Comparing Multiple Models of Reasoning: An Agent-based Approach

A. van der Meulen (s1901974)

August 31, 2016

Abstract

Being able to negotiate under pressure is generally considered to be a useful skill that one can apply to maximize one's gains in a negotiation situation. In this study, we have looked at negotiation situations using a game called Coloured Trails. Coloured Trails is a decision-making board game in which human or agent players are given a finite set of chips they need to hand in to cross corresponding fields on the game board. The players have to reach or get close to a goal position on this board. They are usually unable to reach this goal position with their initial set of chips, and needs to trade with a purely responsive agent to better their position. We consider a version of Coloured Trails in which a human competes with an agent for the opportunity to trade with a responding agent in order to obtain chips that will bring them closer to their goal.

We have looked at the effects of different models on the learning behaviour of both participants and agents. To achieve this, we have implemented two models: a model that uses a belief-adjustment theory of mind system and a model that uses parameter-based input. We have implemented two variations of both of these models with Java agents. After this, participants were asked to play 10 one-shot games of Coloured Trails against each of the four agents, after which the participant performance was evaluated.

In the end, we found that while it is hard to highlight differences in the participants' performance across the models, participants in the study showed that they were clearly improving their position and speed throughout the experiment. This shows that Coloured Trails may be an efficient game in teaching players to balance their own goals versus those of potential competitors and negotiation partners. This in turn teaches people to negotiate more efficiently. In the future, it may be beneficial to increase the contrast between games played against different agents, in order to accurately highlight the differences of the respective models in terms of learning.

Contents

1 Introduction

There are many models that try to emulate or simulate the human strategic decision-making process in social situations. In this graduation project, we will explore multiple agent-based models seeking to emulate this decision-making capability. Each of these models has tried to find a method to explain the human reasoning capabilities through simple, logical means, but these methods are theoretically very different and may all emulate just a part of human decision-making. It is unclear which model can be seen as an effective representation of the real world, especially when one has to deal with situations in which one does not know what other persons or agents think. However, most of the models concerning a lack of knowledge about others seem to have one thing in common: they assume human decision-making in social situations is based on theory of mind. This so-called theory of mind is a concept that can be described as the ability that one uses to reason about the thoughts of others. The skill to use theory of mind is commonly attributed to healthy humans above the age of about five years old (Frith & Frith, 2005).

While this age of five is a fair indicator of the moment at which humans develop theory of mind, theory of mind is usually considered to be a spectrum. Studies by Wimmer and Perner (1983) revealed that children under the age of four do not show theory of mind, whereas 57% of the five and 86% of the six year old children possessed theory of mind when posed with questions after hearing a story. Wimmer and Perner showed this by providing the children with a scenario in which a person *x* put down an object in a certain location, after which another person *y* moved the object to a new location, unbeknownst to person *x*. Children without theory of mind reasoned that person *x* would look at the new location, showing that they were not able to correctly reason about the thoughts person *x* would have concerning the object.

1.1 Theory of Mind

The concept of theory of mind is divided into multiple segments, each containing its own reasoning pattern. The most commonly used segments are zeroth level, first level and second level. If one has a zeroth level theory of mind, one will only reason about observable events in the physical world. On the higher levels, however, things get more interesting: theory of mind level one theories assume that one is able to reason about what others might think of these observable events (like the children in our example from the Wimmer and Perner study), and theory of mind level two theories assume that one is able to reason about what others might think oneself thinks about the observable events (the study by Wimmer and Perner found reasoning patterns such as these to be likely at the age of > 6, but not so likely at the age of 5).

Theory of mind is best explained by use of an example: Suppose that you are at the store, and that the clerk announces that there is only one box of breakfast cereal left. You quickly realize that you need a box of cereal. If you lacked theory of mind, you would now assume that there is not really a problem: after all, you only need one box of cereal. There's one box left, so this is enough. However, if you used theory of mind, you would realize that there may be other persons in the store that also want this box of cereal, e.g., you realize other persons also have thoughts and desires, and it would be wise to adjust your route to get the box of cereal. If you had an even higher level of theory of mind, you would realize that you may wish to hurry to get your box of cereal, and head to the cereal directly, as you realize the other persons in the store may realize that there are others that want that box of cereal, and will also adjust their route to get to the cereal, meaning you will have to hurry if you wish to get that last box.

The importance of studying theory of mind lies in understanding human cognition: if we understand why and how certain decisions are made, and on which reasoning patterns these decisions are based, it becomes easier to predict human behaviour. This prediction of human behaviour can be applied to many scenarios: strategic decision-making, studying social phenomena such as pretend play (Dore, Smith, & Lillard, 2015) and bullying (Sutton, Smith, & Swettenham, 1999) amongst children, understanding the limitations of autism (Baron-Cohen, Leslie, & Frith, 1985) and, less specifically, looking at general reasoning skills applied in everyday life.

1.2 Strategic Reasoning in Dynamic Games

Out of the scenarios that use theory of mind we have mentioned, the field that is studied the most is the use of theory of mind in strategic decision-making. This decision-making is usually studied in the context of games, as games often provide an accurate abstraction to measure the development and growth of strategic decision-making capabilities amongst humans. An example of this can be found in a study that showed how one can see smallscale improvements in the strategic decision-making capabilities of elementary level students (Bottino, Ferlino, Ott, & Tavella, 2007). This improvement occurred when they were exposed to computer games that simulate strategic board games. The study identified the game properties that are important to stimulate teaching strategic decision-making capabilities to children, which included giving them the ability to backtrack, giving thorough, direct, feedback and a gradual increase in difficulty. Studies like these can help us understand how people learn best, by directly tying the learning process to the simulation of a multitude of board games.

The study by Bottino et al. is far from the only study that has looked into the effects of board games on learning to sharpen one's reasoning skills. A different study used the game of Mastermind to investigate the effects of theory of mind on logical communication (Verbrugge & Mol, 2008). This was one of the earlier studies in a sequence of studies that studied theory of mind using various board games. This particular research found that some participants are potentially able to make distinction between pragmatic and logical communication, and that in terms of strategic decisions a lot of people tend to resort to theory of mind level one assumptions, with only a small number of players showing the ability to reason with theory of mind level two assumptions before practice. In other words: the Mastermind research showed that in a strategic setting, a lot of humans tend to default to assuming that the opponent does not actively model their behaviour.

1.3 Agent-based Simulations with Higher Cognition

One of the ways to study the strategic decision-making capabilities of humans when it comes to theory of mind, is to have them compete against an agent-based simulation. However, before we can let humans compete against agent models in a simulation, we need to build the agents. So-called agent-based simulations with a higher level of cognition have, for example, been developed to study strategic decision-making in the games of Rock, Paper, Scissors and its various more complicated variations (De Weerd, Verbrugge, & Verheij, 2013). In this research by De Weerd et al., agent models show how higher levels of reasoning agents can outperform their lower level opponents. The research looked into theory of mind levels up to and including level four, showing that the process becomes fairly complicated as the models assume higher levels of cognition (reasoning about your opponent reasoning about you reasoning about your opponent reasoning about your goals and beliefs is also a process that is hard to grasp in practice). They ultimately concluded that any pay-off beyond theory of mind level two may not be worth the effort (the extra pay-off is minimal, while higher level reasoning demands a lot more resources).

The main advantage of using agent simulations is that people can reason in a closed environment, enabling us to study any theorized effects within a specific setting against an agent. This does, however, mean that the agent model needs to be realistic, or needs to have an added benefit when compared to studying social situations where theory of mind is useful with human participants only. As such, one of the things we will look into during this research is how useful agent models can be in learning to make strategic decisions.

1.4 Mixed-motive Situations

It has been shown that theory of mind is advantageous in a multitude of settings. This is not only argued for both competitive games (Byrne & Whiten, 1989) and cooperative games (Vygotsky, 1980), but also shown in practice for combinations of the two (Verbrugge, 2009; De Weerd, Verbrugge, & Verheij, 2015a). When we only consider competitive games, we end up with zero sum games, as the player and its opponent directly oppose each other for a certain number of points (an example of this is the aforementioned well-known Rock, Paper, Scissors game). In this type of games, it is often relatively easy to find a counter strategy to the opponent strategy, and in agent-based simulations that seek to emulate theory of mind, the model only has to reason about picking the choice best suited to their opponent. This is why combinations of competitive and cooperative games are generally more interesting: they are not a zero-sum game. We are dealing with a mixed-motive situation in this case: while there may be an optimal solution for a game, the optimal solution is not the best possible outcome for either of the players. Cooperation could improve their situation, but competition could potentially improve their overall situation even more. Human and agent players will have to actively weigh the benefits of their potential choices against each other.

A well-known example of this is the Prisoner's Dilemma, a game theory concept that has been previously used to study theory of mind (Press & Dyson, 2012). In the prisoner's dilemma two players are faced with a prison sentence, that they can get rid of by ratting out the other player (defecting). The other player would get an extension of their prison sentence in the event that this were to happen. However, if both players choose to rat each other out, they will both be put into prison. If neither of the players chooses to rat out the other, both players will be put in prison for a shorter time than they have been otherwise, but will not go free. Ideally, neither of the players would rat their fellow player out, as this would lead to the shortest sentence overall. However, a player can improve their score by preying on their fellow player, and hoping that they are the only one to report their fellow player to the authorities, thus, walking free.

The dilemma in this particular mixed-motive situation assumes that players are faced with this choice only once in their lives, which according to game theory leads to the Nash equilibrium in which both players choose to rat out their fellow player. After all, if neither player defects, it is always beneficial for their opponent player to defect anyway, reducing their sentence. Both players will defect as that will be better than not defecting and getting a higher sentence due to their opponent's choices. However, if both players had chosen to cooperate, their overall sentence would be lower. This cooperation seems to be a far more likely strategy when posed with players that play a repeated set of games involving the Prisoner's Dilemma (Andreoni & Miller, 1993), the so-called Iterated Prisoner's Dilemma. The Nash equilibrium in the Prisoner's Dilemma seems to suggest that when posed with both a competitive and a cooperative option in a game, players tend to prefer a competitive choice, but when posed with severely negative consequences for this choice (such as the loss of reputation in the Iterated Prisoner's Dilemma), prefer to cooperate instead, which is only possible with a higher level of theory of mind as it requires one to reason about the thoughts of the opponent: a higher level of theory of mind helps one with performing in a competition versus cooperation situation. Another example of a mixed-motive games is the game of Coloured Trails (Section 1.5). This is the mixed-motive game that we used in our study, and the game is an example of a mixed-motive situation where the refusal to cooperate will lead to a player strongly reducing their chances to improve their score.

1.5 Coloured Trails

Coloured Trails is a game in which a number of players try to reach a certain goal position on a game board, from a starting position. The aim of the game is to get as close to the goal position as possible, to gain the highest possible score. This is the case for every player of the game. The goal position can be reached by moving over the game fields. However, there is one catch: in order to move over a field, one has to hand in a chip of a corresponding field colour. The player only has a limited number of chips, limiting the player's ability to reach the goal position.

When the colours of the player's chips do not correspond to the field colours on the game board, the player will require an additional number of chips, or different chips entirely, if the player wishes to reach the goal. A player can solve this problem by negotiating with certain other players: an exchange of chips between players may be beneficial for both of them. An example of a game of Coloured Trails between two players can be found in Figure 1. Therefore, Coloured Trails is a game with both a competitive and a cooperative component: players will try to beat one another by gaining the highest possible score, while at the same time trading with one another, in a way that benefits both, to obtain the chips they need to get as close to the goal position as possible.

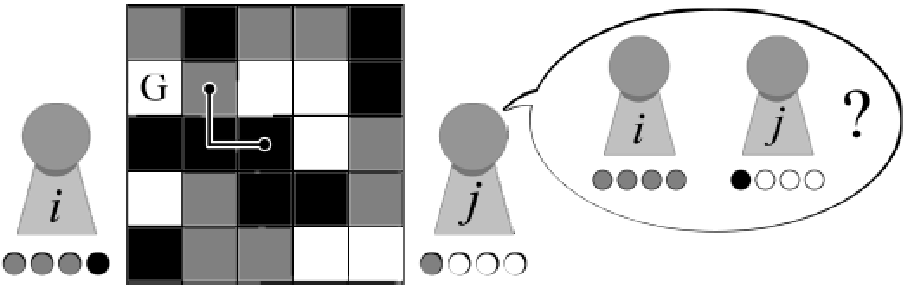


Figure 1: An example of negotiation in Coloured Trails. Agent j is trying to trade a chip with agent i in order to get close to its goal location, G (De Weerd et al., 2015b)

The Coloured Trails game has two types of players, in our simulation referred to as agents. The first type of agents can vary in number, but there will only be one agent of the second type. The two types of agents in the Coloured Trails game are:

1. A number of proposers: Proposers are agents that actively affect the game. They will try and trade chips with a responder agent, the second type of agent, and will try to gain a score that is as high as possible.
2. A responder: The one responder agent in the game is a relatively passive agent. The responder will receive offers from all of the proposer agents, and will decide whether they have proposed a trade that is worth considering. Eventually, the responder either accepts the offer

from *one* specific agent, or decides to reject all offers and keep the chips it currently has.

The proposer agents can use different chip exchange strategies to reach their specific goal position, and can use information about the chips of the responder to reason about the offers they propose. Each proposer can only make one offer to the responder, and as such should find the most optimal offer if it wishes to gain anything from the exchange. Both the proposers and the responder are aware of the whole game board: they can reason beforehand about what chip would be needed two actions later, as the available information about the game board is complete. The proposers and the responder are also aware of one another's goal state and starting point. However, proposers do not know what chips the other proposers possess, which means that the game is a partial information game for the proposers. This causes proposers to require a certain strategy while playing the game, rather than simply calculating whether they can gain a higher score than the other proposers with the chips they possess.

The responder receives all the proposers' offers at the same time, as proposers simultaneously hand in their chip exchange offer to the responder. Due to the fact that the responder will only accept one offer (or reject all of them), the proposers also have to take into account what the offers of their opponent proposers are, because a better opponent offer can beat their own offer, which would result in zero additional points for the proposer itself, as their situation does not change and no additional moves on the board can be made when compared to their situation before the offer.

1.6 Different Models of Coloured Trails

There are multiple strategies that can be used by the proposers, which have been explored by researchers in multiple experiments. Two examples of these experiments are an experiment by Ficici and Pfeffer and an experiment by De Weerd, Verbrugge and Verheij. Both of these experiments are based on the premise that efficient play of Coloured Trails can be achieved by reasoning about the intentions and beliefs of other agents, rather than making a seemingly random proposal.

The first experiment makes use of a weight-based model, which includes certain properties of the Coloured Trails game for the proposer agent to determine how to reason (Ficici & Pfeffer, 2008). These properties include the score change if the offer were to be accepted, and the number of offers the proposer can make given the set of chips available to the proposers and responder. The model of Ficici and Pfeffer has different levels of reasoning with this weight-based method. For example, the proposer model may consider the possibility that the other proposers would also use these weighted properties of Coloured Trails to reason about offers in the game, rather than considering itself to be the only entity with any reasoning capabilities. This can be chained to obtain certain levels of the aforementioned theory of mind: a 'first' proposer may come to the conclusion that another proposer/responder may realize the 'first' proposer realizes that they can also use the weighted properties, and so on. This model is also called the level-n model, referring to the *n* levels of model chaining.

The second experiment makes use of a belief integration model, which includes a factor of belief in another player's use of a certain level of theory of mind (De Weerd, Verbrugge, & Verheij, 2014). These models by De Weerd try to find the proposers' level of theory of mind by looking at the trades they have performed with the responder versus their opponent in the previous games of Coloured Trails. The points a proposer can gain from the current game are then integrated with their belief in how likely it is that their opponents are using a certain level of theory of mind. This integration is to a certain extent an abstraction of the proposer and responder performance when using different levels of theory of mind.

Both experiments have shown that the method that they have implemented increases the proposer score as the proposers show a deeper level of understanding for the other agents. In other words: a deeper level of understanding of theory of mind increases a proposer's performance. In this thesis, we will analyze what the performance differences between these two methods are, highlighting how the differences in the models affect the final score.

1.6.1 Ficici and Pfeffer's Research

The studies performed by Ficici and Pfeffer were specifically designed to find agents that were capable of higher level reasoning. These agents were then used to evaluate whether other players' reasoning capabilities under uncertainty, and how these other players reason within the mixed-motive situation that Coloured Trails provides, mainly focusing on whether people reason about other players while also trying to satisfy a responder. Ficici and Pfeffer first ran an experiment that pitted humans against humans, in order to model how humans would respond against one another in a number of circumstances in the Coloured Trails game. In this experiment, they obtained a total of 268 games of two human proposers trying to bargain with a human responder (over 69 participants in total). Next to this experiment, they also collected data by having 221 games in which a human responder decided between two hand-crafted offers. Using the responses from these two data collection experiments, they taught a model how to reason according to a level-n human mind by use of the offers, their agent implementation and gradient descent.

These so-called level-n agents were then evaluated by having them play Coloured Trails with the same technical game setup (but with different situations) against 59 unique human proposers. The responder model they used was an optimal response model: it simply accepted the situation that would give it the most benefit. During these evaluation trials, Ficici and Pfeffer found that generally speaking, the more the model fits the human data, the better the responses become (in other words: human reasoning outperforms the default agent reasoning), while at the same time, the higher level models performed better than the lower level models (level-$(n+1) >$ level-n). From this, they concluded that their agents were sufficiently capable of emulating human reasoning with their parameter-based models, in addition to concluding that humans reason not only about satisfying their trade partners, but also reason about ways to deal with any potential opposing offers to their trade partners.

1.6.2 De Weerd's Research

The study performed by De Weerd was an exploratory research project looking into modeling agents with specific theory of mind reasoning capabilities, looking into emerging higher level theory of mind from reasoning patterns, rather than to find the data by use of previously inputted human data. In their research, De Weerd et al. simulated multiple one-shot games of Coloured Trails between agents of different levels of theory of mind. Two proposer agents had to try and convince a responder agent to trade with them, in order to reach their goal, similar to the Ficici and Pfeffer studies. De Weerd's models were tested for theory of mind level zero up to and including theory of mind level four. They tested two different types of agents: agents with a best-response strategy and agents with a utility-proportional belief strategy.

The models by De Weerd found that zeroth level theory of mind agents are outperformed by first level theory of mind agents, which are outperformed by second level theory of mind agents, and so on. However, the study also showed that reasoning using a basis beyond the second level of theory of mind has diminishing returns: while the effort and time required to make the calculations serving this higher cognitive function increase quite significantly, the actual performance does not increase that much. In other words, levels of theory of mind beyond level two are often counterproductive if we factor in time and effort spent on reasoning. The models by De Weerd did show that it was possible to emulate theory of mind reasoning without the need of specific human data. However, as they were never tested on human data within the specific Coloured Trails (mixed-motive) situation that was offered, it is unknown how the models would fare against humans, and if they would show similar results when pitted against humans.

1.7 Research Question

In our research, we will look at the differences between the theory of mind-based De Weerd model and the parameter-based Ficici and Pfeffer model, to find whether the use of theory of mind in agents can help humans when learning to deal with a strategic decision over an agent fitting certain parameters to optimize a decision. In short, our research question is 'Can the use of theory of mind in agent-based models help with the improvement of strategic decision-making capabilities in humans over parameter-based models?'. In order to answer this research question, we will pit participants against two variations of both models: a directly applied theory of mind level one model, a directly applied theory of mind level two model, and two models inspired by the level-1 and the level-2 Ficici and Pfeffer model respectively. They will play a game of Coloured Trails, in a variation inspired by a version previously used by Ficici and Pfeffer (2008).

1.8 Hypothesis

We hypothesize that the theory of mind model will help humans in making strategic decisions over the parameter-based models, as the theory of mind models are more directly in tune with and have been directly modeled based on the strategic reasoning capabilities that humans show through use of theory of mind. As the task is rather complicated, we also expect to find that participants do not start off completely mastering the task, and as such will improve while performing the task. This improvement is expected to occur both the speed of their answers and their score improvement when looking at the score they would have gotten when no chips change hands compared to the score they have gained through the trial.

2 Methods

In this section, we will explain the methods behind our experiment, the implementation of the models we have used and the ways in which we will analyze our experiment. We will first start by explaining the setup we have used for our simulation of Coloured Trails, as we cannot model the game without knowing its parameters first.

2.1 Game Setup

The Coloured Trails simulation that we have implemented mostly sticks to the settings used in Ficici and Pfeffer's parameter model. This means that the game setup adheres to the following rules:

1. Each agent is given 5 chips.
2. There are 5 unique tokens that correspond to the chips.
3. The board will consist of 16 tiles, in a layout of 4 by 4.
4. Agents start in a corner, meaning that certain chips are always required in order to get onto the gameboard at all. This gives certain chips more value than others. The goal will always be opposite of this corner.
5. Proposers start in a different corner than the responder, to increase the mutual benefit opportunities that arise from trading.

This setup of Coloured Trails makes use of two proposers and a responder. Some of the previously discussed versions of Coloured Trails have two agents that try to outreason one another to get to their goal, without the interference of a responder. As such, the contact that two proposers have with one another is only indirect. This makes it harder to reason about each others beliefs, as it is much easier to consider the entity with whom you are trading, than the entity who may influence the direct trade you are trying to make with your negotiation partner.

Next to the fact that there are two proposers and one responder, the game setup both the De Weerd and the Ficici and Pfeffer studies adhered to was that the humans and/or agents had to play one-shot games, which meant that there is no room for error or exploration, as the offer will be final as soon as the offer is made. It is impossible for the relatively passive responder to clarify its intentions, which also means that one cannot deduce whether the opposing offer may be better than theirs. This provides a layer of uncertainty.

Proposers do not know each other's chips. As such, the game is not only about estimating the desires of one's opponent and passive trading partner, but also about estimating the assets one's opponent possesses. This complicates the game further, but also, in the case of success, shows that the models can reason even when faced with a major degree of uncertainty.

In order to reduce the complexity of a random corner, proposers always start in the same corner, as do responders. This choice was made to reduce potential confusion for human players, as this is a problem that implemented agents will not encounter (whereas aspects such as reasoning under uncertainty are a problem for both humans and agents).

The score that an agent can obtain is calculated by giving an agent a 100 points if they can reach the goal. Additionally, they get 25 points for each step they come closer to the goal, starting with 0 points. In our 4x4 board, this implies that agents can get a maximum of 200 points from reaching their goal alone. If the agent retains chips after approaching its goal, the agent is also given points: the agent receives 10 points for each chip that has not been used.

This setup has a few consequences:

1. The minimum score an agent can obtain is 0 points. This is because the agent is not on the field when it starts, thus gets no points before handing in the first chip to enter the field. This situation occurs when an agent chooses to settle for 0 chips.
2. All potentially useful, goal-reaching, paths that a player can walk are between four and seven steps long.
3. The maximum score an agent can obtain is 260. This situation occurs when an agent chooses to settle for all the chips, and can actually reach its goal within a minimum number of moves (four).

2.2 Simulating Coloured Trails

The simulation of Coloured Trails that we have implemented makes use of the Java Programming Language (Arnold, Gosling, & Holmes, 1996). This simulation has been divided in four separate modules:

1. A game play simulation module, which runs the Coloured Trails game and keeps track of things such as the number of games that have been played.
2. A game board simulation, which generates and shows the game board that is used to play the game on.
3. A player behaviour simulation, which contains a model of the proposers in the game.
4. A responder behaviour simulation, which contains a model of the responders in the Coloured Trails game.

These modules have been divided into submodules, that implement for example different types of proposer and responder modules.

2.2.1 Game Play Simulation

The game is simulated by first initialising two proposers and a responder, along with a randomly generated game board. After this initialisation, both proposers will play the game for themselves and will see how the game would turn out if they possessed both their own chips and the chips of the responder agent.

In each game play simulation, all agents (all proposers and the one responder) are handed five randomly selected chips. These chips each have one of five different colours, with each colour corresponding to the colour of a potential field on the game board. It is possible that, during this selection process, an agent receives multiple chips of the same colour. An example distribution can be found in Table 1. This table shows the five chips that we use in our game play simulation: a plus chip, a stripe chip, a wave chip, a diamond chip and a lattice ship. Other properties that are initialised are, for example, the ID number of the agent, and which type of response module the agent uses to outreason its opponent(s).

After the game has been played, the score for the game will be evaluated by using the aforementioned criteria for both agents (Section 2.1), after which the proposer agent that has managed to gain the highest number of points is announced to both players and the responder. This process is repeated for *n* rounds, each containing randomised chip distributions, and randomised board layouts. The strategy that either agent applies is kept constant.

Table 1: An example set of chips.

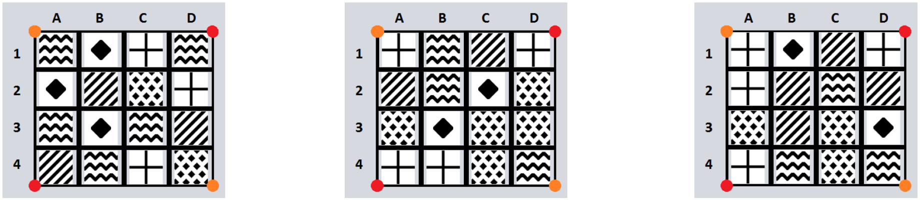
Agent	Plus ⊞	Stripe ▨	Wave ≋	Diamond ◆	Lattice ▦
Proposer #1	2	1	0	1	1
Proposer #2	1	1	1	1	1
Responder	1	2	1	1	0

2.2.2 Game Board Simulation

The game board module is initialised in the Game Play module, but is a separate entity that is also used in order to provide an easily accessible visual representation of the game board.

The game board is initialised by randomly picking *4x4* fields that represent the game board. Examples of these game board can be found in Figure 2. The random values used in these fields correspond to the five potential colours that a proposer has been given with their chips. The starting point of the proposers is in the top left corner, and the goal is at the opposite side of the field, in the bottom right corner (in future figures, represented with orange dots). The starting point for the responder agent is in the top right corner, with the responder goal sitting in the bottom left corner (in future figures, represented with red dots). These field values are not explicitly implemented in the board itself, and are instead represented in the starting point and the goal state of the proposers and the responder.

The differences between the proposers and the responder in terms of the start and goal position should promote chip exchange situations, as the agents potentially require different chips to reach their goal. At the same time, the competitive element between the proposers remains, as they require some of the same chips to reach their goal from their start. The starting points and the goals of all agents are known to one another, which means that all the information about the board is known to all the agents.



(a) First Setup

(b) Second Setup

(c) Third Setup

Figure 2: Example Board Setups for Coloured Trails. The orange dots represent the starting points for the two proposers; the red dots represent the starting point for the responder.

2.2.3 Proposer Behaviour Simulation

The proposer behaviour simulation consists of different types of proposers, that have been implemented as substructures of the main proposer structure. The main proposer structure contains general proposer functions, such as calculating the overall score that the proposer can obtain with the chips it possesses, and the function that moves a proposer over the board with the chips the proposer has.

The proposer firstly calculates the score that it would currently obtain from the chips it possesses. The way in which this is done, is by first checking whether the proposer can play at all. Since the player starts in the top left corner, this means that one of the five chips the proposer has been given at the start needs to have a colour that corresponds to the top left field of the game board. If this is the case, then the proposer will start walking over the board to see where it can end up with the chips the proposer has been given.

Walking over the board is performed by recursively checking the fields the proposer can move toward from the current position field, and checking whether the player actually has a chip that will allow the proposer to move to this field. Moving over the board can be done in a horizontal, vertical and diagonal manner. If the agent can move to the subsequent field, the recursive process is repeated until either of three conditions has been reached: the proposer has reached its goal, the proposer has expended all its chips, or the proposer cannot perform any additional moves. The final score for all of the finals positions is calculated, and the highest eventual score determines the path that the agent will choose.

This approach implies that the algorithm is path-based, rather than set-based. We have made this choice since walking over potential paths actively shuts down possibilities, greatly reducing the time it requires for an agent to decide which (most beneficial) paths yield which scores. If we were to evaluate all the potential sets, we would first have to define all the sets an agent can use in the game given the 10 chips it can access, leading to 10!, 3628800, evaluations for each possible set of chips. When we assume we do not evaluate chips that have a similar token, but are different physical objects (e.g.: there are 2 diamond chips in game, which are interchangeable in terms of the set of 10 chips), the least ideal situation would still leave us with 10 chips that have 5 different tokens, which translates to 10!/(10/5), 1814400, evaluations. Even when considering the fact that some chips are double, this huge number becomes a problem when an agent needs to evaluate more than 100 possible opponent chip sets.

If the proposer is already able to reach the goal without exchanging any chips with the responder, the proposer is handed new chips, to force a potentially beneficial negotiation situation. This redistribution situation is an unlikely scenario due to the fact that it always requires at least four chips to reach the goal state, and that the chance of an individual field matching a given chip is 1/5, a chance that will decrease as more chips are used to perform steps on the game board.

After evaluating the proposer's own score, the proposer will evaluate the score it could reach if it were possible to access the chips of the responder agent, with whom it will have to bargain or trade chips with. The proposer agent will then choose a strategy to obtain an score with the available chips (the number of chips can range from 0 to 10 chips). These strategies differ corresponding with the kind of methodology that has been used: different models for a utility-based approach, which uses weights to determine the utility of a move, and different models for a theory of mind-based approach, which uses explicit reasoning about the utility of oneself and others. These models are explained in more detail in Section 2.3.

2.2.4 Responder Behaviour Simulation

The responder behaviour module only deals with incoming offers. These offers are represented by the chips that the responder will have when it accepts the offer of a proposer, which in theory could be any offer ranging between zero and ten chips. In dealing with these offers, the responder has three options: it can reject the offer for both proposer *one* and proposer *two*, retaining the original setup, it can accept the offer of proposer *one*, and it can accept the offer of proposer *two*. The responder's score for the chips is calculated in a similar way to how the proposers do their calculations: the score that can be obtained with the chips that the responder originally had, is considered and compared to the scores that can be obtained with the offers the agents have made to the responder. The models for the evaluation of the player offers are explained in more detail in Section 2.3.

2.3 General Decisions in Coloured Trails

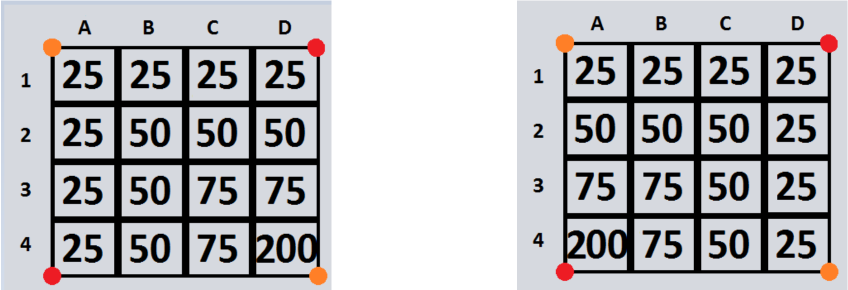
In this section, we will describe the general decisions one (and by extension our agent models) has to make in order to 'solve' the game of Coloured Trails. These decisions include score calculations, determining the paths one can walk and what offers an opponent could make against the responder.

2.3.1 General Score Calculations

The score one can gain by playing our variant of Coloured Trails is calculated by considering the distance to the goal in terms of tiles required to reach said goal. The agent is given 25 points for each tile that is closer to the goal, and will get another 100 points if this goal is reached. These additional 100 points serve as an incentive, to represent the benefit of actually reaching the intended goal and to encourage a competitive play style. For this same reason, any unused chips yield an additional 10 points: preventing as many chips as possible from being used, while still owning them, incentivizes competitiveness.

With the version of Coloured Trails that we play, *4x4* fields, this gives us a proposer score template as seen in Figure 3a, and a responder score template as seen in Figure 3b. In these templates, we have assumed the start and end positions used in Section 2.2.2: top left corner to bottom right corner for the proposers, and top right corner to bottom left corner for the responders. In addition to calculating the score gained from the tile on which the proposer, its opposing agents or the responder lands, the score calculation also takes the previously mentioned 10 points per chip remaining into account. We can capture these two score calculations in Formula 1 and in Formula 2:

$$S_p = 100 - 25 \max_{1 \leqslant x,y < bl} \{bl - y; bl - x\} + (10n), \text{ and } x < bl \vee y < bl \qquad \text{(1a)}$$



	A	B	C	D
1	25	25	25	25
2	25	50	50	50
3	25	50	75	75
4	25	50	75	200

(a) Proposer Score Grid

	A	B	C	D
1	25	25	25	25
2	50	50	50	25
3	75	75	50	25
4	200	75	50	25

(b) Responder Score Grid

Figure 3: Score Grid Overviews for Coloured Trails

$$S_p = 200 + 10n, \text{ and } x = bl \wedge y = bl \tag{1b}$$

$$S_r = 100 - 25 \max_{1 \leqslant x,y < bl} \{bl - y; x\} + (10n), \text{ and } x \geqslant 1 \vee y < bl \tag{2a}$$

$$S_r = 200 + (10n), \text{ and } x = 1 \wedge y = bl \tag{2b}$$

In Formula 1 and Formula 2, we account for the horizontal board position, represented by x, and the vertical board position, represented by y. The variable bl is used to indicate the length of the board. In our case, this would be 4. An overview of the x,y-grid as used in our variant of Coloured Trails can be found in Figure 4. The variable n represents the number of chips left after reaching the square at grid point [x,y]. For non-goal calculations, this results in the score calculations presented in Formula 1a and Formula 2a for the proposers (S_p) and the responder (S_r) respectively. When the proposer or responder reaches its goal tile, the score calculations default to the ones presented in Formula 1b or Formula 2b respectively.

Example 2.1. Calculating the Proposer Path Score

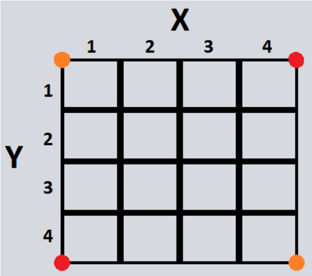
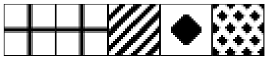
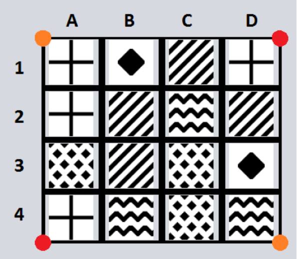


Figure 4: An overview of the grid coordinates as used in the score calculations of our variant of Coloured Trails

Prop. Chip Set





Suppose proposer 1 with its chip set (Table 1) plays on board 3 (Figure 2c). The proposer will iterate over the board to see what fields it can reach. In the event of board 3 we can see that:

- Proposer 1 starts in the top left corner and has to travel to the bottom right corner
- Proposer 1 has 2 plus chips, 1 stripe chip, 1 diamond chips and 1 lattice chip
- Proposer 1 has to use its 1 plus chip to enter the game board

As proposer 1 has a plus chip, it can move on the board. After performing this move, it will check which moves are possible from this field. The fields it can access from (A1; plus) are (B1; diamond), (B2; stripe) and (A2; plus). Proposer 1 is able to move diagonally downward right and downward. As such: proposer 1 can only move in two directions when reasoning from its current position (A1; plus).

- If proposer 1 were to go diagonally downward right (B2; stripe), it could go (if only looking at coming closer to the goal) diagonally downward right once more (C3; lattice), and diagonally downward left (A3; lattice), adding another two directions it can move in.

  - If the proposer was to go diagonally downward right (C3; lattice), the proposer could make two new moves: moving rightward to (D3; diamond) or moving downward to (C4; lattice). No new fields can be accessed after this. These moves would result in a net loss of points, as one additional chip is spent, while the tile distance to the goal remains the same.
  - The diagonally downward left move (A3; lattice) would result in a point decrease, as the tile distance between the proposer position and its goal increases. This is sometimes beneficial, as it opens up new opportunities for crossing the board, but in this case there is no benefit, as the proposer cannot move any close to the goal after moving to this square.

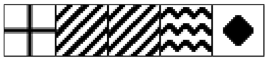
- If proposer 1 were to go downward (A2; plus), it could go rightward (B2; stripe), resulting in a field already accessed in a faster way in the previous scenario, diagonally downward right (B3; stripe) and it could go downward once more (A3; lattice). Moving to either (A3; lattice) or (B2; stripe) are not very fruitful, as in the case of the lattice token only a wave token would bring us closer to the goal (and would result in a redundant move), and in the case of field B2, we already have a shorter route with chips in the chip set. The diagonally downward right move to (B3; stripe) is worth considering however. The only fruitful moves from this field onward are moving to (C3; lattice) or (C4; lattice). However, C3 can already be reached in a shorter number of moves with the chip set available, and C4 will yield the same number of points as C3 with no room for score improvement. In short, most of the moves in the (A2; plus)-route would only result in a net point loss as the tile distance remains the same at the cost of expanding an additional chip.

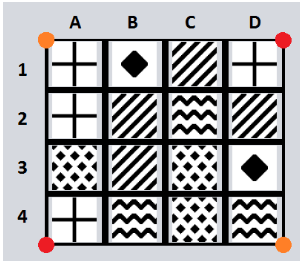
The overall best option for proposer 1 is as such to expend one plus chip, one stripe chip, and one lattice chip to reach field (C3; lattice). According to the rules established in Section 2.2.1, and the score calculation as given in Formula 1 this would result in a score of 95, seeing as the proposer is one square away from its goal and has two chips remaining.

The proposer will calculate the scores for the responder in a similar fashion.

Example 2.2. Calculating the Responder Path Score

Resp. Chip Set





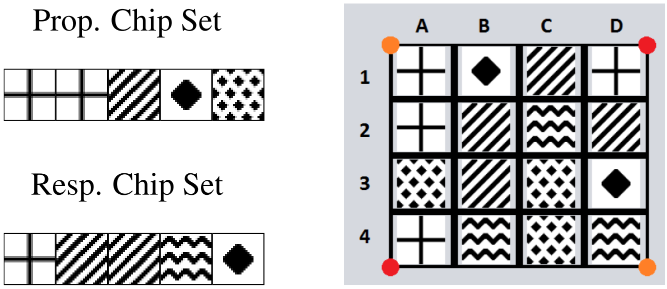
Suppose the responder with its chip set (Table 1) plays on board 3 (Figure 2c). The responder will iterate over the board to see what fields it can reach. In the event of board 3 we can see that:

- The responder starts in the top right corner and has to travel to the bottom left corner
- The responder has 1 plus chip, 2 stripe chips, 1 wave chip and 1 diamond chip
- The responder has to use a plus chip to enter the game board

With these three facts, the responder can also calculate the score it would gain. This calculation occurs in a similar fashion to the proposer score calculation. This will result in the best score path of top left corner start (D1; plus) - (C2; wave) - (B3; stripe). This will result in three chips used for one field away from the goal, a score of 95.

The proposer agent in our algorithm has remembered all the paths that it can take, but also takes into account what paths become viable when it possesses some, if not all, chips of the responder: the agent comes up with the paths that are possible when it possesses both the chips of itself and the responder. Based on these paths, the agent comes up with a proposal that divides the chips between itself and the responder. What this offer will be depends on the score utility of the offer, and the strategy the agent adheres to (described in Sections 2.4.1 and 2.4.2.

Example 2.3. Calculation the Score with an Added Offer



After proposer 1 has calculated its score, it will realize it is not able to reach its goal state yet. In order to do this, it will first have to cross field (D4; wave). What the proposer can do, is request the chip required to cross field (D4; wave) from the responder. As the responder actually owns the chip, the agent can try and get this chip either by trading away its unused chip(s), or by simply asking for the chip. After obtaining this chip, the score of the proposer would be 200 or 210 if it simply obtained the chip by trading. If the proposer were to receive the chip for free, the proposer would reach the goal position and have a remaining chip. This would result in a score of 220. It is theoretically possible that the proposer requests all chips from the responder without offering anything in return, netting a total score of 260, but the responder would then be very likely to reject such an 'offer'. In practice, getting the wave chip from the responder may not be the easiest task: the responder would be unlikely to give a chip it needs to pass field (C2; wave) and thereby reduce its score. The proposer would have to come up with a very good offer to acquire this chip, which may be hard for some of the agent reasoning methods described below.

2.3.2 Optimality Principle

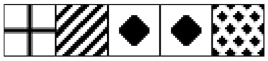
Calculating all paths that result from a chip set is often an intensive process, given the number of paths that can be possible given the conditions in Section 2.1. In order to counter this, we propose the optimality principle. This principle states that given a chip set $\overline{C}$, a chip route found with subset C always be superior to a chip route $C+1$, given that they end up at the same square. As both chip sets make use of the same chips, the chip set that uses an additional chip will never be beneficial for neither the agent nor its responder, meaning we can remove it from the set of possible offers.

2.3.3 Offer Scenarios

The fact that we have reduced the agent's offer space to one that will never contain an offer that is inherently weaker than a different offer, greatly reduces the time spent calculating the scores for the offers an agent can make. However, we are still left with a lot of potential offers, and if we consider the situation as sketched in our previous examples (Examples 2.1 - 2.3), we would still end up with 50+ potential situations to fully work through. As such, we will only use three situations throughout this thesis to illustrate the differences of the algorithms used by the different agents to decide on an offer. These three scenarios can be found in Example 2.4.

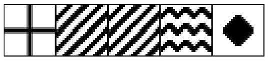
Example 2.4. Three Potential Offer Scenarios

Prop. Chip Set

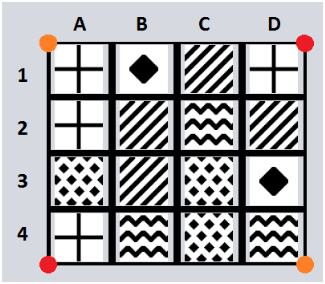


Initial score: **95**

Resp. Chip Set



Initial score: **95**



Suppose the agent with its chip set (Table 1) plays on board 3 (Figure 2c). As we have already seen in Examples 2.1 - 2.3, the agent can (despite using the optimality principle) reach a lot of squares given the chips of both itself and the responder, even when we assume the agent only uses the chips it requires to reach the square it wishes to reach. Three situations have been illustrated below.

1. Scenario 1: The agent can reach square C2 with three chips: (A1; plus) - (B1; diamond) - (C2; wave). It also chooses to keep a stripe chip. This will yield **60** points for the agent. The responder would receive 2 plus chip, 1 diamond chip, 2 stripe chips, and 1 lattice chips. This will land the responder at square A4 ((D1; plus) - (C1; stripe)- (B2; stripe) - (A3; lattice) - (A4; plus), with a total of **210** points. If we look at the situation as given by Examples 2.1 and 2.2, we will see this implies a score change of **-35** for the agent and **+115** for the responder when looking at the initial

situation.

2. Scenario 2: The agent can reach square C3 with three chips: (A1; plus) - (B2; stripe) - (C3; lattice). It also chooses to keep two diamond chips and an additional stripe chip. This will yield **105** points for the agent. The responder would receive 2 plus chips, 1 stripe chip, and 1 wave chip. This will land the responder at square A4 ((D1; plus) - (C2; wave) - (B3; stripe) - (D4; plus)), with a total of **200** points. If we look at the situation as given by Examples 2.1 and 2.2, we will see this implies a score change of **+10** for the agent and **+105** for the responder when looking at the initial situation.
3. Scenario 3: The agent can reach square D4 with four chips: (A1; plus) - (B2; stripe) - (C3; lattice) - (D4; wave). It also chooses to keep a diamond chip. This will yield **210** points for the agent. The responder would receive 2 plus chips, 2 stripe chips, and 1 diamond chip. This will land the responder at square B2 ((D1; plus) - (C1; stripe) - (B2; stripe)), with a total of **70** points. If we look at the situation as given by Examples 2.1 and 2.2, we will see this implies a score change of **+115** for the agent and **-25** for the responder when looking at the initial situation.

One should note that variations of this theme are possible when you consider that the agent can also choose to withhold varying chip sets from the responder to gain varying sets of 10 points per chip. These chips would preferably be chips that are not very beneficial for the responder, but whether the agent chooses the 'more intelligent' chips to keep depends on the strategy of the agent. If, for example, the agent would have chosen to keep a plus chip over a diamond chip in Scenario 1, the agent score would remain at **60** points, whereas the responder score would change from **210** to **95**.

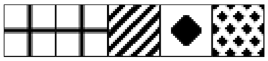
2.3.4 Offer Similarity

The previously discussed fact that board and chip combinations often lead to more than 50 combinations to work through has another consequence: a lot offers will potentially yield the same outcome, despite being different offers. It is even possible for the same offer to have multiple ways leading to the same score. As the algorithm that iterates through the paths is path-based, and not set-based, the set is considered twice in this case. For example, in Example 2.4, Scenario 1, we saw an agent score of -35 points and a responder score of 210, with the agent possessing a plus chip, 2 diamond chips and a wave chip. If, however, we consider a different route with the chips given to the responder, namely (D1; plus) - (D2; stripe)- (C3; lattice) - (B3; stripe) - (A4; plus), the agent would still have a score of -35 points, and the responder would still have a score of 210 points.

In the event that an agent would prefer this so-called score tuple of {agent: -35, responder: 210} over all the other options, we know that there are at least two possibilities to obtain this score tuple, which means that there are at least two offers to choose from that share the same score. In some emulation models, such as the Ficici and Pfeffer model that we will discuss later on, offers with a commonly occurring score tuple are more likely to be chosen, even if it is a bad offer, simply because there is a stronger possibility an agent will come up with this offer purely based on the frequency of a particular offer tuple.

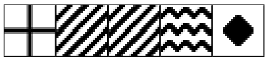
Example 2.5. Offer Similarity Scores for the Three Scenarios

Prop. Chip Set

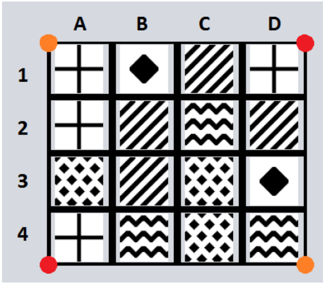


Initial score: **95**

Resp. Chip Set



Initial score: **95**



Scen. 1: Ag. (60), Resp. (210)
Scen. 2: Ag. (105), Resp. (200)
Scen. 3: Ag. (210), Resp. (70)

In this example, we will indicate all the possible tuples that have the same score as the three scores obtained in our given scenarios.

1. Scenario 1 {60; 210}: There are only two possible offers that fulfill this score tuple's criteria are the two offers already outlined. These are the two paths that are possible if Agent (1x plus, 1x stripe, 1x wave, 1x diamond) and Responder (2x plus, 2x stripe, 1x diamond, 1x lattice) is true. The score tuple frequency of {60,210} equals 2.
2. Scenario 2 {105,200}: The only offer that fulfills this score tu-

ple's criteria is the one already mentioned in Example 2.4: Agent (1x plus, 2x stripe, 2x diamond, 1x lattice) and Responder (2x plus, 1x stripe, 1x wave). The score tuple frequency of $\{105,210\}$ equals 1.

3. Scenario 3 $\{210,70\}$: There are two possible offers that fulfill this score tuple's criteria, in this case two unique offers. The first one is already highlighted in Example 2.4: Agent (2x plus, 1x stripe, 1x lattice, 1x wave) and Responder (1x plus, 2x stripe, 2x diamond). The second scenario is exchanging the two redundant chips for one another: Agent (1x plus, 1x stripe, 1x wave, 1x diamond, 1x lattice) and Responder (2x plus, 2x stripe, 1x diamond). The score tuple frequency of $\{210,70\}$ equals 2.

2.3.5 Finding the Opponent Chip Set

One of the main aspects of our Coloured Trails simulation is that the game we use is a partial information game for the agent and its opponent: they do not know each other's chips. This means that we will need to emulate some sort of mechanism that allows the agent and the opponent to make predictions about each other while possessing a minimal amount of information about each other.

There is only one way to robustly deal with this uncertainty: assume the opponent could have any possible set of chips that is possible. In order to calculate the size of this set, we need to know the number of unique chips in the game and the total number of chips the set should possess, after which we can apply Formula 3.

$$n_{chipsets} = \frac{(n_{unique} + n_{chips} - 1)!}{(n_{chips}! * (n_{unique} - 1)!)} \tag{3}$$

With 5 unique chips (plus, stripe, diamond, wave, lattice), n_{unique}, and 5 total chips per agent, n_{chips}, this gives us 126 unique combinations of chips ($n_{chipsets}$). These combinations are then combined with the chips of the responder, which are known, to calculate the 126 unique possibilities that are potentially accessible for the opponent. Each of these 126 unique combinations can have their own number of offers, but calculating and processing all these offers would take a long time. As such, the algorithms we have applied use a maximisation principle: the agent always assumes that the opponent with its chip set will go for the best possible offer with the chip set the opponent has (in other words: the offer with the highest possible utility for the opposing agent). This means the agent assumes the opponent will choose what would be the least beneficial for the agent, but not necessarily the best for the opponent itself.

2.4 Offer Heuristics for Coloured Trails

2.4.1 Weights-based Method (Ficici and Pfeffer)

In the following section, we will describe the theory behind and the implementation of the models by Ficici and Pfeffer (2008).

Algorithm The model by Ficici and Pfeffer (2008) makes use of a fully utility-based function, with a level-n model. In this level-n model, reasoning is simulated by performing *n* steps in a utility-based function, based on different levels of reasoning. This level-n model is described by different proposer and responder models, that can be chained to simulate a deeper level of reasoning. These levels of reasoning describe whether proposers reason about other proposers' thoughts about the thoughts of the proposers themselves, to the n^{th} level. The model explicitly integrates the belief that an opponent is playing according to a certain strategy, and learns to improve based on gradient descent beforehand. This means that, in the different levels of reasoning, the model uses different utility-based approaches, but does not require any tweaking of beliefs in the level of theory of mind an opponent uses (and thus, does not adjust its strategy during its run).

Implementation The Ficici and Pfeffer implementation contains both models for the proposer agents, and for the responder agent, which differ slightly depending on the function the model has in the game. The responder utility function mainly makes use of two basic utility variables: a *self benefit*, specifying the increase in score of a responder when it accepts a certain offer, and the *other benefit*, specifying the increase in score of the proposer when the proposer makes a certain offer. These offers are quantified by two corresponding weights, that can range from -1 to 1.

The proposer utility function also makes use of the two aforementioned utility functions, but has an additional basic utility function: *class size*. This variable component specifies the number of offers in a specific category: offers that have the same *self benefit* and *other benefit* combination are grouped together in the same class. This is the offer frequency that we have discussed before. This variable component is also quantified by a corresponding weight that ranges from -1 to 1.

Making use of the basic utility variables and the weights, the base proposer formula used by the Ficici and Pfeffer model can be expressed as:

$$U(O) = w_{sb} * O_{sb} + w_{ob} * O_{ob} + w_{cs} * ln(O_{cs}) \tag{4}$$

In this formula, we account for the *self benefit* (sb), *other benefit* (ob) and *class size* (cs) respectively, to calculate the utility of an offer. As the *class size* holds no importance for the responder, seeing as they receive one offer per proposer (likely resulting in unique offers), the *class size* for the utility of the responder's choices is set to 1 when looking at the responder model.

Base Proposer Models In order to take into account that the responder can receive multiple offers, and may as such reason about the utility of a move, the proposer can also reason about the utility a responder may attribute to a move. This calculation is, in its most basic form, represented as:

$$Pr(responder\ accepts\ O) = \frac{e^{(U(O)_{agent})}}{e^{v_{agent}} + e^{U(\varphi)_{agent}} + e^{(U(O)_{agent})}} \tag{5}$$

This model does not assume that the responder can reason about other proposers' thoughts yet, and the proposer does not reason about the mental status of other proposers yet either. As a result, the value for the utility score of the other proposers is denoted by an estimated generic utility value, v_{agent}: the value the proposer agent assigns to emulate the other proposers' behaviours. The second value denoted is the utility of the proposer would attribute to the game situation when no chips change hands, denoting the utility a proposer attributes to the retaining the status quo. As the proposers know the chips of the responder, they can calculate this value without having to take a guess. The value is denoted as $U(\varphi)_{agent}$: the value the proposer agent assigns for the responder's initial utility score. The final value used in the model is the utility of the proposer itself, given the proposal it has just sent to the receiver, denoted by $U(O)_{agent}$: the value the proposer agent assigns for its own utility score.

In order to find the maximally beneficial move, a proposer can use the general utility score of agents of its type (t_i) under this circumstance. This (total) expected utility function can be used for the expected utility in the following way:

$$EU(O)_{agent} = U(O)^{t_i} * Pr(responder\ accepts\ O) \tag{6}$$

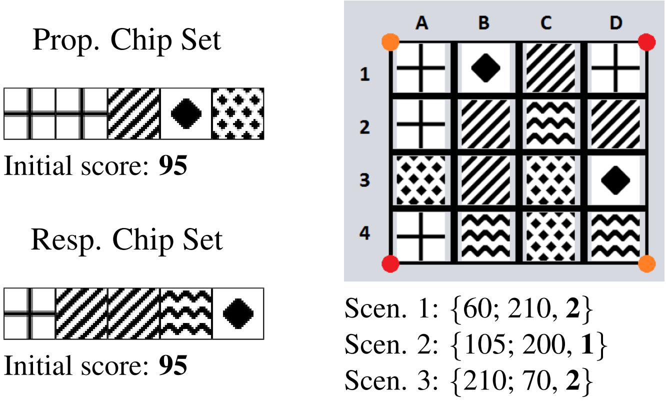
This expected utility is used to calculate the best offer, but the calculation only works if the proposer using it is the only one that reasons about its fellow proposers' and the responder's game plan. In order to account for the fact that the agent may not be the only reasoning entity in the game, the models of Ficici and Pfeffer take into account probability of offer success, $Pr(O|\overline{O})$, rather than the expected utility. This probability can be calculated by having the proposer also take into account the opponent's move utilities calculation.

We can now calculate the overall score over all agent types, given that the expected utility has been normalized ($(EU(O)_{agentNorm})$:

$$Pr(O|\overline{O}) = \sum_i (EU(O)_{agentNorm} * \rho_{agents_i}) \tag{7}$$

In this notation, ρ_{agents_i} is used to denote the number of agents (ρ) adhering to a certain reasoning level i. If this is also the reasoning level of the proposer, this number includes the proposer itself.

Example 2.6. Calculating a Base Parameter Offer



Scen. 1: {60; 210, **2**}
Scen. 2: {105; 200, **1**}
Scen. 3: {210; 70, **2**}

Suppose we wish to know which of the three scenarios is preferred by the Base Parameter Offer Model. In order to do this, we first need to revisit Formula 5, the main formula the Base Parameter Offer Model uses:

$$Pr(responder\ accepts\ O) = \frac{e^{(U(O)_{agent})}}{e^{v_{agent}} + e^{U(\varphi)_{agent}} + e^{(U(O)_{agent})}}$$

In order to analyze the three scenarios, we first need to calculate the three values for $U(O)_{agent}$ and $U(\varphi)_{agent}$ (the other value, v_{agent}, is a parameter, for which we will assume $+50$). For this, we need to use the utility formula:

$$U(O) = w_{sb} * O_{sb} + w_{ob} * O_{ob} + w_{cs} * ln(O_{cs})$$

In order to make the utility formula work, we will first need an estimation of the weight parameters. For the sake of this example, let's suppose the agent is fairly self-focused ($w_{sb} = 0.8$), and does not really care about the thoughts of the responder, but still realizes helping it can be beneficial to itself ($w_{ob} = 0.1$) The agent does not case about the size of its offer ($w_{cs} = 0$). We already know the actual offer values O_{sb} and O_{ob} for the three scenarios. The offer frequency is either 2 or 1 depending on the scenario (for the responder it is always 1, as established in Section 2.4.1). We can now calculated the $U(O)$ for both the agent ($U(O)_{agent}$) and the status quo according to the agent ($U(\varphi)_{agent}$). Assuming scenario 1, we would get:

$$U(O)_{agent} = 0.8 * 60 + 0.1 * 210 + 0 * ln(2) = 69$$

$$U(\varphi)_{agent} = 0.8 * 95 + 0.1 * 95 + 0 * ln(1) = 85.5$$

We can now calculate the value for the likelihood that the responder would accept the offer:

$$Pr(responder\ accepts\ O) = \frac{e^{69}}{e^{50} + e^{85.5} + e^{69}} \approx e^{-16.5}$$

If we do not consider the generalisation that can be provided by looking at what a general Base Proposer Agent would do, we as such find that Scenario 1 has a probability value of $\sim e^{-16.5}$. Repeating these calculations for Scenario 2 and Scenario 3, yields the probability values of $\sim e^{-9.24}$ (Scenario 2) and ~ 1 (Scenario 3). Given the base model, the proposer will clearly prefer the option in Scenario 3, as this is the offer has the highest probability value. This model would offer Agent (2x plus, 1x stripe, 1x lattice, 1x wave) and Responder (1x plus, 2x stripe, 2x diamond). The responder would never accept this offer, given that the offer would result in a negative position for the responder when compared to the initial responder position.

Level-N Proposer Models The level-n models make use of the same formula as the base models, except for the fact that instead of knowing the proposer model, we have to take $N-1$ steps in order to embed the probabilities that the other proposer agent is using a certain level of reasoning. For this formula, the proposer takes four things into account:

1. The probability that an opposing proposer with reasoning level $N-1$ has a certain chip set C in the total set of chips $\overline{C}$
2. The probability that a proposer of reasoning level $N-1$ makes the offer O, $Pr(O|\overline{O})$
3. The probability that the responder (R) accepts the proposer offer O, $Pr(O|O,O,\varphi)$
4. The utility of the proposer's own offer (U(O)).

This yields the following formula for any level-n model ($P\{R\{\},P\{N-1\}\}$):

$$EU(O)=\sum_{C\in\overline{C}}\sum_{(O\in\overline{O})|\overline{C}}(Pr(c)*\overset{P\{N-1\}}{Pr(O|\overline{O})}*\overset{R\{\}}{Pr(O|O,O,\varphi)}*U(O)) \tag{8}$$

Calculating the level-n model's expected utility of an offer requires the utilities of two different types of offers. First of all, there is the offer from the agent itself, denoted by O. The second offer is the offer the opponent would make if it possessed a certain chip set C (in the set of all possible chip sets, $\overline{C}$), denoted by O, part of the set of all possible opponent offers with that chip set, $\overline{O}$.

The final utility still makes use of a score normalization, and the total utility over all agent types can still be summed up as before, with Formula 7. The proposer will select the offer that corresponds with the highest overall utility to select the potential trade offer for the responder.

Parameter Estimation The weights used by the Ficici & Pfeffer model were defined by playing several level-n models against one another, and having them trying to outreason each other. We tested the following parameters for the weights:

1. Weight Offer Self Benefit (w_{sb}): {0, 0.2, 0.4, 0.6, 0.8, 1}
2. Weight Offer Other Benefit (w_{ob}): {-0.5, -0.25, 0, 0.25, 0.5, 0.75, 1}
3. Weight Offer Class Size (w_{cs}): {-1, -0.5, 0, 0.5, 1}

We chose to never include a negative weight for the self benefit, as an agent that was actively looking to improve its own score, likely would not attribute itself a lower utility for obtaining a higher number of points in all possible cases. The Other Benefit and Class Size do not necessarily have this drawback if reasoning from the agent's own view point, especially for the lower tiers of the level-n models.

We have use the parameter sweep three times: one time to find the most optimal parameters for a base parameter model playing against a base parameter model (two base proposer models playing against each other, using any parameter in the set of parameters). This was to find the most optimal parameter setting for a base proposer (Level-0 model). We restricted the parameter space for the base proposer model, as a model unable to reason about its opponents should not give the opponent any advantages over itself. The parameters resulting from this iteration were: $w_{sb} = 1$, $w_{ob} = 0$ and $w_{cs} = 0.5$. For v_{agent}, we always assumed a static 50, as we were more interested in the weights, to use them in the level-n model rather than a base proposer model.

Using the parameters for the base proposer model, we did another iteration, this time playing the base proposer model against the level-n model. The parameter set that was the most capable of dealing with this model was chosen as the level-1 model. The parameters that were produced using this method were:

1. Weight Offer Self Benefit (w_{sb}): 0.8
2. Weight Offer Other Benefit (w_{ob}): 0
3. Weight Offer Class Size (w_{cs}): -0.5

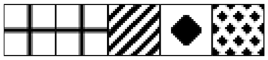
We repeated the approach for the Level-2 model: we selected the best Level-1 model and had it play against a parameter sweep version of the level-n model. This yielded the following parameter for (a likely candidate for) the level-2 model:

1. Weight Offer Self Benefit (w_{sb}): 0.8
2. Weight Offer Other Benefit (w_{ob}): 0.5
3. Weight Offer Class Size (w_{cs}): 0.5

In theory, one could chain this approach to find parameters that suit any n in the level-n model, but the n is theorized to stabilize with higher n's, as there are only so many options to choose from that would result in a decent offer.

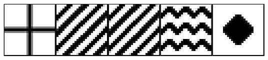
Example 2.7. Calculating a Level-N Model Offer

Prop. Chip Set

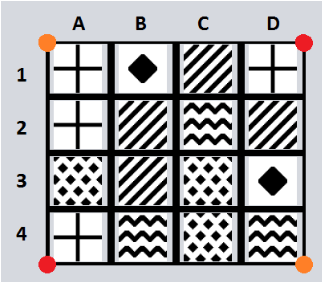


Initial score: **95**

Resp. Chip Set



Initial score: **95**



Scen. 1: {60, 210; **2**}
Scen. 2: {105, 200; **1**}
Scen. 3: {210, 70; **2**}

Suppose we use the weights for a level-2 model, w_{sb} = 0.8, w_{ob} = 0.5 and w_{cs} = 0.5. We can now look at the main formula (Formula 8) for calculating any level-n offer likelihood success:

$$EU(O) = \sum_{C\in\overline{C}} \sum_{(O\in\overline{O})|\overline{C}} (Pr(c) * \overset{P\{N-1\}}{Pr(O|\overline{O})} * \overset{R\{\}}{Pr(O|\overline{O}, O, \varphi)} * U(O))$$

Assuming the chip combination generation algorithm as described in Section 2.3.5, we can assume that the total number of chip combinations is 126. In the event that we were to change the chip combination generation algorithm to allow for chips with the same token to be referred to as different entities (e.g. a set with two chips with a lattice pattern becomes a different set when we change the order of the lattice chips), this number would change for each chip combination, but this would also greatly slow down the currently very generally applicable chip combination approach (specific implementations with preset input would have less issues with this implementation). As our implementation is not specifically designed to allow this diversification of sets, and only allows unique sets, the $Pr(c)$ in the example equals 1/126.

The U(O) is calculated in the same way as with the base proposer model (Example 2.6), but with the new parameter values: the $U(O)$ for scenario 1 equals $0.8*60+0.5*210+0.5*ln(2) \approx 153.35$. In short, the only two real unknown variables are the score the responder would give to any offer in the set of offers, and the score the opponent would give to any offer in the set of offers. The level-n model considers all possible combinations, but as stated in Section 2.3.5, only uses the best potential path for the opponent with all chip combinations, and uses all the path+offer combinations for itself.

A possible combination is for example that the opponent possesses the chip set given in Table 1 (Section 2.2.1), one of each chip (chip combination 126 in our implementation). With one of each chip, the opponent would be able to reach the final square by default (with 210 points), meaning that it would probably choose to hand out none of its chips. It is not possible to improve the opponent's position in any way, meaning the opposing agent will simply choose to offer the initial position. This will result in a $U(O) = 0.8 * 210 + 0.5 * 95 + 0.5 * ln(1) = 215.5$. With a population representation value (ρ) of 1, this means that this probability value is initially attributed $e^{215.5}$. Looking at Example 2.4.1, we can see that the opponent with this chip set would be a fairly big danger.

The agent will also calculate the probability scores according to the responder: $U(O) = 0.8 * 95 + 0.5 * 210 + 0.5 * ln(1) = 181$. Once again, the population distribution is assumed to be $\rho_{agents_i} = 1$, leading to a probability value of e^{181}. As such, we get:

$$EU(O)_{126} = \frac{1}{126} * e^{215.5} * e^{181} * 153.5$$

Do note that this value is only one out of the many values obtained over all the offers. The true $EU(O)$-value will contain $126 * n$ calculations, where n is the number of possible offers the proposer can do that would yield a potentially fruitful path.

2.4.2 Theory of Mind-based Method (De Weerd et al.)

In the following section, we will describe the theory behind and the implementation of the models by De Weerd et al. (2014).

Algorithm The De Weerd et al. model is different from the Ficici and Pfeffer model, in that it does not use raw parameters to decide on an offer, but uses predictions about its opponents' offers by adhering to the belief that they reason in a similar way. While this belief can be likened to parameter values, the approach is inherently different because the model actively models its opponents offers, rather than simply taking into account what the other agents their benefit would be, and tries to find an offer that would work best with both the demands of the responder and the supposed predicted opponent offer. An important aspect in the De Weerd model is that the proposer agent has belief values for which level of theory of mind an opponent is using, and that it adjusts the belief that a certain opponent is using a certain level of theory of mind based on previous experience.

The model designed by De Weerd et al. uses so-called theory of mind chains. These chains emulate a certain level of reasoning (or lack thereof) about the opponent and the responder the agents are playing against.

Theory of Mind Level Zero An agent that uses theory of mind level zero only reasons about itself. This means it will always make the choice that is seemingly the most beneficial to itself. In terms of the game of Coloured Trails, this means it will always choose to make an offer that will yield it the highest score possible with the chips presented in the game setup. In the case of our setup, this would mean the agent would use the 10 chips it can access (its own chips and the chips of the responder) to reach the goal, or get as close to its own goal as possible, without entertaining the possibility that the responder will also wish to reach its goal or get as close its goal as possible. The agent does not even consider the possibility that there is an opponent who is also trying to convince the responder.

Theory of Mind Level One An agent that uses theory of mind level one, reasons about itself and the desires of its opponent. This means that it will make a choice that is beneficial for itself, but also takes into account that the responder would not prefer a negative score for itself and that there is an opponent that will try to maximize its score. In this process, a theory of mind level one agent considers its opponent to be a theory of mind level zero agent: an agent that tries to maximize its own score but does not consider an opposing agent and ignores the fact that the responder would also like to actually reach its goal, rather than just having a goal that provides more points by virtue of proximity.

Theory of Mind Level Two An agent that uses theory of mind level two reasons about itself and the desires of its opponent, while also keeping in mind that the opponent will also reason about the desires of the agent. This means that it will make a choice that is beneficial for itself, while taking into account that the responder would not prefer a negative score and while taking into account that the opponent will also do the same, as it knows the agent also desires to maximize its score. A theory of mind level two agent considers its opposing agent is either a theory of mind level one agent or a theory of mind level zero agent. The theory of mind level two agent considers the responder to be just that: a responder. It simply takes into account the fact it will try and maximize its score. When it comes to the responder, a theory of mind level two and theory of mind level one agent act the exact same.

Theory of Mind Level Three and Beyond While De Weerd et al. do not consider it very likely that humans apply theory of mind chains higher than level two, there are theoretical models that support implementations of theory of mind level three and beyond. De Weerd et al. have implemented the models up to and including theory of mind level four, but the improvements compared to level two models were not significant. A theory of mind level three agent considers its opposing agent is a theory of mind level two, theory of mind level one or theory of mind level zero agent. In other words: a theory of mind level three agent considers the possibility that the opponent realizes that the agent may be predicting what the opponent wants, and actively uses this consideration in its reasoning. As with theory of mind level two and theory of mind level one, the agent's behaviour for the responder is to simply consider it a maximum self benefit responder, an entity that wishes to maximize its own score.

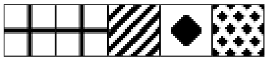
Implementation We have implemented three levels of the De Weerd theory of mind models. Each models adhere to the theory behind theory of mind as closely as possible, meaning that models can explicitly reason about thoughts of others, or are completely unable to do so.

Theory of Mind Level Zero In order to have a working implementation of a theory of mind zero agent, we have to take certain liberties. As the agent does not consider the desires of the responder (let alone the opponent), the agent will always claim the most optimal offer for himself: as close as possible to the goal, and with the highest possible score. The problem with this approach is that the agent will also gain points due to chips that have been left unused. In the most naive theory of mind zero model, this implies that we are left with an agent that always desires all of the chips. This is a side-effect of the game we are using to evaluate the models: games that have no additional benefit from keeping trading assets do not suffer from this point maximisation constraint. Since even the most selfish and uncaring agent will probably realize there is not anything to gain from always wanting all the chips, and always desiring all the chips shows no more intelligence than blatantly grabbing all the chips from the board without any reasoning at all (not even about ones own desires and goal), we propose a slightly tweaked model, theory of mind zero plus. This is the same model as presented by De Weerd et al.

The theory of mind zero plus model works without using the point maximisation constraint, and only looks at the goal at hand. As such, the model will only consider the chips it needs to get as close as possible to the goal, without desiring any additional chips. This way, the model will still reason about its own desires and goals, without automatically claiming all the chips as a consequence of our choice of game. Since there are often multiple ways to reach or get close to the goal with the same number of chips, the agent can choose from offers that yield the same score. Since the only constraint is to get as close as possible to the goal with the smallest number of chips used, and the game itself adheres to certain constraints (Section 2.1), the number of times a tie (due to paths yielding the same score) occurs is fairly big. In the event of these ties, the agent chooses one offer at random.

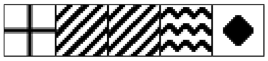
Example 2.8. Calculating a Theory of Mind Level Zero Plus Offer

Prop. Chip Set

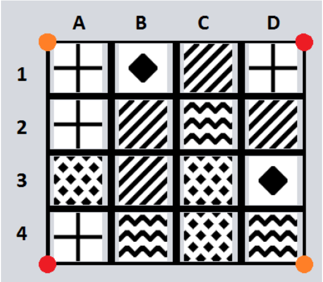


Initial score: **95**

Resp. Chip Set



Initial score: **95**



Scen. 1: {60, 210; **2**}
Scen. 2: {105, 200; **1**}
Scen. 3: {210, 70; **2**}

If we only consider the three offers in our examples, the choice that a theory of mind zero plus agent will make is fairly clear. The agent only considers its own desires, and does not take the fact that there are more players into account. It will as such focus on maximizing its own benefit, which in this case would correspond with Scenario 3. This model would offer Agent (2x plus, 1x stripe, 1x lattice, 1x wave) and Responder (1x plus, 2x stripe, 2x diamond). In reality, this offer would differ and not include Scenario 3, as the agent would never take an additional chip after already reaching its goal.

Theory of Mind Level One The implementation for the theory of mind one algorithm uses the same chip iteration implementation as the Ficici and Pfeffer level-n model (Section 2.6): as the opponent's chip set is unknown, the algorithm considers all possible sets the opponent can have as a potential set. Using the knowledge of the offers the theory of mind one agent can make, and the fact that there are 126 potential opponent sets to work with, the agent construct a set of 'success likelihood' values that equal either 0, 0.5, or 1.

As an agent that includes the reasoning of its opponent and the responder in its behaviour is no longer purely self beneficial, we have added a fourth state for the 'success likelihood' value: -1. This value is attributed to an offer when it turns out that the offer the agent would make would lower its own overall score when compared to the initial situation. Given these four ’success likelihood' values, scores for the combinations are attribute in the following way:

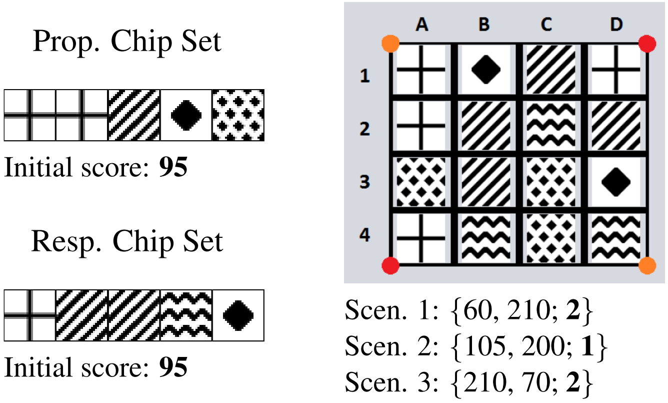
1. Value -1: If the score gain for the proposing agent is negative, attribute a value of -1 by default.
2. Value 0: If the potential opponent offer results in a better position for the responder than the potential ToM1 agent's offer, then the offer will fail (given that the opponent actually goes through with the offer). The agent will attribute a 'success likelihood' of 0 to such an offer.
3. Value 0.5: If both the potential opponent offer and the agent's offer result in the same score benefit for the responder, then the 'success likelihood' is 0.5. After all, given that there are two proposers (one agent and one opponent), the likelihood of the proposer agent's offer being accepted is 50%.
4. Value 1: If the potential opponent offer results in a worse position for the responder than the potential ToM1 agent's offer, then the offer will fail (given that the opponent actually goes through with the offer). The agent will attribute a 'success likelihood' of 1 to such an offer.

After determining how successful each agent offer will be given the potential opponent offer, we can attribute a score to the offer using the theory of mind formula, Formula 9.

$$tomScore_1 = p_{sucToM1} * util_{self_{ToM1}} * conf_{ToM1} + (1 - conf_{ToM1}) * util_{self_{ToM0}} \tag{9}$$

In this formula, the score for using a certain level of theory of mind ($tomScore_1$) is found by multiplying the success value of an offer ($p_{sucToM1}$) with the Coloured Trails score the agent would get when making the offer, $util_{self}$. At the same time, the score also takes into account what the utility for the agent would be given its belief in theory of mind level zero ($util_{self_{ToM0}}$). Both of these utilities take into account that the opponent may not be using theory of mind one, and as such the values are multiplied by the likelihood that the opponent is (or is not) actually using theory of mind one ($conf_{ToM1}$). This likelihood is, as previously mentioned, adjusted by experience.

Example 2.9. Calculating a Theory of Mind Level One Offer



Scen. 1: {60, 210; **2**}
Scen. 2: {105, 200; **1**}
Scen. 3: {210, 70; **2**}

In order to calculate a theory of mind level one offer, we first need to realize that the agent is unable to know the chips of its opponent, meaning that it somehow needs to figure out a way to model its opponent chip set . For this, we have used the method mentioned in Section 2.3.5: as the agent cannot possible figure out its opponent chip set in one turn, it needs to actively model all the possible chip sets an opponent can have. As such, the example outlined is only one of the possibilities for the theory of mind level one model.

One of the chip sets the agents will model is the one outlined in Table 1, the actual opponent chip set: (1x plus, 1x stripe, 1x wave, 1x diamond, 1x lattice). As such, the chips the opponent can access are

(2x plus, 3x stripe, 2x wave, 2x diamond, 1x lattice). With this chip set, the best strategy for the agent to assume (Section 2.3.5) would be to use the main four chips to move to the goal: (1x plus, 1x stripe, 1x diamond and 1x wave). Additionally, the opponent agent would not like to drop in score (it can already reach the goal with its initial chips), so it will also take a random chip from the six remaining chips. In reality, this would not be the case, as the opposing agent would never take more chips than those required for the goal, thus causing a conflicting scenario assuming Scenario 3! As with the level-n model example (Example 2.7), the initial positions cannot be improved in this case.

After establishing that the agent score will be 200, and the responder score will remain 95, given that the opponent proposer refuses to lower its score, it will also work through its own scores. Suppose that the agent proposes Scenario 2 (Example 2.4). This would yield 105 points for the agent itself in the theory of mind level one beliefs, along with 200 points in the theory of mind one beliefs. The agent then decides to give this offer a value. In this case, the value would come down to +1. After all, the responder its score (200 in Scenario 2) yielded by the agent offer far outweighs the responder score obtained when the responder were to stick to the ‘offer’ of the opposing agent.

This +1 is then integrated into the theory of mind level one score (Formula 9):

$$tomScore_1 = p_{sucToM1} * util_{self} * conf_{ToM1} + (1 - conf_{ToM1}) * util_{self_{ToM0}} \tag{10}$$

Assuming we currently believe theory of mind one is fairly likely to be right about the opponent (i.e. the opponent uses theory of mind level zero), we will use a confidence value of 0.9. This is the basic confidence value in our models, before learning from experience.

$$tomScore_1 = +1 * 105 * 0.9 + (1 - 0.9) * 200 = 114.5 \tag{11}$$

As such, the particular offer combination agent (1x plus, 2x stripe, 3x diamond, 1x lattice); opponent (1x plus, 1x stripe, 1x wave, 1x diamond, 1x lattice) would receive a theory of mind one score of 114.5. This offer score will be compared to the other $126 * n - 1$ offer scores, where n is the number of potential offers from the agent, with the highest offer score leading to the offer chosen by the theory of mind level one agent.

Theory of Mind Level Two A theory of mind two agent actually assumes the opposing agent is using the theory of mind one algorithm to decide its offers (while also realizing the opposing agent may use theory of mind zero). As such, it first calculates what its opponent would do given its potential 126 chip sets. This is also where a problem sets in: the theory of mind two agent will soon realize that the opponent does not know the agent's chips either! This would mean that there are a total of 126*126 possible sets, a number which only grows as the theory of mind level increases. In order to counter this we have implemented a slightly adjusted version of theory of mind two.

In this slightly tweaked version, the agent assumes that the opponent does actually know the agent's chips, as this is data that the agent can readily access (after all, this is agent's knowledge about its own position and chips). This serves to greatly decrease the search space the agent would have to go through, as combining 126*126 possible sets with maximisation principle from Section 2.3.5, only assuming the best potential path an opponent and the agent itself can walk, would lead to a great decline in prediction accuracy. This decline in accuracy is caused by the fact that not all paths are iterated over as with the proposer: a human opponent would more than likely consider more than just the path that is most optimal for them. However, not using the maxisation principle would result in fairly long calculation times. By assuming the agent's chips are known to its opponent, the algorithm does not fully emulate the actual thoughts of the opponent, but it does ensure that the agent finds the opponents' thoughts that are most relevant to the actual situation.

After finding the scores that the opponent would have attributed to the potential situations, the agent finds the theory of mind level one scores that the agent itself would have attributed to the potential situations, as it needs these scores later on. The theory of mind level one scores found for the opponent are integrated with the 'success likelihood' system from before, but this time labeled as theory of mind level two scores for the agent:

1. Value -1: If the score gain for the proposing agent is negative, attribute a value of -1 by default.
2. Value 0: If the potential ToM1 opponent offer results in a better po-

sition for the responder than the potential agent offer, then the ToM2 agent's offer will fail (given that the opponent actually goes through with the offer). The agent will attribute a 'success likelihood' of 0 to such an offer.

3. Value 0.5: If both the potential opponent ToM1 offer and the agent's offer result in the same score benefit for the responder, then the 'success likelihood' is 0.5.
4. Value 1: If the potential ToM1 opponent offer results in a worse position for the responder than the potential agent offer, then the ToM2 agent's offer will succeed (given that the opponent actually goes through with the offer). The agent will attribute a 'success likelihood' of 1 to such an offer.

In short: the algorithm for theory of mind level two offers acts in the same way as the algorithm for theory of mind level one offers, only it calculates the 'success likelihood' scores based on opponent ToM_1 offers, rather than opponent ToM_0 offers. These scores are then integrated into a theory of mind score similarly to before (Formula 12).

$$tomScore_2 = p_{sucToM2} * util_{self} * conf_{ToM2} + (1 - conf_{ToM2}) * tomScore_1 \tag{12}$$

This formula is interpreted in a similar fashion as the formula for theory of mind level one, only this time the confidence value for theory of mind level two is used instead, along with the 'success likelihood' of theory of mind level two ($p_{sucToM2}$). As before, both the utilities of the agent itself ($util_{self}$) and the responder ($util_{resp}$) take into account that the opponent may not be the respective level of theory of mind, and as such the values are multiplied by the likelihood that the opponent is actually using a certain level of theory of mind (in this case, $conf_{ToM2}$).

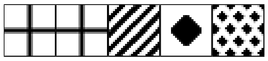
This likelihood is, yet again, adjusted by experience, but there is a big difference now: we are dealing with two likelihoods. A distinct feature of higher levels of theory of mind that we have discussed before, is that they are also able to assume lower levels of theory of mind if necessary. This is why the agent calculated its theory of mind level one beliefs before: the agent also calculates the $tomScore_1$ (Formula 9) and integrates this with the $tomScore_2$. This value is averaged to obtain a final score (Formula 13) on what offer would be best suited to use given the current situation the agent is in.

$$tomScore = \sum_{n=1}^{k} \frac{tomScore_n}{k} \tag{13}$$

In Formula 13, we evaluate all the scores obtained from a certain level of theory of mind reasoning pattern by the agent ($tomScore_n$) based on the number of total theory of mind evaluations performed by the agent (k).

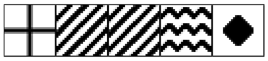
Example 2.10. Calculating a Theory of Mind Level Two Offer

Prop. Chip Set

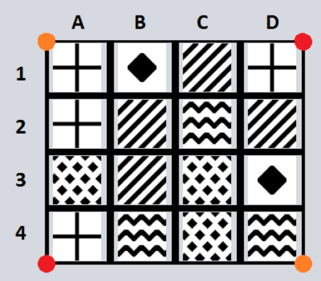


Initial score: **95**

Resp. Chip Set



Initial score: **95**



Scen. 1: {60, 210; **2**}
Scen. 2: {105, 200; **1**}
Scen. 3: {210, 70; **2**}

A theory of mind level two offer is not that different from a theory of mind level one offer. However, it also assumes that the opponent has already applied theory of mind one, i.e. the opponent realizes the offer it comes up with in the hypothetical situation in Example 2.9 may not be the best one. In this case, the agent will realize it has no choice, as every offer it can make will decrease either its own or the opponent's score, leading to an ironically fairly difficult situation for it (despite being able to reach its goal, it cannot improve its score!).

As such, the opposing agent will, despite its theory of mind one beliefs, still come up with the (1x plus, 1x stripe, 1x wave, 1x diamond, 1x lattice) offer. The agent will assign a value for theory of mind level two to this offer. Seeing as the offer is the same, we already know this value will be +1, with the $tomScore_2$ turning out to be the same as the $tomScore_1$, 114.5. These scores are then integrated, which by Formula 13 will also turn out to be 114.5.

If we assume this 114.5 from Scenario 2 is actually the highest score out of the $126 * n$ possible scores, the agent will once again propose (1x plus, 2x stripe, 3x diamond, 1x lattice), likely winning the offer as it is far more advantageous for the responder. When it wins the offer, the agent will realize its theory of mind one beliefs resulted in the same offer as the theory of mind two beliefs, thus attributing this win to theory of mind one beliefs.

Learning Algorithms One of the most distinctive differences between the theory of mind model and the weight-based parameter model is the fact that while the weight-based parameter model is prelearned, and does not adjust its beliefs and strategy during a negotiation session, the theory of mind model adjusts its belief value in whether an opponent is using a certain level of theory of mind to give the most accurate offer for the proposal its opponent will make. The model is adjusted with a learning rate (λ) of 0.6. Both the update algorithm and the learning rate have been inspired by another theory of mind model by De Weerd, that was used to emulate theory of mind in an aforementioned Rock, Paper, Scissors setting (De Weerd et al., 2013).

$$conf_{ToM} = \lambda + (1 - \lambda) * conf_{ToM} \tag{14}$$

$$conf_{ToM} = (1 - \lambda) * conf_{ToM} \tag{15}$$

Theory of Mind Level Zero The theory of mind level zero agent does not consider its opponents, so it does not learn based on experience, either. This agent does not adjust its beliefs about its opponent, as they are non-existent.

Theory of Mind Level One The theory of mind level one agent can do three things: strengthen its current belief in the fact the opponent is using theory of mind level one, weaken this belief or choose to do nothing at all.

1. Strengthen $conf_{ToM1}$: If the agent has won the current bid, it will raise its belief in the respective level of theory of mind for the opponent (Formula 14).

2. Weaken $conf_{ToM1}$: If the agent has lost the current bid to the opponent, it will lower its belief in the respective level of theory of mind for the opponent (Formula 15). In practice, this loss could also be caused due to a situation that the agent could not have won or an unlikely opponent chip set. It would as such falsely change its confidence level in the opponent use of theory of mind. The trials we have used in our experiment (Section 2.5) prevented this design from being a problem.

3. Leave $conf_{ToM1}$ as is: If neither the agent nor its opponent has won the current bid, this could have a lot of reasons. It may be that both offers could never convince the responder without weakening the proposers'

own position, or that both offers were simply badly formulated. No matter the reason, we do not learn enough about the opponent's level of theory of mind to warrant an update, as the opponent offer could be better, weaker or similar to our own in the eyes of the responder.

Theory of Mind Level Two A theory of mind level two agent behaves differently from a theory of mind level one agent in that it takes into account multiple belief values for the opponent's theory of mind. The agent has both a $conf_{ToM1}$ and a $conf_{ToM2}$ to update. As there are multiple belief values to take into account, whereas an opponent can only adhere to one level of theory of mind, the agent assumes that the simplest model that predicted the correct outcome is the correct one. In other words, if both the theory of mind level two model and the theory of mind level one model predicted the same outcome, and this outcome leads to the agent making the correct offer, the agent assumes this was due to its theory of mind level one beliefs.

1. Update $conf_{ToM1}$: The agent first updates its confidence values for theory of mind one. This update occurs in accordance with the guidelines established in Section 2.10.
2. Find $conf_n$ with the lowest correct prediction: The agent accesses its memory to check which theory of mind model made which prediction. The lowest level model with the correct prediction wins.
3. Update $conf_{ToM2}$: The agent then updates its confidence values for theory of mind level two. This update only occurs if the theory of mind level two considerations made a wrong prediction or a correct prediction, while also being the lowest correct prediction (if not, $confToM2$ remains the same). The update for strengthening $conf_{ToM2}$ is provided in Formula 14, whereas the update for weakening $conf_{ToM2}$ is provided in Formula 15. As with theory of mind level one models, if neither the agent nor its opponent traded with the responder, $conf_{ToM2}$ also remains the same.

Updates for a higher level of theory of mind happen in a similar fashion. If for example a theory of mind level four model concludes that theory of mind level two has also provided the correct answer, then the theory of mind level two will be updated, rather than theory of mind level four. A model is only updated when it made the wrong prediction, or the lowest correct one.

2.5 Experimental Setup

We performed both a pilot study and an extended study to see whether participants learned differently against the differeng agent models. These studies were fairly similar in terms of contents, and are both described in this section.

2.5.1 Pilot Study

We initially invited 8 participants (2; all with a higher level education degree) to take part in a pilot study, using the Coloured Trails settings established in Section 2.1. Participants were paid 8 to 15 euros based on their performance. They had to play a game of Coloured Trails against four different agents: a level-1 Ficici and Pfeffer weight-based parameter model, a level-2 Ficici and Pfeffer weight-based parameter model, a theory of mind level one De Weerd model and a theory of mind level two De Weerd model. The pilot was intended to make sure that the models ran properly and that participants were able to complete the task. Participants received an instruction on the game of Coloured Trails, and the specific setup of Coloured Trails they would be playing.

The participants were presented with an abstraction that put themselves in the shoes of a negotiation expert, that had to obtain land from multiple landlords (the respective responder of each round). While performing these negotiations, an opposing entity would also try to obtain land from this landlord, and could potentially outbid the participant. Of course, a complete victory would be achieved when the participant could reach their goal. The paths that a certain chip set yielded were automatically calculated, to remove the potential for participants to mess up because they missed a better route. Of course, the chances of success for the participant increased if the participant and the path-finding algorithm decided on the same path.

We manually designed 13 scenarios with 3 different agent conditions, resulting in a total of 39 potential trials. Each participant had to play 10 randomly selected unique scenarios against all four of the agents (four rounds), meaning that the experiment took the players a total of 40 trials. The other three scenarios were used as practice trials. With 8 participants, this means that each scenario was (recorded as) played against approximately 10 times. Each scenario was developed using the following constraints:

1. There is always a chip available in the set of 10 chips for both the

proposers and the responders to make their first move on the board.

2. There is always room for improvement, e.g. a proposer and the responder can always improve their initial positions by trading with one another, as can the proposer's opponent. The proposer and its opponent are not made aware of this fact.
3. Neither the player nor its opponent will have a major advantage over the other, as that would make the game scenario unfair (we would not learn anything from a player or an agent losing by default).

The three different agent conditions were set up in such a way that one condition had a slight advantage for the opposing agent, one condition had a slight disadvantage for the opposing agent and one condition had the exact same chip set for both the player and its opponent. This was to guarantee the uncertainty, as it left players unable to deduce the chip set of the opponent (which would not be the case if for example the chip set was always the same). Neither the players nor the agents were aware of the fact that we used these three variations.

Participants only had 75 seconds to think of and propose their offer to the responder. This was to add a sense of pressure to the experiment, to emulate the fact that one often needs to reason within a given time (in addition to making sure participants would not take too long to complete the experiment). This time turned out to be a good fit, as the early trials tended to take up a fair amount of time for the pilot participants.

The participants received feedback after each trial. This feedback contained information about their own proposal (score before the proposal and score after the proposal), information about the opponent proposal (score before the proposal and after the proposal), who won the trial (participant, opposing agent or no trade), and gave away the situations for the players and the responder after the potential trade. Additionally, participants were shown what the assets (chips) of their opponent were, making the game a full information game during the feedback (to maximize the learning for the participant). Participants were also shown the paths that were walked after the offer. Participants did not explicitly hear about their overall score improvement (which was tied to the payment): they received this information after the experiment. As such, for all intents and purposes, during the experiment, participants were evaluated per trial, not for the whole experiment.

2.5.2 Extended Study

We invited 34 participants (15 male, 18 female, 1 other; all with a high level of education) to take part in a Coloured Trails experiment following the successful pilot study. The pilot study showed a few minor flaws with the agents, which were fixed for the final run. We also clarified some of the instructions, in response to questions that were asked during and after the pilot participants performed in the experiment. Otherwise, the experiment took place in the exact same way as the 8-man pilot study. Participants were once more paid 8 to 15 euros based on their performance. Participants who had already performed in the pilot study were not allowed to take part again.

The order in which participants played against the agents was randomized, as was the order of the 10 out of 13 randomly selected trials (for each agent). We also transformed the setup for each agent: i.e. we used the same scenario, but interchanged some of the token patterns with one another (for example: diamond tokens became lattice tokens, lattice tokens became plus tokens, stripe tokens remained the same, etc.). This was to prevent the participants from recognizing the scenario when it was used again for the next agent. The difficulty for these scenarios remained the same, which meant that if a participant was faced with scenario 1 with agent 1 with a slight disadvantage, they would also get this setup for agents 2 up to and including agent 4. This was done to allow for a fair score comparison later on.

We also ran a second study, comparing the agent models' performance against each other using the trials we offered to the participants. This was a fairly short simulation pitting all of the agent models against each other, mainly focusing on whether the agent models were comparable in terms of performance (mostly when it came to the level one models versus the level one models and the level two models versus the level two models) and whether their performance was consistent with previous studies (level one models versus level two models).

1. Theory of Mind Level One De Weerd vs. Theory of Mind Level Two De Weerd
2. Theory of Mind Level One De Weerd vs. Level-1 Ficici and Pfeffer
3. Theory of Mind Level One De Weerd vs. Level-2 Ficici and Pfeffer
4. Theory of Mind Level Two De Weerd vs. Level-1 Ficici and Pfeffer
5. Theory of Mind Level Two De Weerd vs. Level-2 Ficici and Pfeffer
6. Level-1 Ficici and Pfeffer vs. Level-2 Ficici and Pfeffer

This analysis is reported on in the Section 2.6.2.

2.6 Recorded Data

We suggest using different recording methods for our two studies 2.5.2, as the data from the human versus agent simulation provides us with different insights than the data from the agent versus agent simulation. The recording methods for both of these studies will be listed below.

2.6.1 Participant versus Agent Experiment Data

We recorded several aspects during the research to find whether there were any differences in the learning behaviour for the participants when faced with a different agent opponent. Additionally, we also recorded the data for the full experiment, to see whether there were any effects from the overall experiment. We measured aspects such as the overall score increase, the time it took to resolve a trial, and the time the participant spent looking at the feedback.

Score Measures

1. Overall score improvement: First off, we measured the overall score improvement that occurred over the 40 trials. This means that we divided the data in four groups: one for each iteration of 10 games (one round) against a random agent that the participant played. By comparing the score improvement (the difference between the number of points a participant receives from their chips after their trade proposal and their initial set of chips) for one participant, over all participants, we can see whether the individual actually learns from a task. As we took out any potential order effects by randomizing the trial order, agent order and transformed every potentially repeated trial, the effect shown should only be caused by learning. We have measured the improvement by use of the overall score per round (per agent played against), as some trials can yield a higher score than others, but the overall score that can be gained per round is the same. Because we expected the participant to learn over the task, the data obtained from the measure is related over the groups.
2. Score improvement per agent: As the agent order was randomized, the effects per agent are different from the round scores found in the overall investigation. For this analysis, we have grouped the data gained

in the experiment in such a way that the scores for playing against a respective agent are placed in the same group. The assumption for this data group is that the different agent groups are unrelated.

Temporal Measures

1. Time required for a trial (overall): Similarly to how we have measured the score, we have also measured the time it required participants to answer the trial. Once again, as we took out any potential order effects, but still expect an overall effect from learning, the data obtained from this measure is related over the data groups.
2. Time required for a trial (per agent): Again, similarly to how we have measured the time required for a trial based on the round the participant is in, we have also grouped together the participant data for each individual agent, e.g. performed an analysis on the time required for a trial against a specific agent. This measure assumes the groups are unrelated.
3. Time required to check feedback (overall): We have measured how long (on average) it took participants to check their feedback for each round, expecting an overall decrease as the trials went on.

Input Measures

1. Find whether there is a shift in clicks from self to the landlord, over time: During the pilot, many participants indicated that they felt like they were slowly shifting their approach from first looking at which chips they themselves needed to which chips the responder needed. This report correlates with a shift in level of theory of mind, making it interesting to look at whether there is a way to show the reported effect. While our pilot did not show the effect in a tangible way, the more extensive study may help shed light on this phenomenon.

Theory of Mind Measures

1. Apply a measure of level of theory of mind used (overall): by adding a score measure for how participant offer corresponds to what a certain level of theory of mind model would come up with, we can see whether the level of theory of mind used increases overall.
2. Apply a measure of level of theory of mind used (per agent): by adding a score measure for how participant offer corresponds to what a certain

level of theory of mind model would come up with per agent, we can see whether the level of theory of mind used differs per agent, showing that different strategies are applied for different agents.

2.6.2 Agent versus Agent Experiment Data

For the agent versus agent simulation, we have only measured the agent score improvement. This analysis was performed by comparing the scores the agents obtained while performing in the scenarios we offered to our participants, and comparing the scores of the agents in both the proposer and the opponent position (with the main score being the proposer role score). By finding the improvement scores gained by the agents, we can find out which of the agents is the best at the negotiation process, and whether any of the agents obtain comparable scores against certain agent types. Parameter sweeps have already taught us that the level-2 Ficici and Pfeffer models outperform the level-1 Ficici and Pfeffer models. We have chosen not to include the time it takes for an agent to come up with an offer in this analysis, as the time it takes to come up with an offer is negligible for lower level models compared to the time required for the general path and score calculations (Section 2.1). For higher level models, both Ficici and Pfeffer (2008) and De Weerd et al. (2014) have shown that the response time increases drastically.

3 Results

We have collected results for both of our experiments: the human versus agent experiment (Section 2.5.2) and the agent versus agent experiment (Section 2.6.2). We will outline the results from both of these studies in the sections below.

3.1 Human Participant Experiment

We performed the experiments according to the description in Section 2.5.2, while recording the data described in Section 2.6.1. First off, we recorded the difference between the number of points a participant would have received if they did not make an offer (their initial position) versus the score after the trade offer with the responder agent (their final position). This measure is also referred to as the score improvement. Secondly, we analyzed the time spent on a trial and the time spent on looking at the feedback. Finally, we also analyzed some of the participants' clicking behaviour during the trials.

We have analyzed the score improvement and response time measures both per round (round 1-4) and per agent (ToM1, named Casey, ToM2, named Jamie, level-1 parameter, named Alex and level-2 parameter, named Riley). The measures per round have been analyzed with a repeated measures anova coupled with paired-sample t-tests, as the assumption for the rounds is that, since we are analyzing the improvement over the participant, the different samples we draw from are from the same population (namely, the same participant). This is different for the agent analysis: the agent order is randomized, so there should be no specific order effects. As we wish to see whether there are significant differences from the individual agent rounds, the assumption here is that the samples are not drawn from the same population, meaning that we have used an anova and regular t-tests while analyzing the differences between the agents.

3.1.1 Overall Score Improvement

We firstly performed a repeated measures analysis of variance for the overall score improvement over the rounds. The one-way within subjects ANOVA was conducted to compare whether the round played by the participant had an effect on their score improvement for all 4 rounds played: round 1 ($mean = 209.71$; $sd = 31.43$), round 2 ($mean = 240$; $sd = 29.74$), round 3 ($mean = 235.29$; $sd = 31$) and round 4 ($mean = 280.74$; $sd = 30.75$). There was an almost significant at the p< 0.05 level for the four rounds ($F(3,99) = 2.137, p = 0.0802$). A significant effect would have implied that the score improvement increased significantly for each round.

As we are particularly interested in the score improvement at the end of the experiment, and not only in the overall score improvement, we performed a few paired-sample t-tests to compare the score improvements over the different rounds played by participants despite an insignificant repeated measures anova (Section 4 discusses this insignificance into more detail). The main effect that we found can be seen (Figure 5) when comparing the score improvement in round 1 to the score improvement in round 4. With the paired t-test, we found that there was a significant difference in the scores for round 1 and round 4; $t(33) = -2.468$, $p = 0.01894$. This seems to suggest that the score improvement increases as the player gets more practice by playing the game.

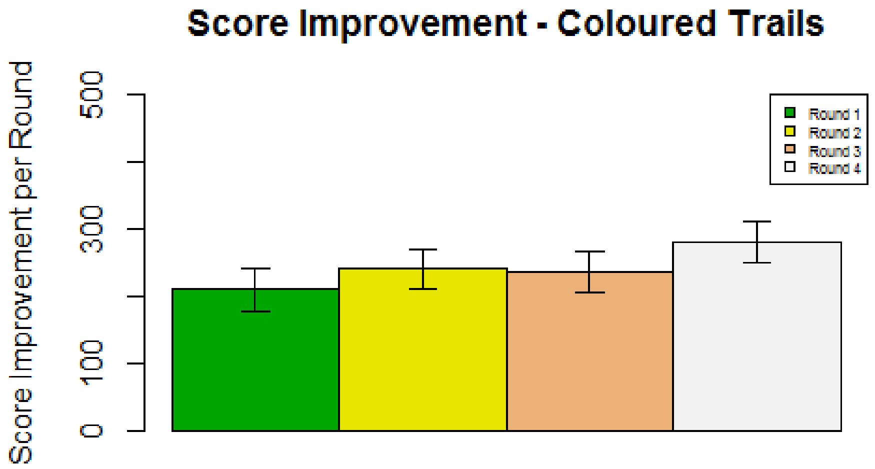


Figure 5: Participant Means for the Score Improvement in Coloured Trails per round

Next to the main effect that we found, we also found that the score improvement in round 2 differed significantly from the score improvement in round 4; $t(33) = -2.3141$, $p = 0.025$. The score improvement from round 3 was not significant when compared to round 4. No other significant effects were found in the overall score improvement data.

3.1.2 Score Improvement per Agent

We performed an analysis of variance for the score improvement differences over the rounds. The one-way ANOVA was conducted to compare whether there was an effect on their score improvement for all 4 agents the participants competed against: the level-1 parameter agent (Alex; $mean = 238.38$, $sd = 32.69$), the level-2 parameter agent (Riley; $mean = 248.53$, $sd = 26.50$) the ToM_1-agent (Casey; $mean = 219.85$,$sd = 30.02$), and the ToM_2-agent (Jamie; $mean = 247.79$, $sd = 32.45$). There was no significant at the p< 0.05 level for the four rounds ($F(3,132) = 0.1913, p = 0.9021$). This severe lack of a significant effect implies that there is no overall different amongst the agents when played against in terms of response times.

We performed post-hoc regular sample t-tests to compare the score improvements players had against the different types of agents, seeing as we were not only interested in an overall significant effect, but also in whether there were significant differences amongst any of the agents. None of these t-tests provided us with significant results ($p \approx 0.50$, $df = 33$). An overview of the means and standard deviations can be found in Figure 6.

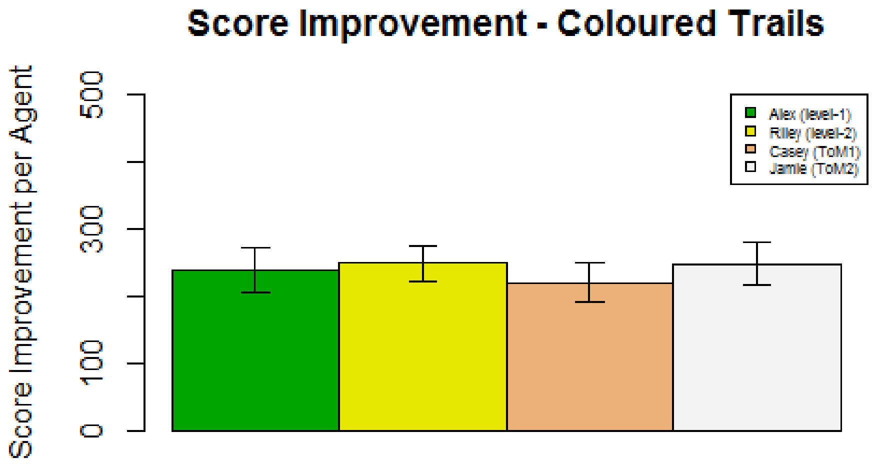


Figure 6: Participant Means for the Score Improvement in Coloured Trails per agent

3.1.3 Overall Response Time Improvement

We performed a repeated measures analysis of variance for the response time differences over the rounds. The one-way within subjects ANOVA was conducted to compare whether the round played by the participant had an effect on their response times for all 4 rounds played: round 1 ($mean = 49.34$; $sd = 1.87$), round 2 ($mean = 45.53$; $sd = 1.88$), round 3 ($mean = 41.20$; $sd = 2.26$) and round 4 ($mean = 37.63$; $sd = 2.14$) . There was an extremely significant at the p< 0.05 level for the four rounds ($F(3,99) = 21.28, p = 1.02 * 10^{-10}$). This significant effect implies that the response times decrease significantly for each round.

We also performed a series of paired-sample t-tests to compare the response time improvements over the different rounds played by participants. The biggest effect that we found can be seen when comparing the time spent on a trial in round 1 to the time spent on a trial in round 4 (Figure 7). With the paired t-test, we found that there was a significant difference in the response times for round 1 and round 4; $t(33) = 6.24$, $p = 4.75 * 10^{-7}$. This significance strengthens the Repeated Measures ANOVA's suggestion that the response time significantly decreases as the player gets more practice by playing the game.

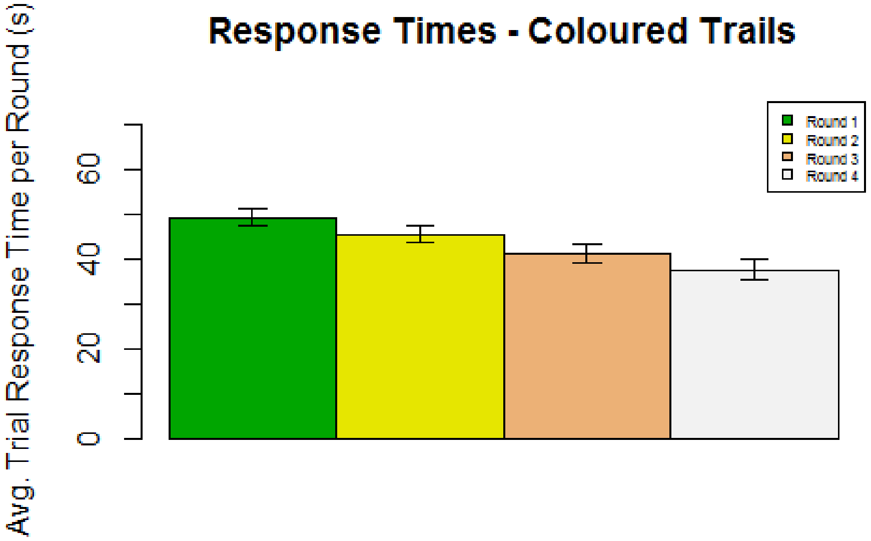


Figure 7: Participant means for the average response times in Coloured Trails per round

The suggestion that is raised by the extremely significant t-test between rounds 1 and 4 is strengthened by the fact that every round shows to be significantly faster when compared to earlier rounds: no matter which round is compared with which round, there is a significant difference between the rounds. The most telling examples are round 1 versus round 2: $t(33) = 2.84$, $p = 0.0076$, round 2 versus round 3: $t(33) = 2.967$, $p = 0.0056$ and round

3 versus round 4: $t(33) = 3.35$, $p = 0.0020$.

3.1.4 Response Time Comparison per Agent

We performed an analysis of variance for the response time differences over the rounds. The one-way ANOVA was conducted to compare whether the agent the participant played against had an effect on their response times for all 4 agents the participants competed against: the level-1 parameter agent (Alex; $mean = 33.30$, $sd = 3.47$), the level-2 parameter agent (Riley; $mean = 36.74$, $sd = 3.23$) the ToM_1-agent (Casey; $mean = 26.18$,$sd = 3.02$), and the ToM_2-agent (Jamie; $mean = 31.37$, $sd = 3.28$). There was no significant at the p$<$ 0.05 level for the four rounds ($F(3,132) = 1.8404, p = 0.1429$). This lack of a significant effect implies that there is no overall different amongst the agents when played against in terms of response times.

Despite the non-significant analysis of variance, we performed regular sample t-tests to compare the response times players had against the different types of agents, as we were not only interested in whether there was an overall difference between the agent groups, but also whether there were any individual differences between the agents. We only found one significant result here, between two agents that were relatively apart in terms of models: the ToM_1-agent and the level-2 parameter model agent: $t(33) = -2.3913$, $p = 0.020$.

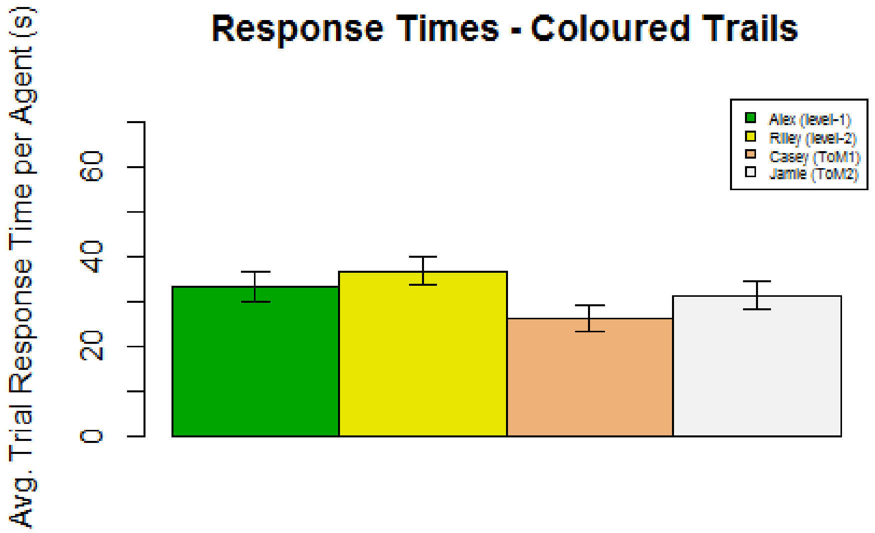


Figure 8: Participant means for the response times in Coloured Trails per agent

The rest of the comparisons yielded p-values between $p = 0.2$ and $p = 0.6$. The bar graph representation of the different agents can be found in Figure 8.

3.1.5 Overall Feedback Time Improvement

The overall improvement in the time spent to check feedback shows effects similar to the overall improvement in response time. We performed a repeated measures analysis of variance for the time spent on checking feedback over the rounds. The one-way within subjects ANOVA was conducted to compare whether the round played by the participant had an effect on their response times for all 4 rounds played: round 1 ($mean = 14.24$; $sd = 1.18$), round 2 ($mean = 8.75$; $sd = 0.82$), round 3 ($mean = 6.88$; $sd = 0.56$) and round 4 ($mean = 5.5$; $sd = 0.64$). There was an extremely significant at the p< 0.05 level for the four rounds ($F(3,99) = 32.77, p = 8.53 * 10^{-15}$). This significant effect implies that the time spent on reading the feedback decreases significantly for each round.

We also performed a series of paired-sample t-tests to compare the time participants used to check their feedback (averaged per round). The biggest effect that we found can, once again, be seen when comparing the time spent looking at the feedback on a trial in round 1 to the time spent on a trial in round 4 (Figure 9). With the paired t-test, we found that there was a significant difference in the scores for round 1 and round 4; $t(33) = 6.80$, $p = 9.415 * 10^{-8}$. This already seems to suggest that the feedback time significantly decreases as the player gets more practice when playing the game.

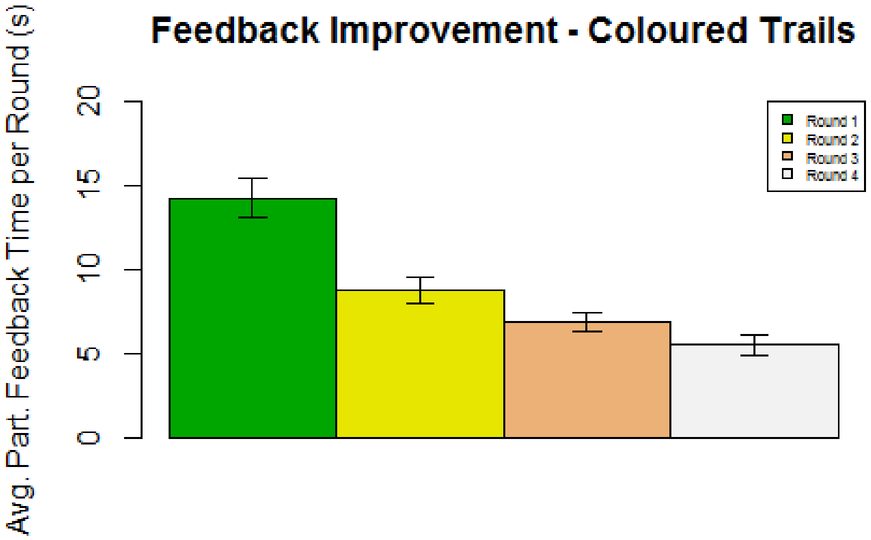


Figure 9: Participant means for the time spent looking at feedback in Coloured Trails per round

The suggestion that is raised by the extremely significant paired t-test between rounds 1 and 4 is strengthened by the fact that participants check their feedback significantly shorter every round: as with the time spent on the trial, no matter which round is compared with which round, there is a significant difference between the rounds. These results include round 1 versus round 2: $t(33) = 5.82$, $p = 1.652 * 10^{-6}$, round 2 versus round 3: $t(33) = 2.528$, $p = 0.016$ and round 3 versus round 4: $t(33) = 3.00$, $p = 0.0051$.

3.1.6 Shift in Opponent Click Behaviour

We analyzed the number of clicks participants used to 'solve' a trial over time by using the key stroke data that we collected. This data, for example, consisted of the mean number of clicks they performed per round: round 1 ($mean = 3.39$; $sd = 0.16$), round 2 ($mean = 3.46$; $sd = 0.20$), round 3 ($mean = 3.46$; $sd = 0.21$) and round 4 ($mean = 3.22$; $sd = 0.21$) respectively (Figure 10).

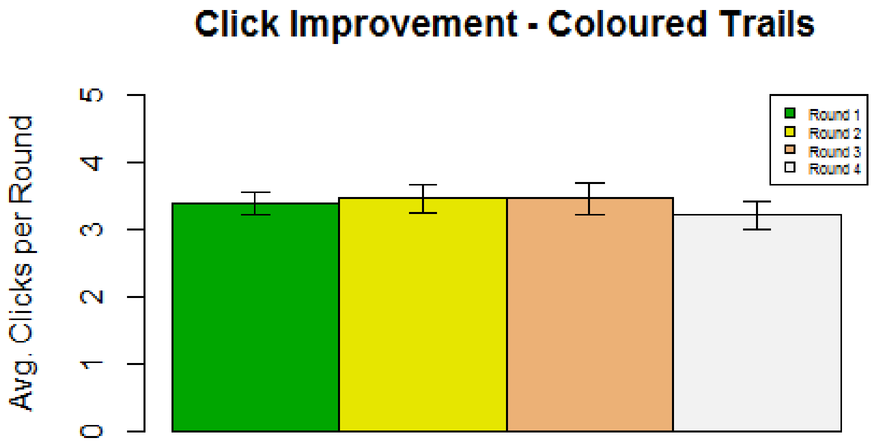


Figure 10: Participant means for the number of clicks per trial in Coloured Trails, averaged per round

We firstly a repeated measures analysis of variance for the difference in the number of mouse clicks over the rounds. The one-way within subjects ANOVA was conducted to compare whether the round played by the participant had an effect on their response times for all 4 rounds played. There was no significant effect at the p< 0.05 level for the four rounds ($F(3,99) = 0.96, p = 0.415$). This lack of a significant effect implies that the number of mouse clicks did not decrease for each round.

Additionally, none of the round combinations showed any significant effects when we performed a paired t-test (which is already slightly indicated by Figure 10, which shows signs that the groups are fairly similar even when assuming not all groups have to be significantly different). Most comparisons showed p-values ≈ 0.6)

We also analyzed whether we could find the participant to landlord click shift (left click to right click) that we theorized would exist based on the pilot (Section 2.6.1). In order to find this effect, we checked whether the participant first gave a chip to the responder or first gave a chip to themselves. We registered these values in the percentage of times participants started by giving a chip to the landlord, rather than to themselves. Unfortunately, the one-way repeated measures analysis of variance did not show any significant shift in starting to reason from the viewpoint of the participant to the viewpoint of the landlord: ($F(3,84) = 0.499, p = 0.0.684$). We did not find any significant increase in the number of times a participant started by giving a chip to the landlord in any of the individual rounds either (most p-values ≈ 0.5, as per paired t-test). The means and standard deviations (round 1: ($mean = 38.28$, $sd = 4.89$); round 2: ($mean = 34.83$, $sd = 5.12$); round 3: ($mean = 37.93$, $sd = 4.74$), round 4: ($mean = 40.68$, $sd = 5.03$)) can be found in Figure 11. We also noticed that the participants tended to reason from their own viewpoint first: more than 50% of first clicks between left and right was a click on the left side, the participants' own side.

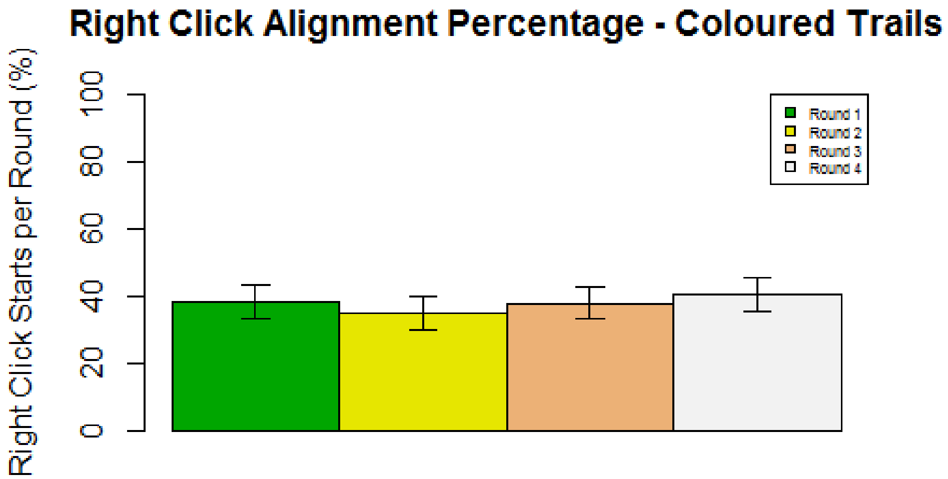


Figure 11: Percentage of starting landlord clicks (right aligment) over starting self clicks (left alignment) in Coloured Trails, averaged per round

3.1.7 Measuring the Participant Level of Theory of Mind

We measured the level of theory of mind the participants were using by applying the offers made by the De Weerd et al. model to the offers made by the participants. This was done by checking how close the offer was to the theory of mind offer that the De Weerd model produced given the same situation. We then assigned a proportion value to how likely the offer was to be of a certain level of theory of mind by subtracting the offer from the De Weerd offers (using the Hamming Distance, (Hamming, 1950)). The similarity scores were divided over theory of mind level zero, theory of mind level one and theory of mind level two likelihood. The problem with this analysis was, however, that the offers differed too much from the offers predicted by theory of mind to extract any meaningful data: any data extracted here proved to be too different to provide a realistic approach (as elaborated on in Section 5.2).

3.2 Agent Simulation

Paired t-tests on the performance records of all agents showed that the agents generally did not obtain scores that differed significantly from one another. Most of the agent comparisons showed p-values around 0.60 ($dF \approx 75$), meaning that there is no significant performance difference between the agents when looking at their performance against each other. We did however perform another analysis, to see what the individual score differences between the respective agents were. These are listed in Table 2.

Table 2: An overview of the mean score improvement per agent condition for all trials

	Opp. ToM_1	Opp. ToM_2	Opp. lv.-2 F&P	Opp. lv.-1 F&P
Player ToM_1	15.77	26.92	36.41	27.95
Player ToM_2	42.82	32.69	52.69	49.87
Player level-2 F&P	43.21	40.51	26.41	52.95
Player level-1 F&P	50.13	43.08	40.51	52.95

In Table 2, we see the playthroughs of the four respective agents amongst themselves, with the players in the vertical column taking the role the human players filled in the participant versus agent experiments. The rows list the average score improvement per trial that an agent managed to obtain against each of their individual opponents. Note that this score is by definition positive if the agents do not trade away chips that worsen their position: if the agent does not trade at all, the score improvement is 0, so if the agent makes 9 unsuccessful bids and improves its score in the next 1 trial, the score will be positive, even if it is (most likely) severely being outperformed by its opponent.

Another point one should note is that Table 2 shows the average score improvement for an agent per trial: as the human participant trials were calculated using a round, or 10 trials, a score comparison with the human participants would require multiplying the results by 10. Even then, this is not a fully fair comparison, as participants only faced off against either a similar setup, a slight chip advantage for their opponent, or a slight chip disadvantage for their opponent. As the chip advantage situation is often better for the score gain than the chip disadvantage situation (as the goal yields +100 points and can often be reached due to this chip advantage), the scores should be corrected (lowered) if one wishes to compare the scores to the human participant scores.

As we can only make a full comparison between our agents if we take both their performance as the player and as the opponent, we have accumulated the scores in another table (Table 3). This table compares the games for the four respective agents amongst one another, teaching us which agent is outperformed by which (overall).

Table 3: An comparison of the agents against each other for all trials

	ToM_1	ToM_2	level-2 F&P	level-1 F&P
ToM_1	-	-15.9	-6.8	-22.18
ToM_2	+15.9	-	+12.18	+6.79
level-2 F&P	+6.8	-12.18	-	+12.44
level-1 F&P	+22.18	-6.79	-12.44	-

4 Interpretation of the Results

We found that overall, as participants progress through the experiment, the number of points they gain when compared to their initial position increases per round. These effects are not immediately significantly visible, but the big contrast between the score improvement between round 1 (and 2) and round 4 shows that participants have definitely become better in the task as they progress. This finding is strengthened by the fact that both the time spent checking the feedback and the time spent on the trial decreases quite significantly for each round. As we see that the score increases, while the response times decrease, it is fairly clear that the participant becomes better at making the strategic decisions as they practice more against the agent simulations.

However, the repeated measures analysis of variance showed (Section 3.1.1) that there was no overall significant effect ($F(3,99) = 2.137, p = 0.0802$) amongst all different rounds. We suspect that this has to do with the performance of participants in round 2 and 3. The score improvement in these rounds is fairly equal, indicating that the participants did not improve their performance in the middle part of the experiment, which, despite there potentially being a general learning trend, influences the analysis of variance p-value in such a way that the overall effect becomes insignificant. Further investigation showed that when the results for round 2 and 3 are paired as one middle round, instead looking at a general learning trend over the start of the experiment, the middle round and the final round of the experiment, we *do* obtain a significant effect with a repeated measures analysis of variance: ($F(2,66) = 3.905, p = 0.025$). This seems to indicate that there is indeed a positive learning trend over the rounds, but that our initial data analysis has hidden this effect.

Unfortunately, similar studies for the participants, but this time with an eye out for the individual score improvements against the types of agents, showed no significant effects. We were unable to find any significant differences between the agents when it came to the participant performance, meaning that they probably performed equally as well against each agent. One exception can be seen when comparing the response times for the ToM_1-agent and the level-2 Ficici and Pfeffer model: participants took significantly longer to answer against a level-2 parameter-based model than against a ToM_1-agent. It is possible that participants were more confident playing against the ToM_1-agent, prompting faster responses while getting approximately the same score improvement overall. However, since none of the other data seems to indicate that this difference is truly caused by the agent models, this is probably caused by different, confounding effects.

We did not find any significant effects while looking at the change in the clicking behaviour of the participants over time. The number of clicks did not decrease over time either, nor did the relatively self-centered clicking behaviour of the participants.

When we looked at the agent models competing among one another, we did not find any significant performance differences in the task itself, but we did find that the agents that were expected to outperform the other agents, actually did outperform their opponent (Table 3). This is not only seen when looking at the ToM_2-model versus the ToM_1 model and the level-2 parameter model versus the level-1 parameter model, but also when comparing the ToM_2 model to the level-1 parameter model and the level-2 parameter model to the ToM_1-model. Somewhat remarkable is that the ToM_1-model is completely outperformed by the level-1 parameter models. We will elaborate on this in the Discussion (Section 5.3.1).

5 Discussion

There are many factors in our research that are worth discussing: the task parameters (Section 5.1), the methods we used to track the reasoning skills of our participants (Section 5.2), and the influences that the models themselves had on the research (Section 5.3).

5.1 Task Parameters

One of the major aspects that could benefit our research, is to take another look at the task parameters. There are numerous factors that may have hampered the participant performance, such as the task difficulty, the offer path calculation and the method we have used to track participants' exact reasoning skills. We will elaborate on these below.

5.1.1 Task Difficulty

One of the main things we can see in the participant versus agent experiments is that while there are clear overall learning effects, there appear to be no differences in the participant learning when we compare the individual agents. We suspect that these two things may share a slight connection: it is clear that the participants improve over time, which indicates that the task is not one that participants can fully effectively perform after the three practice trials that they are offered. These practice trials may help with *understanding* the game mechanics, but do not teach the participant to perform *well* in the task. During the task, we made participants realize that there are perhaps multiple entities striving to accomplish the same thing, at the cost of the human participants' own performance, and that pleasing your negotiation partner can help in sealing the deal. While this means that our variant of Coloured Trails may be an effective tool in teaching people to negotiate under pressure, it also means that the task may be too difficult to fully see the agent differences come to light, as the participants were not capable enough to recognize the individual differences between the agents.

During the task, participants slowly begin to realize that they need to take their opponent into account, but the fact that their opponent's assets (chips) are unknown, may often lead them to make an estimate of the opponent offer (such as seen in the Base Proposer Ficici and Pfeffer model), rather than fully emulating the responder's potential offers and attributing any efficiency values to them. This is in stark contrast to earlier experiments performed by De Weerd et al.

In the study by De Weerd that we used to build one of our models (2015a), the opponent chips are known, meaning there is no need for reasoning about the potential scenarios that could be going on. Another study by De Weerd includes this uncertainty component, but only has a proposer reasoning with an opponent in order to reach their own goal (De Weerd, Verbrugge, & Verheij, 2015b) within a set number of turns, which means that both the agent and the proposer gave away information about themselves during the process, while at the same time getting more than one chance to get it right and only having one agent to worry about. Both of these studies showed clear differences between the agent models, but they both gave more information to the proposer than in our setup.

5.1.2 Limitations of Offer Paths

In the end, the combination of a one-shot game, with an unknown opponent strategy and multiple agents to consider during the trade, may be detrimental for finding any differences between the agents: the uncertainty human participants and agents face with respect to their opponent, along with being unable to determine an opponent strategy from multiple offers in the same trial, severely complicate the game (Section 5.1.1). A factor that adds to this is that we were constrained in our computing power: we were unable to calculate the full paths for the unknown opponent strategies, instead always assuming the opponent will make the most optimal decision for themselves (which in reality, is not always true). This means that the number of paths the algorithms could choose from was by default limited to 126, which limited the offer space to 126 as well. This issue caused by the maximisation principle 2.3.5 'ensured' that the agents had fairly similar paths to choose from.

There is, however, another limitation to the offer paths, which is probably an even bigger constraint to finding differences between the agents: we played our games on a 4*x*4 board, with only one direction heading directly toward the goal for both the proposer and its opponent (upper right corner to bottom left corner) and the responder (upper left corner to bottom right corner). This severely decreases the number of potential routes, meaning that it is easier to make a decision for participants, but also that there are less routes to choose from for the agents, decreasing the likelihood that for example a ToM_1 and ToM_2 model would make a different choice as the number of choices that improves the score of both the proposer and the responder is fairly low.

5.2 Tracking Methods for Reasoning

One of the major flaws we found with our analysis was that while we were clearly able to see whether participants were able to outreason a model, we were unable to accurately decide what reasoning level participants used, as the offers participants came up were very different than those generated by the models from the experimental data alone. This could potentially indicate that none of the models are able to capture what a human would respond all that well, but one should also take into consideration that the model response may be an optimally playing human, whereas a lot of the humans in our data set did not adhere to this optimal playstyle.

Except for when there were very slight deviations from the initial participant chip set, our human participants did not often come up with the exact same offer (despite the small offer space, as human participants often chose to make chip offers with chips they did not need). In other words: it could be that the models accurately model the theory of mind levels that they are supposed to emulate, but are just like another (more optimal) human player. After all: the agent versus agent experimental data shows that agents are able to outperform each other with their respective levels as they should.

Future studies may be able to find a more precise method for deciding on the reasoning level of a participant. Experiments that are specifically designed to measure the level of theory of mind a proposing participant uses have shown that it is possible to develop agent models that *are* more easily able to decide on the opponent level in several games (De Weerd, Diepgrond, & Verbrugge, 2016). This is something that our experiment could in theory also do, but it would for example require an observer agent to run concurrently with the level-n models to constantly obtain the opponent level of theory of mind, which is something we have not done.

5.3 Potential Modeling Improvements

There are numerous aspects in our models that could be changed. In this section, we will discuss the actual differences between the level-1 and the theory of mind level one model, the differences between our parameter-based models and Ficici and Pfeffer's parameter-based models, and the use of offer-based versus path-based solutions to the game of Coloured Trails. We will also elaborate on a minor error in our theory of mind two model.

5.3.1 Performance of the Level One Models

A consistent trend in both the participant experiment and the agent versus agent experiment was that the performance of the theory of mind level one model was worse than the other three models. For the level two theory of mind model and the level-2 parameter model this is a logical consequence of a lesser reasoning capability. For the level-1 parameter model, this seems a bit stranger. We hypothesize that this big difference, especially in the agent versus agent simulations, comes from the fact that the level-1 parameter model is capable of reasoning on a higher level than the theory of mind level one model.

The theory of mind level one model is still relatively egocentric: the theory of mind level one model reasons about its opponent's state of mind, assuming itself to be superior to the world in any way (after all, according to itself, the theory of mind level one model is the only agent that can reason about the mental state of others). The level-1 model, however, already reasons about the opponent's beliefs about itself, but it diminishes the effects of this reasoning capability with level-1 parameters. This means that the level-1 parameter model is theoretically able to reason from a theory of mind level two perspective, even if this reasoning capability is fairly weak and suppressed. This means the level-1 parameter model may be superior to the theory of mind level one model, explaining the fact that it defeats the level one model when they are pitted against each other. We would have to further suppress the level-1 parameter model to make for a fair comparison, which would mean that we need to influence it by using data from humans that reason with theory of mind level one.

5.3.2 Human Data for the Ficici and Pfeffer Model

One of the main disadvantages of our model is that it is merely inspired by the Ficici and Pfeffer model, not fully based on it. While the models behave in an expected way, as seen when comparing the models' overall performance on the task with one another, one of the reasons we did not manage to find a significant difference in the participant learning capabilities for the parameter models versus the direct theory of mind models may be that we merely emulated the human data in the Ficici and Pfeffer model with parameters, giving the parameter model a bit more of an artificial touch. This is in stark contrast with the data Ficici and Pfeffer had: they used the data from offers made by human participants in two extensive human participant studies. After performing the studies, they used the data to train their models to emulate the human offers. Only then did they run their human versus agent experiments. As we only had 8 participants in our pilot study, and used the data to estimate the weights, rather than specifically train ‘likely’ offers, our model is somewhat limited in terms of human offer data.

The consequence of this artificial touch is that both models are solely based on manipulations of data that are directly derived from the game scenarios. Even if the ideas behind and the eventual implementation of the two models is entirely different, the lack of human data causes them to be a bit more similar. This similarity may be a partial contributor to the similarity found when comparing the distribution of the results in terms of agent groups.

5.3.3 Path-based Trekking versus Offer-based Trekking

One of the choices we made while modeling concerned what counts as an offer and what does not. We chose to use a path-based solution for Coloured Trails, because offer-based solutions (Section 2.2.3) take up a lot of computing time due to having to find all possible offer permutations beforehand. This does mean that in a set with four diamond chips and one plus chip, an agent is just as likely to look for routes for a plus chip than for routes for a diamond chip. For humans, this would be different: while having multiple tokens that refer to the same field does not mean anything for our agents, the higher chip count for one of the tokens makes it more likely for a human to choose a path with that token, due to the fact that these chips have a higher activation value (Verplanken & Holland, 2002).

At the same time, offers that lead to multiple paths (that yield the same score) are counted twice, due to the path-based solution for Coloured Trails. This, however, is something that we do not regard as a bad thing, as participants can also solve a Coloured Trails scenario by looking at the paths and relating these paths to chips that they can hand in to pass the fields on the board. In the event of this strategy, they would be twice as likely to find a solution with that offer (as, after all, there are two solutions), despite the fact that there is only one ‘literal’ offer these solutions have been drawn from.

5.3.4 Theory of Mind Level Two Model Calcuation Discrepancy

After performing our experiments, we found that we had made a small error in our theory of mind level two model. The formula for the $tomScore_2$ (Formula 12) that we used in the experiments, made use of the $util_{self_{ToM0}}$, instead of the $util_{self_{ToM1}}$. This means that the $tomScore_2$ did not directly integrate the $tomScore_1$. However, as we did a double integration step by using Formula 13, all three scores were represented. As both the theory of mind level one and theory of mind level two models integrated the score for theory of mind zero, however, theory of mind zero has been weighed twice in the theory of mind level two model, meaning that the agent may make decisions that seem a bit more selfish. This did not cause any major issues according to our agent versus agent simulations (Section 3.2), but should be rectified in future iterations of the theory of mind level two agent algorithm for Coloured Trails nevertheless.

5.4 Future Research

When looking at the future, there are two obvious paths to take: the first path is to continue looking into the benefits of Coloured Trails in making strategic decisions, one that is likely to be fruitful looking at the results found by our thesis. The second path is one of improvement: by improving our current models it may be possible to diversify the models in such a way that their differences become clear during the experiments, rather than only from the theoretical framework. There are two other options for future research to consider: we could tweak our version of Coloured Trails and see whether the learning effects that we found remain the same, and we could look into models that allow agents to deduce their opponents use a similar level of reasoning. We will discuss all four of these options below.

5.4.1 Learning Potential for Coloured Trails

The learning benefits our Coloured Trails application provide could easily be expanded on towards different scenarios (testing more strategically interesting scenarios than only our original 13), and in different abstractions, to see whether the effects found with our Coloured Trails study also hold up when applied in real life situations. Whereas we have shown that there is a clear strategic decision-making improvement curve for the theoretical Coloured Trails game, we have not shown that the effects can be translated to any practical applications of theory of mind.

This could be accomplished by performing social studies that compared managing tasks before and after practising by playing Coloured Trails. As our studies indicate that participants had not fully mastered the task yet (their performance was still increasing in the last round), these studies would have to use more than the 40 trials that our task adhered to, meaning we would have to have multiple training sessions to determine whether there are any lasting effects (as our study already took up to an hour). Additionally, it would be interesting to see whether these multiple training sessions would affect every day decision-making as well.

5.4.2 Highlighting the Agent Differences

As there are some minor differences between our parameter-based methods and the Ficici and Pfeffer model because of the lack of human data, making our agents more distinct from one another can be accomplished by tackling that problem by following Ficici and Pfeffer's approach to include human versus human trials and translating the data into the level-n model. However, as mentioned before, most of the diversification would have to come from changing the experiment to give away more information about the opponent. The only way to truly distinguish between the agents based on our experimental setup is probably to do multiple training sessions with different scenarios, to truly make sure participants have mastered the task. Only then, they can tackle the agents and the chip estimation based on an individual level (the estimations as performed by our simulation would require a lot of computing power from an inexperienced participant), rather than by use of a low level Ficici and Pfeffer (base level model) score estimation strategy.

5.4.3 Diversifying Coloured Trails

The third option for future research lies in the overall applicability of our Coloured Trails model. While we already considered using different scenarios and applying the models to training for real world decision-making (Section 5.4.1), it is currently not known how dependent our results were on our specific setup. We could further complicate the task by increasing the game board size from $4x4$ to $5x5$, could consider an agent starting point other than a top left corner start, could add uncertainty about the goals of the responder and the opposing agent, and could change our setup from a one-shot to a multiple-shot negotiation game or inform the players about each other's chips.

We expect that increasing the board size will only help when the human participants are already adequate at solving the task. The more board fields and potential routes are involved, the harder the task will get as the number of options grows significantly. The number of offers will remain approximately the same, but the consequences of these offers will increase. An agent starting point different from a corner position, and the lack of information concerning the responder goal, would both also increase the number of potential routes, especially on boards bigger than $4x4$, complicating the task even further.

The opposite is true when we change our game from a multiple-shot to a one-shot negotiation game. If the participants could make multiple offers to their responder, they would be able to deduce the opponent strategies more easily, giving the participants more to work with when compared to having to determine their opponent strategy from post-trial feedback. This option would probably help in determining the participant learning differences against different agents, as the fact that they can try to outreason their opponent becomes far more obvious to them, in addition to giving them more chances to learn in the first place.

The same holds when we remove part of the uncertainty, and tell the participant which chips the opponent possesses and the other way around. While this would help the participant with learning to determine their opponent strategy, and thus learning how to complete the task, the opportunities for any abstraction to the real world decrease, as in the real world humans often do not know what the assets of their opponent are.

Overall, changing up the settings we have used may be an interesting choice to see whether effects of learning persist or improve, but one needs to carefully consider which options to add, as adding more uncertainty may lead to human participants reasoning even less about their opposing agents. This would result in the participants getting better at the task in general, instead of actively becoming negotiation experts in a multitude of situations (i.e. agent strategies).

5.4.4 Equality of Theory of Mind

A fourth, less obvious, research direction can be found in looking at the realism of both models. One of the major flaws of both the direct theory of mind model by De Weerd et al. and the Ficici & Pfeffer parameter model is that an agent is by definition a higher level agent if it reasons against something that could have been its peer. Whereas humans can reason that someone is their equal, the agent models are always one step ahead of their opponent. It is not possible to have a theory of mind level one agent realize that its opponent is a theory of mind level one agent, nor is it possible to have a level-2 parameter model realize its opponent is also a level-2 parameter model. This detracts from the realism of the models, and it may be possible to expand on the models and/or theory behind the models in such a way that considering an opponent may be on an equal level becomes possible as well.

5.5 Final Words

While it is definitely possible to improve our already successful models, the true improvement for showing differences between the parameter principle and the direct applied theory of mind principle is to adjust the task offered to the participants accordingly. Future research with a different research approach, such as a simplified Coloured Trails task, may show more of the individual model differences, but at the same time runs the risk of diminishing the overall learning effects due to the task's increased simplicity.

Our model has shown that we can use the two player-one responder-variant of Coloured Trails to teach players how to negotiate under time pressure. This shows that Coloured Trails can be an efficient tool to help teach humans how to balance their own goals versus those of potential competitors and negotiation partners, helping them deal with mixed-motive situations.

References

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, *103*(418), 570–585.

Arnold, K., Gosling, J., & Holmes, D. (1996). *The Java Programming Language* (Vol. 2). Addison-Wesley Reading.

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

Bottino, R., Ferlino, L., Ott, M., & Tavella, M. (2007). Developing strategic and reasoning abilities with computer games at primary school level. *Computers & Education*, *49*(4), 1272 - 1286.

Byrne, R., & Whiten, A. (1989). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans (Oxford Science Publications)*. Oxford University Press, USA.

De Weerd, H., Diepgrond, D., & Verbrugge, R. (2016, July). Estimating the use of higher-order theory of mind using computational agents. In *LOFT2016.*

De Weerd, H., Verbrugge, R., & Verheij, B. (2013). 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). Agent-based models for higher-order theory of mind. In *Advances in Social Simulation* (pp. 213–224). Springer.

De Weerd, H., Verbrugge, R., & Verheij, B. (2015a). Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Journal of Autonomous Agents and Multi-Agent Systems*.

De Weerd, H., Verbrugge, R., & Verheij, B. (2015b). Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 1–38.

Dore, R. A., Smith, E. D., & Lillard, A. S. (2015). How is theory of mind useful? perhaps to enable social pretend play. *Frontiers in Psychology*, *6*.

Ficici, S. G., & Pfeffer, A. (2008). Modeling how humans reason about others with partial information. In *Proceedings of the 7th*

International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1 (pp. 315–322).

Frith, C., & Frith, U. (2005). Theory of Mind. *Current Biology*, *15*(17), R644–R645.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, *29*(2), 147–160.

Press, W. H., & Dyson, F. J. (2012). Iterated prisoners dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, *109*(26), 10409–10413.

Sutton, J., Smith, P. K., & Swettenham, J. (1999). Bullying and theory of mind: A critique of the 'social skills deficit' view of antisocial behaviour. *Social Development*, *8*(1), 117–127.

Verbrugge, R. (2009). Logic and Social Cognition. *Journal of Philosophical Logic*, *38*(6), 649–680.

Verbrugge, R., & Mol, L. (2008). Learning to apply theory of mind. *Journal of Logic, Language and Information*, *17*(4), 489–511.

Verplanken, B., & Holland, R. W. (2002). Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology*, *82*(3), 434.

Vygotsky, L. S. (1980). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

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