which of the agents is initiating. An example is a ToM1 agent negotiating with a ToM2 agent: the former received a slightly higher gain when initiating than when the ToM2 initiated.

The results for the experiments with a learning speed of 0.5 can be found in Figure 4.29. Again, the error bars indicate one standard error. The average gains were very similar to those found for the experiments with a learning speed of 0.2. One difference is that now, the initiator of a ToM1-ToM2 negotiation obtained a significantly higher gain on average that its trading partner. This gain was also higher than that obtained by the non-initiating agent of the ToM0-ToM2 type. Furthermore, initiator of a ToM2-ToM1 negotiation type now obtained a higher average gain that the initiator of a ToM2-ToM0 negotiation.

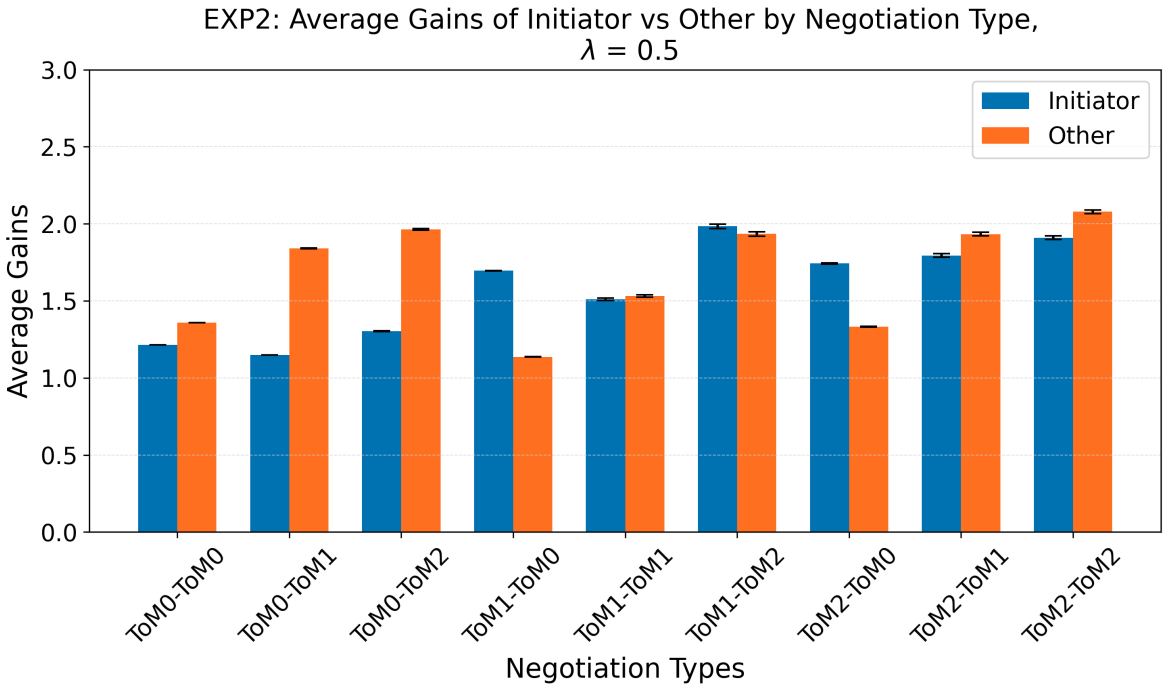

Figure 4.29: The average gains of the initiating agent and its trading partner for each negotiation type of experiment 2, plotted for the experiments with a learning speed of 0.5. The error bars indicate one standard error.

The average gains for the experiments with a learning speed of 0.8 can be found in Figure

4.30. The results are mostly similar to those from the experiments with a learning speed of 0.5. The main difference is that now, the initiating agent of a ToM1-ToM1 received a higher average gain that its trading partner. The obtained average gain of the ToM1 agent in the ToM1-ToM2 type increased further compared to the results for the learning speed of 0.5.

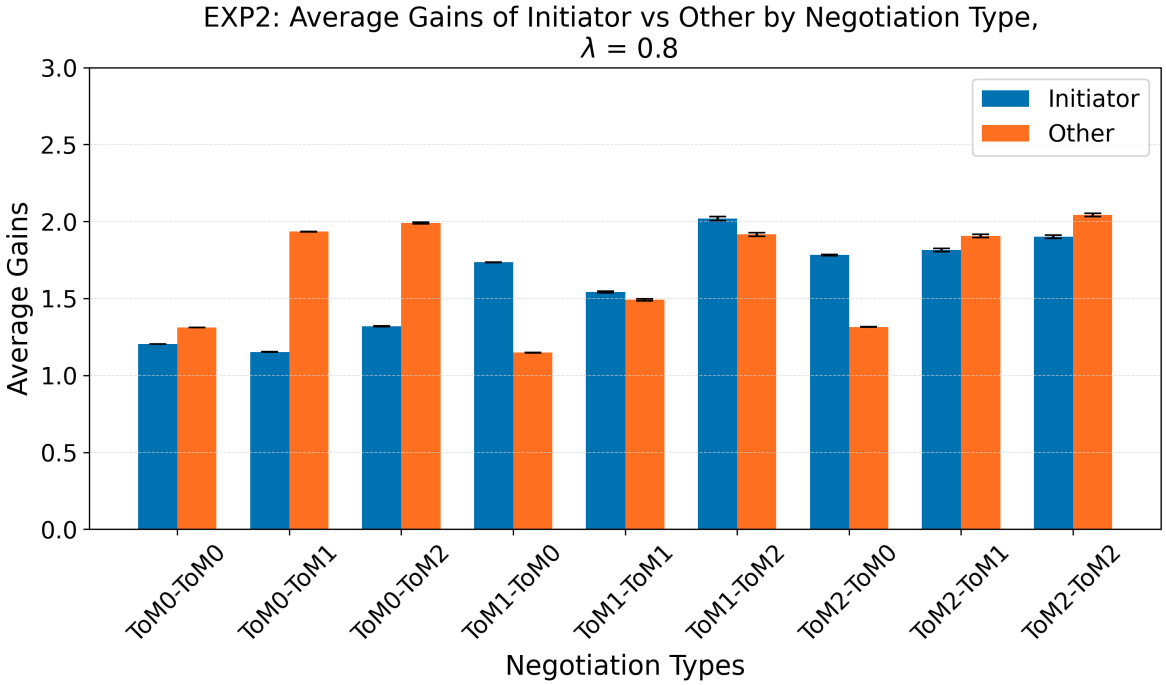

Figure 4.30: The average gains of the initiating agent and its trading partner for each negotiation type of experiment 2, plotted for the experiments with a learning speed of 0.8. The error bars indicate one standard error.

Chapter 5

Discussion

Humans are remarkably good at projecting mental content to others, i.e., at using Theory of Mind (ToM), compared to other animals. For this quality to develop, there must have been an evolutionary benefit, especially for the computationally highly expensive higher-order ToM, where ToM is used recursively. Previous research suggested that mixed-motive settings may have been the reason for humans to develop higher-order ToM. Mixed-motive settings are those settings where a combination of competition and cooperation is required, like in a negotiation. However, previous research focuses solely on simulations with pairwise interactions.

This study aimed to find out if the conclusions that were previously drawn about the social advantages of using ToM in pairwise interactions in a mixed-motive setting also translate into advantages on a population level. A population containing agents with various orders of ToM was simulated. The research question was: *How do various orders of Theory of Mind evolve in a population of agents placed in a mixed-motive environment?*

To answer this research question, two experiments were conducted. The first experiment was constructed to represent a mixed motive setting where negotiations are used to collect resources that are needed to survive in an evolutionary process. The environment contained a population of agents that could have zero-order, first-order, or second-order ToM that they could use during the negotiations. Agents could negotiate in pairs, and these individual negotiations were an adaptation of the implementation of Colored Trails (CT) in De Weerd et al. (2017). When the agents did not reach the resources threshold, they were replaced by a new agent whose ToM order was sampled from the population, with a change of mutation. The results of the negotiation thus played a role in the dynamics of the population.

The second experiment was constructed to more closely resemble a real scenario. As an example, if you recently had an unsuccessful negotiation with a company, you will likely avoid a new negotiation with this company in the near future. Furthermore, you may have learned some company-specific information during a negotiation in the past, like their preferred purchasing quantities. In a future negotiation, you may use this remembered information. These two

features were reflected in the second experiment. Apart from these alterations, the experiment was identical to the first experiment.

This section will first present the interpretation of the results that were found (Section 5.1). Then, the limitations of the research are discussed (Section 5.2), followed by suggestions for future research (Section 5.3). Finally, the implications and insights of this research are discussed in the conclusion (Section 5.4).

5.1 Interpretation of the Results

The first experiment has shown that ToM0 agents were far more evolutionarily successful than ToM1 and ToM2 agents. This became evident from both the average final distributions and the average dominance frequencies of the species. In the vast majority of the cases, ToM0 agents were the only type to survive. ToM2 agents generally had the upper hand in the first moments of the evolutionary process but then went extinct soon after. If multiple species lasted until the end of the experiment, these were always ToM0 and ToM1. The success of the zero-order agents also became evident from the ages of the agents. Although many agents lasted around 50,000 ticks (20 generations), the ToM0 agents had the most 'dinosaurs': agents that lived for more than a million ticks.

The negotiation-specific results provide further insight into the evolutionary process of the ToM orders. The gains that the agents received after a trade indicate that qualitatively, ToM1 agents made the most rewarding deals. This was especially true for agents with a high learning speed (0.8). In this scenario, ToM1 agents obtained a higher average gain than their trading partners in each of the negotiation types. These results differ from those found by De Weerd et al. (2017), who found that ToM2 agents received the highest negotiation gain. De Weerd and colleagues furthermore found that ToM1 agents received lower gains than ToM0 agents because the ToM1 agents suffered from trying to reach a cooperative solution. Our results show that ToM2 agents generally caused the gain of their trading partner to be higher than if that trading partner were to negotiate with another agent type. This is similar to the findings by de Weerd and colleagues that ToM2 agents increase social welfare when they maximize their personal gain. In the current research, the least rewarding trades were made by the ToM0 agents, who obtained a lower gain than their trading partners in each negotiation type. These results suggest that first-order and second-order agents are more likely to reach their resource threshold than zero-order agents.

However, the results furthermore showed that the average negotiation length for negotiations with two agents with an order above zero could last up to eight rounds longer than those where a ToM0 agent participated. An exception was the negotiations with two zero-order agents, which took longer than if this agent negotiated with an agent of another type. In the CT setting used by De Weerd et al. (2017), the negotiation length was also influenced by the ToM of the par-

ticipating agents: the research showed that whilst a negotiation took a maximum of 15 rounds on average, there was an exception for negotiations with a ToM2 agent, which took longer. In the current research, the learning speed of the agents also played a large role in the negotiation length: a lower learning speed resulted in longer negotiations. Both factors, i.e., ToM order and learning speed, are also reflected in the termination reasons of the negotiations: negotiations with non-ToM0 agents frequently ended without a successful deal due to the negotiation limit being reached. For experiments with a learning speed of 0.2, this limit was even reached for negotiations with a zero-order agent. Another frequent outcome was a withdrawal. The qualitative research showed that often no good deal exists, which is why this high withdrawal rate is intuitive.

The agents with an order above zero, but especially the ToM1 agents, thus tend to make better deals than ToM0 agents. However, it takes significantly more time for them to reach this negotiation outcome. Given that in the majority of the negotiations there may not even be a mutually beneficial deal, the time it takes higher-order agents to terminate a negotiation decreases their chances of survival. This is a key difference to the research by De Weerd et al. (2017), where each negotiation at least had the potential of being successful. The results showed that ToM0 was the most successful species, even though they are at a relative disadvantage in pairwise negotiations when compared to ToM1 and ToM2 agents. In this experimental setting, the time pressure to reach the threshold was too high for ToM to provide an advantage, causing the skill to disappear over time.

The results of the second experiment were for the most part similar to those of experiment one, but in some ways, they differed. The number of experiments where multiple ToM orders survived until the end of the experiment was larger than for the first experiment. Apparently, the remembrance of past experiences increases the survival chances of ToM1 and ToM2 agents. This was also reflected by the observed final distributions, which showed that both ToM1 and ToM2 occasionally survived until the end of the experiment.

Another interesting difference compared to the first experiment is that the negotiation lengths were generally shorter. Additionally, the effect of the learning speed on the negotiation length of negotiations between two ToM1 agents was different than observed previously: the negotiation length was now highest for a learning speed of 0.5. The frequency of the negotiations where the limit was the termination reason was lower compared to the previous experiment. The frequency of either of the agents accepting increased slightly. This was the case for each of the three learning speeds.

These results show that saving the trading-partner-specific information mainly allowed agents to make a decision earlier in the negotiation process: prior information like the producing resource enabled the agents to withdraw immediately if necessary or make a good offer right away. However, also in a more true-to-life experimental setting, ToM0 agents were most fit to survive, although the survival chances of the other species were slightly higher now. The negotiation lengths were impacted, but the outcomes of these negotiations showed little difference.

The reduction of negotiation lengths was not large enough to give ToM orders above zero an evolutionary advantage.

5.2 Limitations

The research that was conducted contains a few limitations. These limitations are described, as well as the influence they may have had on the results.

Observable environment To make the experimental environment manageable, the design choice was made so that all agents can view which resources their trading partner has. This choice was partly made to lower the computational complexity of the program, but also to resemble the observable nature of the CT setting of De Weerd et al. (2017). However, it is not highly realistic in this evolutionary application. In a real-life trade, we do not necessarily reveal to our trading partner which resources we have. Instead, we may use it to our benefit that the other person does not know have much it can ask. This may have impacted the results: possibly more 'eager' offers were made, in turn resulting in fewer acceptances.

Limited offer possibilities Another concession that was made to reduce the computational complexity of the program, was limiting the number of resource types and the maximal stock. Ideally, many more offers would be possible to allow for more negotiations with a successful outcome. If the proportion of negotiations with a predetermined negative outcome were to be smaller, maybe the time cost of ToM1 and ToM2 would have been worth the increased gain compared to ToM0. The current results may thus favor ToM0 agents.

Producing resource Another limitation of this research is the implementation of the producing resource. The idea was to set the stock for this resource to four (the maximum) to allow for free trade where everyone has at least one type of resource to offer. In practice, this made it very easy for the agents to guess the producing resource of the trading partner, since the agent always has four of this type. Also, ToM2 might normally want to mislead their trading partners by asking for their producing resource. Due to the implementation, this was not possible. This may have reduced the advantage of ToM2 agents in negotiations.

Population size The population size of 60 was a design choice that likely influenced the results. A smaller size would have caused fewer negotiations, since agents encounter fewer other agents. The small qualitative inspection that was done suggests that this smaller population may cause the ToM1 and ToM2 agents to have a better chance at survival: the ToM0 agent may not benefit enough from its fast negotiations in this setting.

Despite these limitations, some patterns of benefits in pairwise interactions that we observed were in line with previous literature (De Weerd et al., 2017). More specifically, we also found that ToM1 can help a negotiation reach a successful outcome, as well ToM2 increasing the gain of the trading partner. We also found that ToM0 can perform relatively well in negotiations, benefiting from the ToM of their partner. However, the use of the (adapted version of the) ToM model by De Weerd et al. (2017) led to some irrational behavior in this new setting. These are not necessarily limitations since we wanted to compare the results to those found by De Weerd and colleagues, but they are important to consider when drawing conclusions based on the findings.

Repeating offers The main drawback of the ToM implementation is that repeated offers occur frequently. The qualitative observations showed that this repetition was a consequence of the confidence values, which start off with a value of one. Agents did not use their acquired beliefs, which is why they could repeat offers an infinite number of times. The repeated offers were especially a reoccurring issue for ToM1 and ToM2 agents, and they may have greatly influenced their evolutionary success. The repetition may have been avoided by using a time penalty in the negotiations as was done by De Weerd et al. (2017). Instead, in this research we opted for a mechanical time pressure, forced by the evolution in the environment.

Offer beliefs Another point for attention is the non-intuitive use of beliefs about offers. Due to the implementation of belief updates, frequently encountered offer types receive lower beliefs than others. As a result, agents may make greedy offers, which are less likely to be accepted. In combination with the belief updates for specific offers, which decreases the beliefs for offers where it gives less of a resource than it was asked for, this may lead to unnecessary withdrawals.

5.3 Future Research

The current research expands Theory of Mind (ToM) studies by applying it to simulated populations rather than just to pairs of simulated agents. Previous research by Lenaerts et al. (2024) also focused on the evolution of ToM in a population, but they modeled the evolutionary process using evolutionary game theory. The biggest difference to their research is that the current research modeled a population of individuals, where each agent had their own beliefs, making each individual unique. The order of ToM determined the species and behavior of the agents, but the behavior was also influenced by their own unique past experiences. Since the study of the evolution of ToM in a population in this sense is a (to our knowledge) new area of research, this study can be expanded and used in numerous ways.

The first suggestion for future research is to validate the findings of this research. Several design choices, specifically those mentioned in the previous section, may have impacted the results. We suggest focusing especially on the repeating offers since we expect that this is the reason for the large number of negotiation rounds for first-order and second-order agents. We propose initializing the confidence scores differently so that offer beliefs will impact the action choices of the ToM1 and ToM2 agents. Optionally we propose creating a new version of the ToM model. The original implementation by De Weerd et al. (2017) was developed especially for the CT setting. In this research, only some slight adaptations were made to apply the same model to a new setting where there is not a complete redistribution of the resources/chips. Developing a ToM model specifically for this setting may give more realistic results.

Another insightful adaption of the current research would be to use the same environment, but replace the negotiation interactions between the agents with pairwise CT games. Since the logic of the pairwise interactions would then be the same as in De Weerd et al. (2017), the results could show whether our finding that the pairwise advantage of ToM did not translate into population-wise advantages is a consequence of our implementation of the negotiation setting. We expect that in this scenario, the pattern of repeating offers would be omitted, thereby reducing the negotiation lengths. In turn, this may cause ToM1 and ToM2 agents to have a better shot at survival.

This research only implemented three orders of ToM, and thus only one species with higher-order ToM: ToM2. However, previous research has shown that in some scenarios, higher-order ToM like ToM4 may provide an additional social advantage De Weerd et al. (2014). The ability of humans to reason with ToM4 also highlights its evolutionary benefit. A suggestion for future research is therefore to expand the current research with additional species, i.e., ToM orders. An analysis of the evolutionary dynamics can provide further insight into the validity of the Mixed-motive interaction hypothesis.

As a further step towards the application of ToM research to multi-agent systems, it would be interesting to extend this research by allowing groups of agents to negotiate. The current research does focus on ToM development in a population, but the negotiations are still pairwise only. This may not fully capture the complexity of real-world social interactions where multiple agents may negotiate simultaneously. Building a framework for mixed-motive group interactions within a population would make the simulation more realistic. These group negotiations represent scenarios like team meetings or markets. The complex and dynamic environment that this would create may make the evolutionary advantage of ToM more apparent.

Finally, another interesting direction for future research is to make the agents more goal-oriented. In this research, the agents have a stock goal to reach, but apart from that, their behavior within the environment is quite random. The agents are simply moving in a straight line until they encounter an agent or a wall. If they move away from either of the two, their new direction is random. The second experiment of this research reduced some of that randomness by allowing agents to avoid negotiating with certain trading partners, but this can be taken

further by allowing agents to physically avoid or search certain trading partners. This would lead to more goal-directed movement within the environment. Consequently, the results of the renewed simulation could be more representative of the true evolutionary process.

5.4 Conclusion

This research was conducted to address the gap in understanding why humans developed (higher-order) ToM, whilst other animals do not seem to have this cognitive ability to this extent. It builds of the work of De Weerd et al. (2013b), De Weerd et al. (2017) and De Weerd et al. (2014) where pairwise interactions were used to test the Mixed-motive interaction hypothesis. That is, a mixed-motive setting was established to check which agent types, and thus which ToM orders, gave an advantage in that setting. This research extends the previous research by testing the benefits of ToM in a mixed-motive setting where the agents exist in a population.

To answer the research question of this study, two sub-questions were constructed. The first question was: *Is there an order of ToM that is the 'winner' in this environment, or is a dynamic equilibrium reached?* The answer to this question is that in the majority of the cases, only one species survived. There were situations where multiple species lasted until the end of the experiment, but we cannot speak of an equilibrium. In these cases, there was always one species that had over 50 members, whilst the other had just a few members left. There was thus a clear winner of the evolution. The second question was: *Do lower orders of ToM (ToM0, ToM1) go extinct over time? In other words: does higher-order ToM provide an evolutionary benefit in this negotiation environment?* The answer to this question is no. In this experimental setup, it was the other way around. ToM1 and ToM2 went extinct, whilst ToM0 survived. (Higher-order) ToM thus did not provide an evolutionary benefit in this negotiation environment.

Some of the findings in this research are in line with previous findings. An example is that ToM0 agents can perform well in a negotiation, depending on their partner. We found that ToM0 agents can receive a relatively high gain when a ToM1 or ToM2 agent initiates a negotiation with them. Within the negotiations, we also saw that ToM1 and ToM2 agents benefit from their ToM, and ToM0 agents benefit from that too. A difference was that in previous research, ToM0 agents received higher gains than ToM1 agents. In this research, this was not the case.

The previous research argued for the Mixed-motive interaction hypothesis, since the ToM benefits agents in mixed-motive settings. Our results suggest that even though in the negotiations ToM benefits agents, the time cost of using this ToM is relatively high. This prevented the ToM agents from collecting enough resources in time. This research thus shows that the advantages that were observed by De Weerd et al. (2013b), De Weerd et al. (2017), and De Weerd et al. (2014) in pairwise interactions do not translate into population-wise advantages.

This research contributes to a broader understanding of the evolution of ToM in humans. While the Mixed-motive interaction hypothesis suggests that social settings that combine co-

operation and competition drove the evolution of higher-order ToM, our results suggest that population-level dynamics may not necessarily favor higher-order ToM in the way previously thought. Performance in pairwise interactions did not translate into performance at the population level. The research raises questions about alternative social settings that may have caused humans to develop advanced ToM. One promising framework is the Social Brain Hypothesis by Gamble et al. (2014), originating from the ideas presented by Dunbar (1996). Dunbar proposed that language may have developed as a way to create social cohesion in groups through gossip. The groups served as a mutual defense against predators. Gamble et al. (2014) build upon this view that the complexities of managing large social groups may have caused us to develop language. They additionally suggest that this may have been why ToM developed. By examining these broader social contexts, we may expand our understanding of how and why higher-order ToM emerged specifically in humans.

References

Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current Anthropology*, *36*(2), 199–221.

Anbugeetha, D., & Nandhini, B. (2021). Evolution of money: From barter system to digital money. In S. C. B. Samuel Anbu Selvan (Ed.), *The New Era of Digital Payments* (pp. 55–65, Vol. 1). Tamil Nadu, India: Thiagarajar College.

Arre, A. M., & Santos, L. R. (2021). Mentalizing in nonhuman primates. In M. Gilead & K. Ochsner (Eds.), *The Neural Basis of Mentalizing* (pp. 131–147). Cham, Switzerland: Springer.

Arslan, B., Verbrugge, R., Taatgen, N., & Hollebrandse, B. (2020). Accelerating the development of second-order false belief reasoning: A training study with different feedback methods. *Child Development*, *91*(1), 249–270.

Austin, J. L. (1962). *How to do Things with Words* (J. Urmson & M. Sbisà, Eds.). New York City, NY: Oxford University Press.

Avery, J. S. (2003). *Information Theory and Evolution*. River Edge, NJ: World Scientific.

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration*. Princeton, NJ: Princeton University Press.

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 C. Sierra (Ed.), *Proceedings of the Twenty-sixth International Joint Conference on Artificial Intelligence* (pp. 4684–4690). Melbourne, Australia: International Joint Conferences on Artificial Intelligence.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Bowler, P. J. (2000). Evolution. In G. B. Ferngren (Ed.), *The History of Science and Religion in the Western Tradition* (pp. 549–557). Abingdon-on-Thames, England: Routledge.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341.

Bryson, J. J., Ando, Y., & Lehmann, H. (2007). Agent-based modelling as scientific method: A case study analysing primate social behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1685–1699.

Byrne, R. W., & Whiten, A. (1988). *Machiavellian Intelligence: Social Epertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford, England: Oxford University Press.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development*, *70*(2), 381–395.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192.

Center for Information Technology. (2024). *Hábrók Documentation*. `https://wiki.hpc.rug.nl/habrok/start`

Colman, A. M. (2003). Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, *7*(1), 2–4.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life.* London, England: J. Murray.

De Lamarck, J. B. D. M. (1809). *Philosophie Zoologique: Ou Exposition des Considérations Relatives à l'Histoire Naturelle des Animaux* (Vol. 1). Paris, France: Dentu.

De Weerd, H., Broers, E., & Verbrugge, R. (2015). Savvy software agents can encourage the use of second-order theory of mind by negotiators. In D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. Jennings, & P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 542–547). Austin, TX: Cognitive Science Society.

De Weerd, H., Verbrugge, R., & Verheij, B. (2012). Higher-order social cognition in rock-paper-scissors: A simulation study. In W. van der Hoek, L. Padgham, V. Conitzer, & M. Winikoff (Eds.), *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1195–1196, Vol. 3).

De Weerd, H., Verbrugge, R., & Verheij, B. (2013a). Higher-order theory of mind in negotiations under incomplete information. In G. Boella, E. Elkind, B. Savarimuthu, F. Dignum, & M. Purvis (Eds.), *PRIMA 2013: Principles and Practice of Multi-Agent Systems* (pp. 101–116). Heidelberg, Germany: Springer Berlin Heidelberg.

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. (2014). The effectiveness of higher-order theory of mind in negotiations. In J. van Eijck & R. Verbrugge (Eds.), *Reasoning About Other Minds: Logical and Cognitive Perspectives: RAOM 2014. CEUR Workshop Proceedings* (pp. 35–39, Vol. 1208). Aachen, Germany: EUR-WS.org.

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*, 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.

De Weerd, H., & Verheij, B. (2011). The advantage of higher-order theory of mind in the game of limited bidding. *Proceedings Workshop 'Reasoning about other Minds', CEUR Workshop Proceedings*, *751*, 149–164.

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

Dunbar, R. (1996). *Grooming, Gossip, and the Evolution of Language*. Faber; Faber.

Emery, N. J., Dally, J. M., & Clayton, N. S. (2004). Western scrub-jays (Aphelocoma californica) use cognitive strategies to protect their caches from thieving conspecifics. *Animal Cognition*, *7*, 37–43.

Epstein, J. M., & Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Washington, DC: Brookings Institution Press.

Finin, T., Labrou, Y. K., & Mayfield, J. (1996). Evaluating KQML as an agent communication language. In M. Wooldrifge, J. P. Müller, & M. Tambe (Eds.), *Intelligent Agents II – Agent Theories, Architectures, and Languages* (pp. 347–360, Vol. 1037). Berlin, Germany: Springer-Verlag.

FIPA. (1997). *Specification Part 2- Agent Communication Language* (tech. rep.). Technical report, FIPA-Foundation for Intelligent Physical Agents.

Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, *17*, 417–442.

Gamble, C., Gowlett, J., & Dunbar, R. I. M. (2014). *Thinking Big : How the Evolution of Social Life Shaped the Human Mind*. London, England: Thames & Hudson.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, *68*(3), 571–581.

Hedden, T., & Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition*, *85*(1), 1–36.

Hemelrijk, C. K. (1999). An individual–orientated model of the emergence of despotic and egalitarian societies. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*(1417), 361–369.

Hemelrijk, C. K. (2002). Self-organizing properties of primate social behavior: A hypothesis for intersexual rank overlap in chimpanzees and bonobos. In C. Soligo, G. Anzenberger, & R. D. Martin (Eds.), *Primatology and Anthropology: Into the Third Millennium* (pp. 91–94, Vol. 11). New York City, NY: Wiley.

Hemelrijk, C. K., Wantia, J., & Isler, K. (2008). Female dominance over males in primates: Self-organisation and sexual dimorphism. *PLos One*, *3*(7), e2678.

Hogeweg, P. (1988). MIRROR beyond MIRROR, puddles of life. In C. G. Langton (Ed.), *Artificial Life* (pp. 297–316, Vol. 6). Redwood City, CA: Addison-Wesley.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, *57*(3), 567–582.

Humphrey, N. K. (1976). The social function of intellect. In P. Bateson & R. A. Hinde (Eds.), *Growing Points in Ethology* (pp. 303–317). Cambridge, England: Cambridge University Press.

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.

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204.

Klügl, F., & Bazzan, A. L. (2012). Agent-based modeling and simulation. *AI Magazine*, *33*(3), 29–40.

Kraus, S. (2001). *Strategic Negotiation in Multiagent Environments*. Cambridge, MA: MIT Press.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1503.

Kummer, H., Anzenberger, G., & Hemelrijk, C. K. (1996). Hiding and perspective taking in long-tailed macaques (Macaca fascicularis). *Journal of Comparative Psychology*, *110*(1), 97–102.

Lenaerts, T., Saponara, M., Pacheco, J. M., & Santos, F. C. (2024). Evolution of a Theory of Mind. *iScience*, *27*, 108862.

Lewicki, R. J., Barry, B., & Saunders, D. M. (2020). *Negotiation*. New York City, NY: McGraw-Hill.

Liddle, B., & Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, *4*(3-4), 231–244.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica: Journal of the Econometric Society*, *60*(4), 803–836.

Meijering, B., Taatgen, N. A., van Rijn, H., & Verbrugge, R. (2014). Modeling inference of mental states: As simple as possible, as complex as necessary. *Interaction Studies*, *15*(3), 455–477.

Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 639–648.

Niazi, M., & Hussain, A. (2011). Agent-based computing from multi-agent systems to agent-based models: A visual survey. *Scientometrics*, *89*(2), 479–499.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, MA: Harvard University Press.

Osborne, M. J. (1990). *Bargaining and Markets*. San Diego, CA: The Academic Press.

Patil, R., Fikes, R., Patel-Schneider, P., McKay, D. P., Finin, T., Gruber, T., & Neches, R. (1992, August). The DARPA Knowledge Sharing Effort: Progress Report. In B. Nebel (Ed.), *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning* (pp. 103–114). Burlington, MA: Morgan Kaufmann.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 731–744.

Pojeta, J., & Springer, D. A. (2001). *Evolution and the Fossil Record*. Alexandria, VA: American Geological Institute; The Paleontological Society.

Povinelli, D. J., Eddy, T. J., Hobson, R. P., & Tomasello, M. (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, *61*(3), 1–189.

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Sciences*, *7*(4), 157–160.

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, *19*(1), 1–28.

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

Puga-Gonzalez, I., Hildenbrandt, H., & Hemelrijk, C. K. (2009). Emergent patterns of social affiliation in primates, a model. *Plos Computational Biology*, *5*(12), e1000630.

Ray, J. (1691). *The Wisdom of God Manifested in the Works of the Creation*. London, England: Smith.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, *50*(1), 97–109.

Schelling, T. C. (1969). Models of segregation. *The American Economic Review*, *59*(2), 488–493.

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, *23*(8), 842–847.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, England: Cambridge University Press.

Tager-Flusberg, H., & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of Autism and Developmental Disorders*, *24*(5), 577–586.

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Sciences*, *7*(4), 153–156.

Van der Vaart, E., & Hemelrijk, C. K. (2014). 'Theory of mind' in animals: Ways to make progress. *Synthese*, *191*, 335–354.

Van der Vaart, E., Verbrugge, R., & Hemelrijk, C. K. (2012). Corvid re-caching without 'theory of mind': A model. *Plos One*, *7*(3), e32904.

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

Verbrugge, R., Meijering, B., Wierda, S., Van Rijn, H., & Taatgen, N. (2018). Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and Decision Making*, *13*(1), 79–98.

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

Wallace, A. R. (1858). On the tendency of varieties to depart indefinitely from the original type. *Journal of the Proceedings of the Linnean Society of London, Zoology*, *3*, 53–62.

Wang, I., & Ruiz, J. (2021). Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction*, *37*(17), 1648–1673.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

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. (2009). *An Introduction to Multiagent Systems*. Chichester, England: John Wiley & Sons.

Appendix A

Experiment 1

Figure A.1 shows the distribution of the agent ages for all the agents of experiment 1 where the learning speed was 0.2. The ages are categorized based on the order of Theory of Mind of the agent.

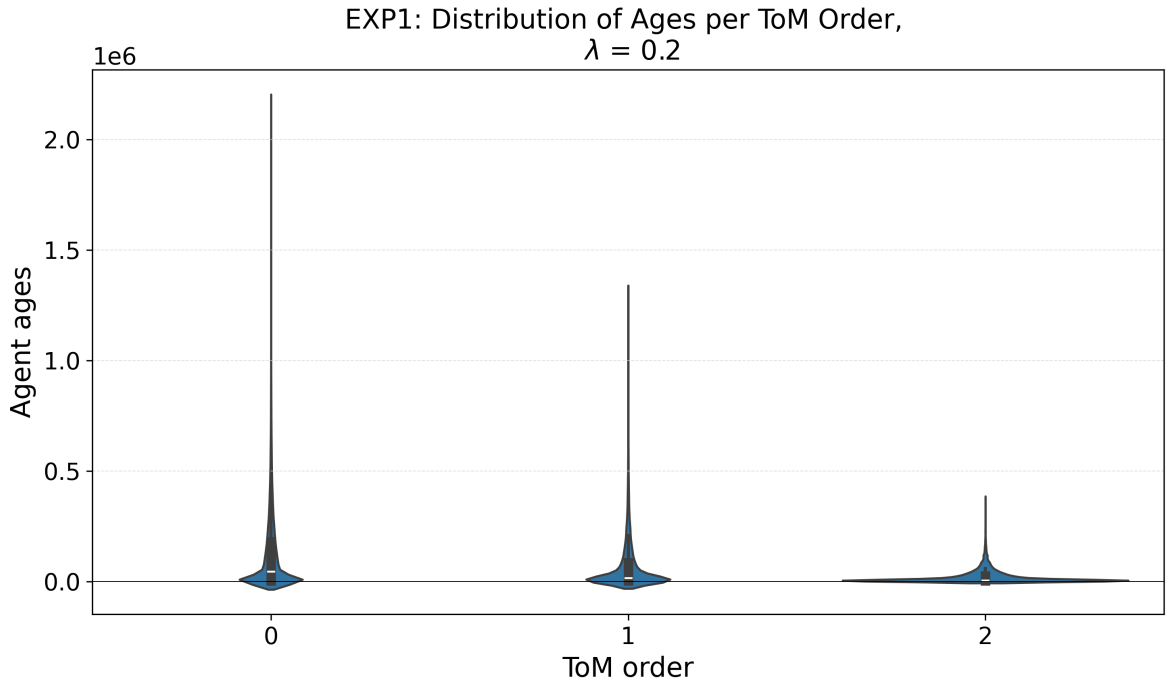

Figure A.1: A violin plot of the spread of agent ages of experiment 1 per ToM order, for the experiments with a learning speed of 0.2.

Figure A.2 shows the distribution of the agent ages for all the agents of experiment 1 where the learning speed was 0.8. The ages are categorized based on the order of Theory of Mind of the agent.

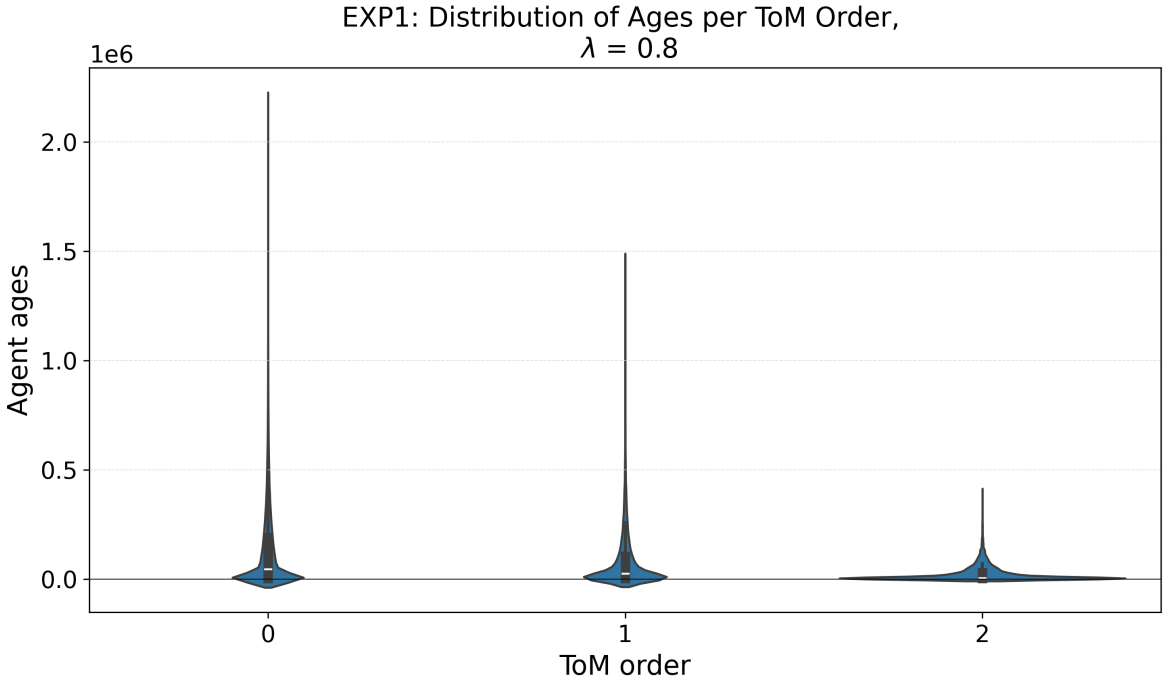

Figure A.2: A violin plot of the spread of agent ages of experiment 1 per ToM order, for the experiments with a learning speed of 0.8.

Figure A.3 shows the frequencies of the reasons for a negotiation termination per negotiation type for experiment 1 where the learning speed was 0.8.

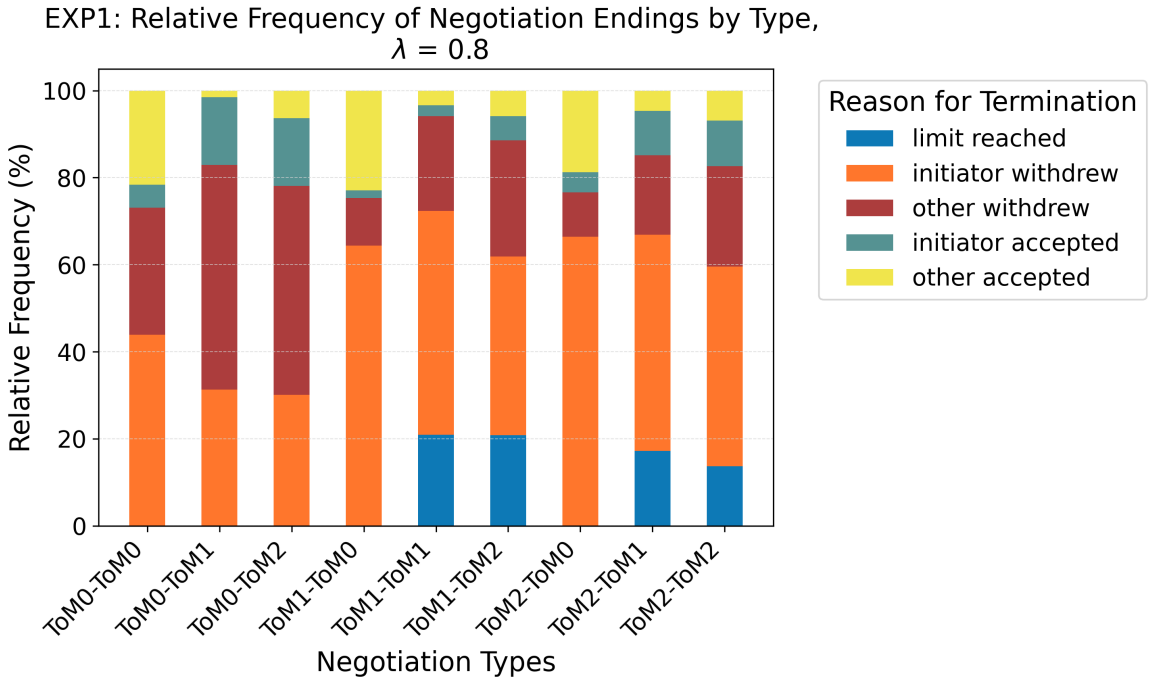

Figure A.3: The reason for negotiation termination for experiment 1 per negotiation type, for the experiments with a learning speed of 0.8.

Appendix B

Experiment 2

Figure B.1 shows the distribution of the agent ages for all the agents of experiment 2 where the learning speed was 0.2. The ages are categorized based on the order of Theory of Mind of the agent.

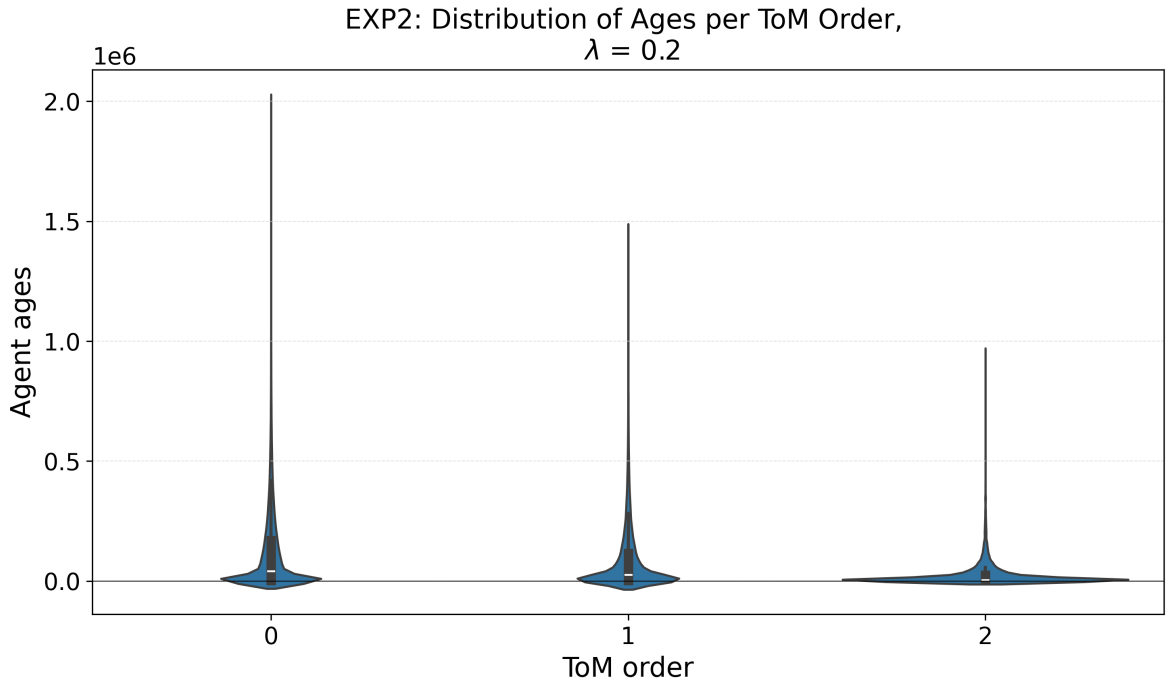

Figure B.1: A violin plot of the spread of agent ages of experiment 2 per ToM order, for the experiments with a learning speed of 0.2.

Figure B.2 shows the distribution of the agent ages for all the agents of experiment 2 where the learning speed was 0.8. The ages are categorized based on the order of Theory of Mind of the agent.

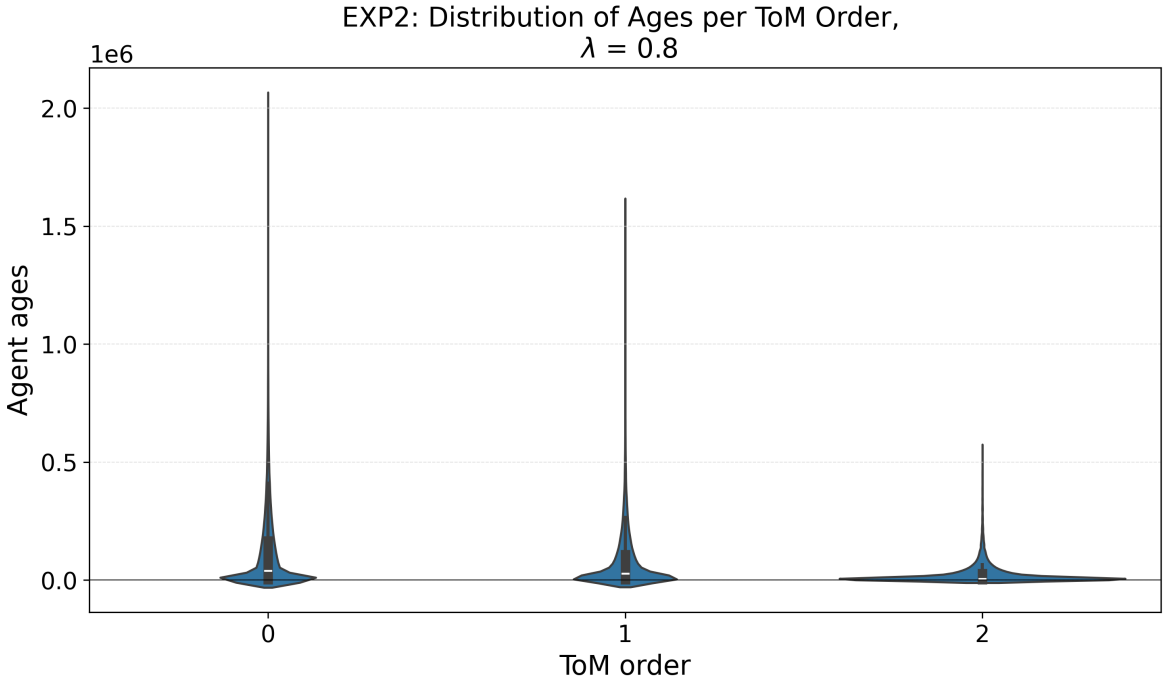

Figure B.2: A violin plot of the spread of agent ages of experiment 2 per ToM order, for the experiments with a learning speed of 0.8.

Figure B.3 shows the frequencies of the reasons for a negotiation termination per negotiation type for experiment 2 where the learning speed was 0.8.

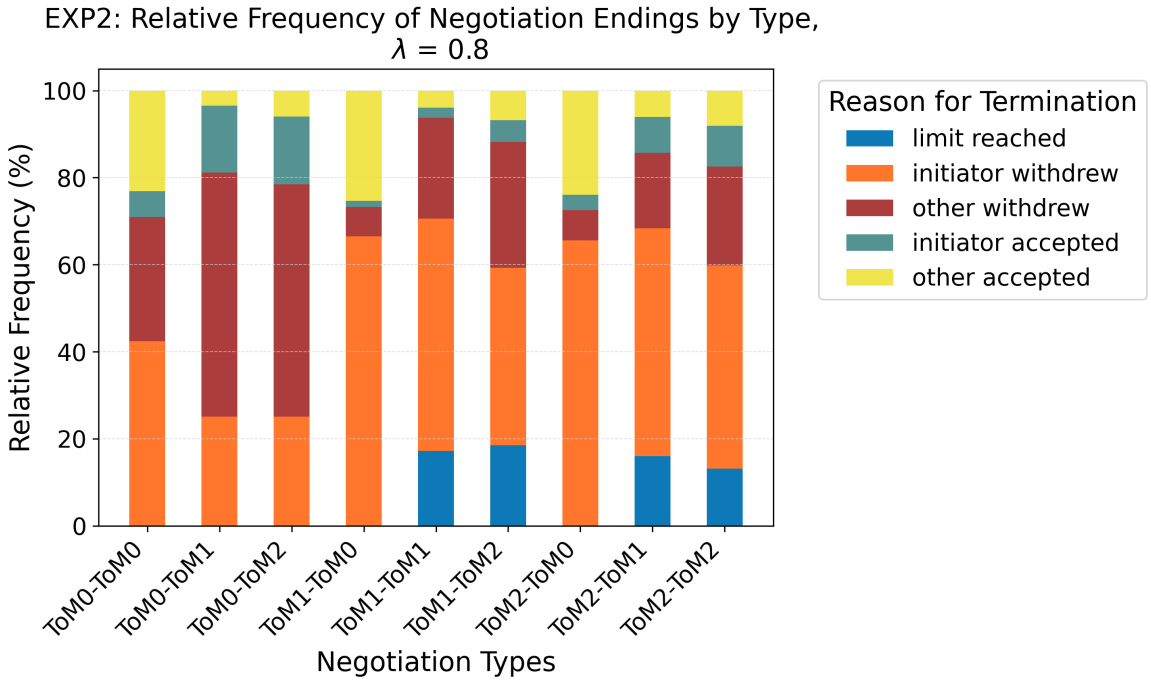

Figure B.3: The reason for negotiation termination for experiment 2 per negotiation type, for the experiments with a learning speed of 0.8.