university of groningen / faculty of science and engineering / artificial intelligence

Human behaviour analysis of human-human and human-agent interactions in the Mod-Signal Game

Artificial Intelligence Bachelor's Project Thesis

Abi Raveenthiran, s4010132, a.raveenthiran@student.rug.nl
Supervisors: mr. J.D. Top, MSc & dr. H.A. De Weerd

Abstract: The resent surge in popularity of AI in recent years has led to AI systems becoming more and more integrated in our daily lives. AI systems have also shown a lot of improvement in recent years making it hard to distinguish the output of AI systems from human output. Previous research has found that humans behave differently when they know they are interacting with an AI, but what happens if they do not? This experiment explores the human behavioural differences when playing with a human and when playing with an agent in the Mod-Signal Game. In the Mod-Signal Game two players simultaneously choose a number from 1 to 24. A player gains a point when their number is exactly one higher than the other player's number. Before every round there is a signalling phase, which allows one of the players to signal which number they are going to choose before the round starts. The player does not have to choose the number they signalled. It is up to the other player to decide whether to trust this signal. In this experiment participants play the Mod-Signal Game for 3 blocks, in which 2 of the blocks they play with humans and in the other block they play with an agent. The participants are made to believe that they are playing with a human, when they are actually playing with the agent. This is also the first time the Mod-Signal Game has been studied in a human-human setting. The results show that despite the increased payoffs for playing cooperatively, humans still tend to play competitively. The results also show that participants had a hard time distinguishing in which blocks they played with a human and that unknowingly playing with an agent did not have a significant effect on the participants behaviour in comparison to their behaviour when playing with a human.

1 Introduction

Artificial Intelligence (AI) has had a surge in popularity in recent years. This has led to AI systems becoming more and more integrated in people's daily lives. The AI systems have also become better over the last years, as the output of the AI systems are becoming increasingly harder to distinguish from human output (Gunser et al., 2021). Previous research has found that humans behave differently when they know they are interacting with an AI system (Bos, 2022; Bellaiche et al., 2023), but how do they behave if they do not know they are interacting with an AI system? If the output of AI systems and human output becomes harder and harder to distinguish, you would also be expect that people would behave the same when they are interacting with an AI or a human. This thesis builds on a bachelor's thesis written by Bos (2022) to analyse human behaviour in human-human play in the Mod-Signal Game as well as comparing that to the human behaviour shown in human-agent play. In the thesis by Bos (2022) an experiment is conducted in which participants play against multiple agents with different strategies. To research whether participants respond differently when they believe that they are playing with a computer compared to playing with a human. The participants are tricked into thinking that they are playing with a human, while they are actually playing with one of the agents. For the other agents, the participants are specifically told that they are playing with an agent. This is done to see whether the perception of whether a participant thinks they are playing with agent or an human affect their behaviour, as it was found that this perception is important within competi-

tive game settings (McGloin et al., 2016). The game that is utilised in the experiment is the Mod-Signal Game (Bos, 2022). The Mod-Signal Game is a game based on the Mod Game (Frey & Goldstone, 2013), where n players have to choose a number from 1 to m simultaneously. Players gain a point for each participant that picked a number that is exactly one lower than the number that they picked. The Mod-Signal Game has the same rules as the Mod game but is played with $n = 2$ and $m = 24$. It also adds a signalling phase before every round, where one of the players has to signal a number that they might choose to the other player. The players do not have to tell the truth and are allowed to play a different number, it is up to the other player to decide whether they trust the signal. A more detailed explanation on the Mod Game and the Mod-Signal Game can be found in section 2.

This bachelor's thesis will be a continuation on the previous bachelor's thesis written by Bos (2022) to further improve the validity of the experiment, as there were uncertainties about whether the participants knew that they did not actually play with a human. This experiment will also introduce a new implementation of the Mod-Signal that allows for human-human play, which has not been done before. It is interesting to study the strategies that humans use when playing with each other. This implementation can be used as a tool for future research in Theory of Mind (see 2.1) within a human-human setting. The implementation of the Mod-Signal Game should be used for this as the signalling phase and the large action space makes it a lot easier to interpret the strategies and Theory of Mind orders of the participants.

This experiment also determines whether the behaviour of the participants differs when playing with another human and playing with an agent and determines whether the participants were able to identify when they played with a human or an agent. This is comparable to a Turing Test (Turing, 2009) on human behaviour in the Mod-Signal Game. The remainder of this thesis will describe the Mod-Signal Game in detail (section 2), the setup of the experiment in detail (section 3), the results of the experiment (section 4) and finally section 5 will conclude it.

2 The Mod-Signal Game

In this experiment, the participants played a mixed-motive game called the "Mod-Signal Game". The Mod-Signal Game introduced by Bos (2022) is an adaptation of the Mod Game (Frey & Goldstone, 2013). In the Mod Game, n-players have to choose a number from 1 to m, where n and m are both larger than one. A player gains a point for every other player that chose the number that is exactly one lower than their chosen number. For example, if participant A chooses number 3 and participant B chooses number 4, then participant B gets a point. There is one exception to the scoring rule, the number 1 is considered exactly one higher than m. In the Mod Game every number has a number that it beats and a number that it gets beaten by, similar to the rock-paper-scissors game. In fact, the Mod Game is a non-zero sum numerical version of rock-paper-scissors if $n = 2$ and $m = 3$.

The Mod-Signal Game uses the same rules as the Mod Game with $n = 2$ and $m = 24$ and adds an extra signalling phase before every round. In the signalling phase one of the players has to signal a number to the other player. From now on, the player that signals a number will be referred to as the signaler and the player that receives the signal will be referred to as the responder. The signaler does not have to adhere to the number that they signalled when choosing a number. They can either choose to play truthfully and choose the number they signalled or they can try to deceive their opponent by choosing a different number. It is up to the responder to decide to trust the signaler. The players take turns signaling every round (e.g. player A is the signaler in round 1, then in round 2 player B is the signaler and in round 3 player A is the signaler once again). In figure 2.1 the user interface of the game is pictured. The numbers are drawn in a circle, this makes it easier for the players as they do not have think about that number 1 is exactly one higher than 24. In the case of the user interface it will always be the next number when going through the circle clockwise. To make it clear that the signaler has sent their signal, the signalled number will turn red for the remainder of the round on both players' screens.

The Mod-Signal Game was chosen over the Mod Game as the Mod Game is only researched for competitive behaviour, whereas the Mod-Signal Game

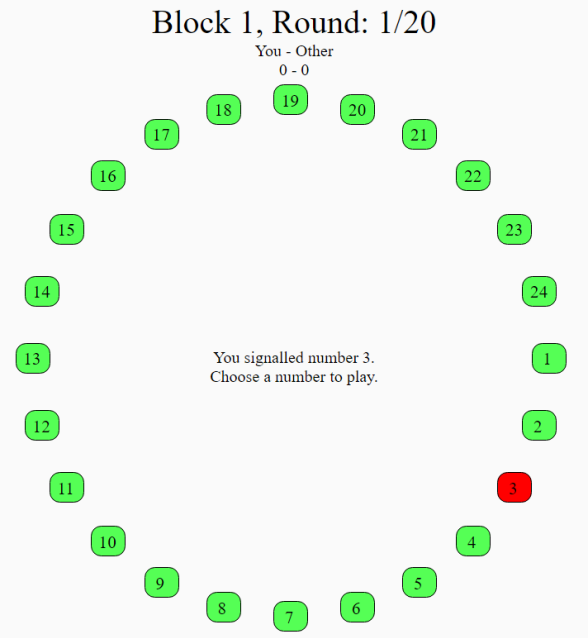

Figure 2.1: The user interface used in the experiment for the Mod-Signal Game

also facilitates cooperation. The Mod-Signal Game is similar to a repeated Prisoner's Dilemma, where cooperation gives $\frac{1}{2}$ points per round. This is because in a fully cooperative Mod-Signal game the signaler plays the number that they signalled. This allows the responder to gain a point in that round, each player is the responder every other round meaning that they get a point every 2 rounds. Similar to the Prisoner's Dilemma, cooperation can be exploited which would result up to 1 point per round in the Mod-Signal Game. Playing the game randomly gives $\frac{1}{24}$ points per round.

2.1 Theory of Mind

In the Mod-Signal Game players will most likely make use of a concept called Theory of Mind (Premack & Woodruff, 1978) Theory of Mind (ToM) is the ability to reason about the mental states and contents of other people. An example of second-order ToM in the context of the Mod-Signal Game goes as follows, if Alice and Bob were to play the game cooperatively and Alice (signaler) would play and signal number 5, Bob (responder) would then play number 6. However, Alice could choose to deceive Bob by playing number 7, but if Bob believes that Alice is going to deceive him he will play number 8. Then if Alice believes that Bob believes that Alice will deceive Bob, Alice will play number 9, etc. This reasoning process using ToM can go on recursively and can lead to interesting behaviour patterns. The large action space of the Mod Game allows to interpret the Theory of Mind order of the participants more easily compared to rock-paper-scissors for example. In rock-paper-scissors there is only an action space of 3, so it is hard to tell the order of Theory of Mind that is used.

3 Methodology

3.1 Participants

A total of 21 people played the Mod-Signal Game in this experiment. The participants were required to have a proper understanding of English to ensure that they understood the instructions of the experiment. The participants' ages ranged from 16 to 25 (M = 20.3). There were 9 female participants and 12 male participants. All participants were students that were enrolled in a study programme that requires the use of computers ensuring that they would be able to grasp the virtual environment of the experiment. The participants did not have any experience in the Mod-Signal Game prior to the experiment. However, there were participants that were knowledgeable about related topics such as the Tacit Communication Game and Theory of Mind (ToM).

3.2 Experiment setup

The experiment was conducted with groups of 3 participants. Each session consisted of 3 blocks, where in each block two of the participants played the Mod-Signal Game with each other and the other participant played with a simulated agent. The games in a block lasted 20 rounds so that both players signalled 10 times. To prevent the participants from thinking that they are playing with a simulated agent, a confederate was also present at the experiment pretending to be the fourth participant. Throughout the experiment the participants are also asked to fill in questionnaires. One session lasted approximately 30 to 45 minutes.

3.2.1 Before the start of the experiment

The experiment took place in a room where the participants are separated by barriers. Firstly, this was done to minimise the number of ways that the participants could communicate with each other as they should only communicate through the signalling phase in the game. Secondly, it also hid the confederate from the participants so that they could not see that the confederate was not actually playing the game. Each computer was assigned a number from 1 to 4, which determined the participant number of the participant that sat at that computer. The participant numbers were used to determine and keep track of which participants played with each other in each block. The participants were aware of their participant number as this is also on the questionnaires, but they did not know when they played with which participant. This was to avoid their strategies being influenced by biases caused by participants knowing each other. The confederate was always assigned number 4 as this computer contained the fake setup, which only showed the same introduction screen to fool the other participants, but was not used to play with the other participants. Instead, the agent that played with the participants was setup on another computer which was not visible to the participants.
When a participant arrived the experimenter assigned them to a computer, making sure that they did not sit at the fake setup. The participant was asked to fill in an informed consent form that explained the goals of this project and requested their consent to use the data that would be collected from them. Once all participants signed the informed consent form the experiment started.

3.2.2 Experiment process

At the start of the experiment the participants were instructed to put on headphones and read the instructions on the starting screen. The headphones were used to make mouse clicks less audible to the participants, as mouse clicks can be used by the participants to figure out who they are playing with or possibly figure out that they are playing with the agent. This is done by determining whether the mouse clicks that they hear, synchronise with the game updating. So, it is important to make them less audible. Then the participants were asked to read the starting screen. This screen contains information about the rules of the Mod-Signal Game and further instructions on what they are going to do in the experiment. After reading the instructions they proceed to the next screen where they play 4 trial rounds of the Mod-Signal Game, to get a better understanding of the user interface and the game. The trial rounds were played against a predictable agent that only played cooperatively, where the agent started as the signaler in round 1. Once they were done with the trial rounds they could start with block 1, where 2 of the participants would play with each other and one participant would play with the agent. After playing 20 rounds the block ended and the participants were led to an intermediary screen which asked the participant to fill in a questionnaire that was provided to them. This questionnaire asked about the strategy that they used in the last block and what they thought of the strategy of the player they played with. The process for block 2 and 3 are identical to block 1 except that the players play with a different participant (or agent) in each block. At the end of the session, each participant had played with each other and also had played with the agent once. Additionally, to prevent order effects, every participant started as the signaler once and started as the responder once, when playing with humans. All participants started as the responder when playing with the agent. After block 3 had ended and the participants had filled in the questionnaire of block 3, they were asked to fill in a final questionnaire that contained some general questions such as asking for their age, sex and what they are studying. This questionnaire also revealed to the participants that some of them had played with an agent in some of the blocks. They were then asked in which blocks they thought they played with a human. The general questionnaire also contained a p-beauty contest (Nagel, 1995), this is a contest where the participants were asked to pick a number from 1 to 100, which they thought would be the average of the numbers picked by all participants multiplied by 2 and divided by 3. This contest is used to show which ToM order the participants have. Once all participants had filled in the last questionnaire, the session ended. The informed consent form and questionnaires can be

found in appendix D.

Throughout the session, to make sure that the participants were tricked into thinking that the confederate was also a participant and that they are not playing with an agent, the confederate pretended to play the Mod-Signal Game by clicking the mouse periodically and also pretended to fill in the questionnaires once a block ended. To make sure this process happens smoothly, the experimenter messaged the confederate to start clicking when the participant that the agent is playing with was ready. Once a block ended, the confederate was notified so that they knew when to stop clicking and pretend to fill in a questionnaire.

3.2.3 Implementation

The Mod-Signal Game implementation from Bos (2022) is used as a basis for this experiment. This implementation modifies the JavaScript implementation of the Mod Game which was created by Veltman et al. (2019). The modifications made by Bos add three new agent types and the signalling phase of the Mod-Signal Game, which allows one of the players to signal the number that they might choose to the other player before every round. The implementation by Bos does not allow for playing with other people. In this experiment, modifications are made to the implementation to make it compatible for online play with 2 players. There is also one modification that was done to make the user interface easier to follow. At the end of every round the numbers that were chosen in that round turn blue until they click the continue button. This was done to make it easier for the players to observe the results of the round and take in the other player's strategy.
In the implementation the in-game text and the questionnaires were specifically written in a way that avoids priming the participants of either competitive or cooperative behaviour.

3.2.4 Agent behaviour

The experiment conducted by Bos (2022) studies how the behaviour of participant differs when playing with several agent implementations that show different behaviour. In one of these conditions, the participants were tricked into thinking that they played against a human, while they were actually playing against an agent. The agent implementation used in this condition is also used in my experiment with some modifications, as this agent showed the most human-like behaviour. In the experiment by Bos (2022) the agent cheats if it has not gained a point for three consecutive rounds, by waiting for the choice of the participant and choosing one higher than that number. In the implementation of my experiment this is not the case. The agent observes the same information as the other participants and is therefore not able to cheat.
In this experiment, as well as in the experiment by Bos (2022), the agent always cooperates in the first round by choosing the number it has signalled. In the other rounds the agent cooperates based on a 'cooperation probability'. When cooperation occurs, the agent plays honestly or trusting when signaling or responding respectively. Otherwise the agent tries to deceive the other player. The cooperation probability is described by formula 3.1, where C is the number of rounds that were played cooperatively in the last 3 rounds and N is the current round number. In the experiment by Bos (2022) only the last 2 rounds were considered for C.

$$P_{coop} = \frac{C}{\min(3, \mathrm{N})} \tag{3.1}$$

The behaviour of the agent when playing cooperatively is very simple. The agent plays the number equal to the current signal when the agent is signaling and plays the number equal to the current signal + 1 when the agent is responding that round. If the agent is trying to deceive the player the behaviour differs. The agent looks at what the cooperative choice was last round for the player and the number the player actually chose. This cooperative choice is the signal if the participant was the signaler or signal + 1 if the participant was the responder that round. The absolute difference between the actual choice and the cooperative choice is taken and 1 is added to get the 'change rate' defined by Bos (2022). The agent then looks at the signal of the current round to calculates its choice by adding the change rate to the current cooperative choice, modulo 24.
The agent itself in the experiment is implemented as if it were another participant. When the agent is signalling it signals a random number from 1 to 24. When the agent is choosing a number it makes

its decision by looking at the information it has on the current round and from the last 3 rounds, it then chooses the number by simulating a click. It is also important that the agent behaves similarly to a human, therefore the time it takes for the agent to make its decision is a random number drawn from an uniform distribution between 1000 and 5000 milliseconds. This reaction time was found to be similar to human reaction times in a separate pilot session of this experiment as well as in the experiment conducted by Bos (2022). The confederate is also told to click periodically approximately within this time interval during the experiment to make it more believable that the participants are not playing with an agent.

3.2.5 Measured data

For every game that is played throughout the experiment the signal and choices made by the players in every round are recorded. The time it takes for the player to choose a number after signaling or receiving a signal is also recorded. The honesty and trust levels of the participants can then be derived by looking at the signal that round and the chosen number of the participant. Participants are considered to play honestly if they play the same action as their own signal. Conversely, participants are considered to exhibit trust whenever they play the action that is one higher (modulo 24) than the signal of their co-player. With this data, several comparisons can be made by looking at the differences between the participant data when playing with other participants and playing with the agent. For example, how honest and trusting the participants are between playing with a human or the agent. In the final questionnaire, one of the questions asks the participant to give a percentage on how confident they were that they played with a human in each block. This data is also interesting to look at to see if participants' strategies differ whether they thought they were playing with a human or the agent.

4 Results

4.1 Surface level results

4.1.1 Signal and choice distribution

In the signalling phase the signaler is asked to signal a number to the responder. There is no strategy involved during this phase, as the number that is signalled does not affect the responder's strategy nor does it express the signaler's strategy. The signal is only an indication of what the signaler might choose. Therefore, you would expect the distribution of the signals to be approximately uniform. However, in figure 4.1 it can be seen that there is a clear preference for the numbers 1, 7 and 13.

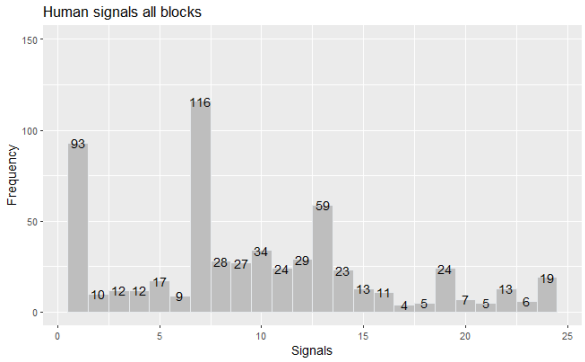

Figure 4.1: Signal distribution of human signalers over all blocks

At first glance there seems to be no correlation between these numbers, but when looking at the user interface in figure 4.2 it can be seen that these numbers are exactly on one of the cardinal points of the circle. Number 19 is also exactly north on the circle, however does not seem to have a high preference unlike the numbers on the other cardinal points. This is because there also seems to be a preference to numbers that are on the bottom half of the circle (2-12) compared to the numbers that are on the top half of the circle (14-24) by looking at the graph. Number 19 is still a peak when the frequency of the numbers surrounding are taken into account, this peak is due to it being exactly north on the circle.

The distribution of the choices (figure 4.3) is more evenly distributed compared to the signal distribution. The peak at number 7 from the signal distribution can still be seen in the choice distribu-

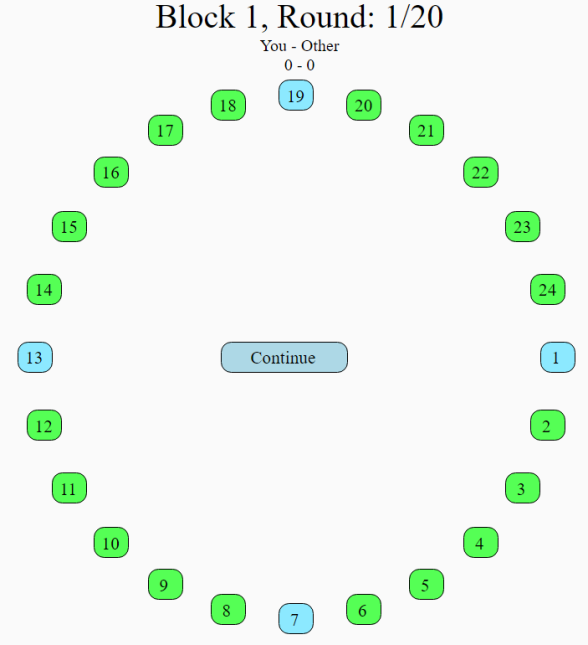

Figure 4.2: User interface with the cardinal points of the circle highlighted

tion. Although number 1 and 13 are still the second and third most occurring choices, they do not peak as highly above the other choices as compared to the signal distribution. The more balanced distribution is expected, because the signals are not always followed by an honest choice. This can be seen by looking at some of the dishonest choices for the most occurring signals in figure 4.1. The dishonest choices are most commonly signal +2,+4 etc. To explain this consider the following example, Alice and Bob are playing the Mod-Signal Game and Alice is currently the signaler. If Alice is to play honestly she would play signal + 0 so that Bob can play signal + 1 to get a point. However, if she wants to gain a point herself, she can be dishonest and play signal + 2. Bob could however choose to not trust Alice and expect her to play dishonestly with signal + 2, so Bob plays signal + 3 in return. For Alice to gain a point when Bob is distrusting, Alice would have to play signal + 4. This goes on recursively, where the signaler has to play signal + 6, + 8, + 10 etc. and the responder has to play + 7, + 9, + 11 etc. Therefore the dishonest choices for the signaler are signal +2, +4 etc.
In the graph it can be seen that number 1 has only been chosen 40 times compared to the 93 times it was signalled. This means that at least 53 times, a signal of 1 was followed by a dishonest choice. Some of the dishonest choices for 1 are number 3 and 5 for example. These numbers have a higher frequency in the choice distribution graph compared to the signal distribution due to the dishonest choices from signalling 1. The dishonest choices therefore spread out the choice distribution more evenly.

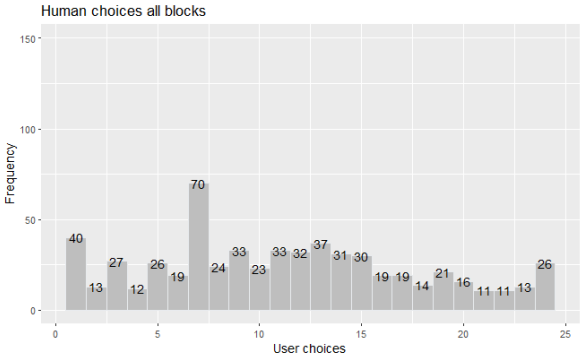

Figure 4.3: Choice distribution of human signalers over all blocks

4.1.2 Distribution of choice - signal difference

In figure 4.4 the distribution of the choice - signal differences is shown when playing with a human, where choice means the signaler's choice. The choice - signal difference indicates whether the signaler played honestly or dishonestly. The graph shows a similar distribution as the graph of the choice - signal differences when playing with the agent (appendix C). This indicates that the signaling and choosing behaviour of the participants seems to be the same when playing with a human and an agent. In the graph a difference of 0 indicates the the signaler played honestly, any other choice is interpreted as the signaler playing dishonestly. After the peak at 0, there are also peaks that happen at a difference of 2,4,6 etc. These peaks are to be expected (see 4.1.1). This also relates to ToM, showing increasing orders of ToM with larger distances between the signal and the chosen number. The graph also shows a small peak at 12, this could be due to location of the number being exactly on the opposite side of the circle.

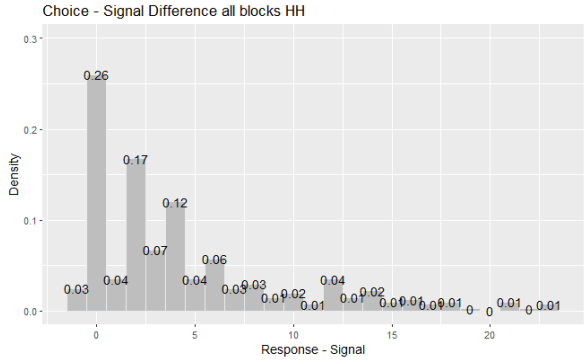

Figure 4.4: Distribution of choice - signal difference when playing with a human (rounded to two digits)

What is also interesting is that the signaler also plays by choosing a number that is the signal + an odd number. This is not a logical strategy. As the signaler you either expect the responder to trust you by playing signal + 1 or to not trust you by playing signal + an odd number, except for 1. If the signaler chooses signal + an odd number, there is a chance that the signaler and responder choose the same number. This results in none of the players getting a point. Therefore it is only logical for the signaler to choose a number whose distance to the signal is even. These uneven differences are most likely due to participants clicking on the wrong number, misunderstanding of the game, miscalculations or even a response to the other player playing illogically. The graph differences range from -1 to 23, this was specifically chosen because playing signal - 1 was also a strategy that was found in the previous bachelor's thesis by Bos (2022). This strategy happens when participants think it is up to the responder to give the point to the signaler by playing - 1 rather than the responder taking a point by playing signal + 1 which is the most common understanding of the game.
The graph for the choice - signal when playing with the agent can be found in appendix C, as well as the graphs for the response - signal difference.

4.2 Analysis

4.2.1 Honesty and trust

The honesty level of a participant is measured by taking the percentage of the number of times the participant played honestly. The concept of honesty only applies when the participant is the signaler in the round. A participant plays honestly when the number that they chose to play is the same as the number that they signalled that round. It is interesting to see whether participants are more honest towards the other participants or the agent. In figure 4.5 the honesty levels of the participants are shown, comparing between when the participants played with another participant and with the agent. Every participant signalled 20 rounds to other participants and 10 rounds to the agent. The figure shows that there seems to be little to no difference between the honesty levels of the participants when signalling to a human (HH) and signalling to the agent (HA). A comparison of the HH condition ($M = 26$, $SD = 24.64$) and the HA condition ($M = 28.57$, $SD = 22.64$) is done using a two-sample chi-squared test. The test shows that there is no significant difference between the honesty levels of participants when signalling to a human or signalling to the agent, $\chi^2(1) = 0.88$, $p = 0.3476$. The trust

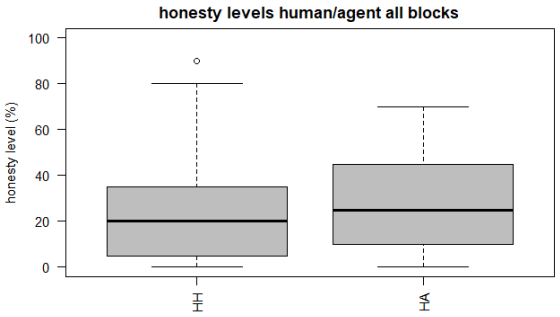

Figure 4.5: Honesty levels of participants when playing with a human (HH) and when playing with the agent (HA)

levels of the participants can be used to determine whether the participants are more trusting of the other participants or the agent. The trust level of a participant is measured by taking the percentage of the number of times the participant trusted the signaler. The concept of trust only applies when the participant is the responder in the round. A participant shows trust when they choose to play a number that is exactly one higher than the number that the signaler signalled. In figure 4.6 the trust levels of the participants are shown when playing

with a human and when playing with the agent. Similarly to the honesty levels of the participants, there also seems to be almost no difference in trust levels between the conditions. When comparing the HH condition (M = 29.5, SD = 22.82) and the HA condition (M = 28.57, SD = 28.16) with a two-sample chi-squared test. The test shows that there is no significant difference between the trust levels of the participants when playing with a human or playing with the agent, $\chi^2(1) = 0.64$, $p = 0.4233$.

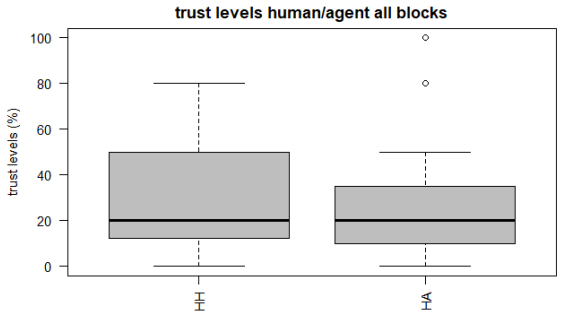

Figure 4.6: Trust levels of participants when playing with a human (HH) and when playing with the agent (HA)

The mean honesty and trust levels are never below 30%, this indicates that the participants played more competitively than cooperatively throughout the experiment.

4.3 Scores

The mean scores of the participants give an indication of whether the participants played cooperatively, competitively or randomly. If the mean scores are closer to 10, the participants played more cooperatively. If the participants played randomly the mean would be closer to 0.83 ($\frac{1}{24} \times 20$). This is a game-theoretic rational strategy as the Mod Game has a mixed-strategy Nash equilibrium (Veltman et al., 2017). The mean scores of the participants (M = 3.05, SD = 2.16) shows that the participants did not play randomly. This is to be expected as people do not play rationally and are generally bad at randomizing. It also shows that the participants seemed to play more competitively, which aligns with the results of the honesty and trust levels of the participants from section 4.2.1.

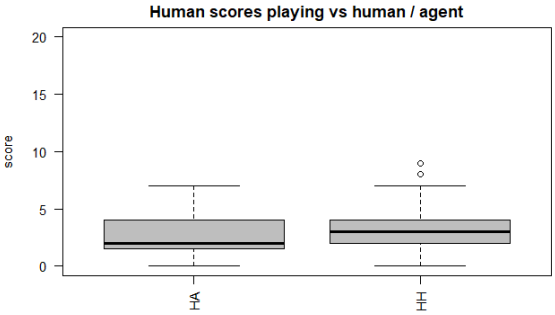

Figure 4.7: Participant scores when playing with a human (HH) and when playing with the agent (HA)

In figure 4.7 the scores of the participants are shown for when they played with a human and with the agent. The graph shows that there seems to be no difference in the scores when playing with a human (HH) or with the agent (HA). A comparison of the HH condition (M = 3.18, SD = 2.29) and the HA condition (M = 2.8, SD = 1.91) was made with a two-sample chi-squared test. The results further confirm the observation that there is no significant difference between the scores of the participants when playing with a human or playing with the agent, $\chi^2(1) = 0.59$ and $p = 0.4434$.

The participant scores also show that there seems to be no order effect for the blocks. The mean scores in each block seem to be approximately the same. It also shows that there is no order effect on the starting order of the participants (signaler/responder assignment in round 1). The graphs that show these results can be found in appendix A, which contains graphs related to the scores of the participants.

4.4 Questionnaires

4.4.1 Confidence levels playing with a human

In the final questionnaire, the participants are asked how confident they are that they played with a human in each block. This data shows how well the participants are able to differentiate human behaviour from the agent's behaviour. Figure 4.8 shows how confident the participants were that they played with a human. The confidence levels were separated by the times that they played with

a human and the times they played with the agent. By looking at the graph, we can see that there is a slight difference between the two conditions, where the human condition has a higher confidence level. The data for the human condition is however slightly skewed. Some participants had used information outside of the virtual environment of the game to answer the question. A couple participants had said in the questionnaire that they closely listened to the mouse clicks of the other participants and were able to tell that they played with a human, because the mouse clicks would immediately be followed by their game screen updating. When this happens for 20 rounds it can be quite easily determined that you are playing with a human. This resulted in a couple of answers being a 100% confident without basing it on the strategy of the player they played with. Even with the skewed data, the graph still shows that the participants had a hard time telling when they played with the agent. Comparing the human condition (M = 65.79, SD = 29.42) and the agent condition (M = 64.37, SD = 26.40) with a two-sampled paired t-test shows that there is no significant difference between the confidence levels of the participant when playing with a human and when playing with the agent, $t(18) = 0.18$ and $p = 0.8564$.
These results could explain why there is no difference in the honesty levels or trust levels of the participants when playing with a human or the agent. The participants already had difficulty with telling in which blocks they thought that they were playing with an agent, but this was asked after all the blocks were played. It is most likely the case that the participants did not think about whether they were playing with an agent or not during the experiment, as they expected to only play with humans due to the experiment setup. Only after seeing the final questionnaire were they asked to reason about whether they played with the agent or not.

Figure 4.9 shows the confidence levels of the participants that they were playing with a human for each block of the experiment. From the graph it can be observed that in block 2 the participants considered their co-player significantly less human than in the rest of the blocks. This is very interesting, because in the experiment all of the blocks have an identical process. In each block two players play with each other and one player plays with the agent. The only difference between the blocks is that you play with a different player each block. An ANOVA test was performed to show that there is indeed a significant difference between the confidence levels of the participants in block 1 (M = 72.89, SD = 25.73), block 2 (M = 48.68, SD = 27.78) and block 3 (M = 74.37, SD = 24.59), $F(2) = 5.816$ and $p < 0.01$.
The order effect for the confidence levels of the participants is unusual as the process of each block is identical. In the questionnaires some participants mentioned that they tried to experiment with new strategies in block 2 compared to their strategy in block 1. It could be the case that the experimented strategies were not as successful and therefore also seemed not human-like to their co-players. Another reason Block 1 and 3 are remembered more clearly than block 2 could be due to the primacy and recency effect. Block 1 happens first and is put in the long-term memory (primacy effect), while block 3 happens last and is put in the short-term memory (recency effect). Block 2 is therefore less memorable and this could lead to participants having a lower confidence level for that block.

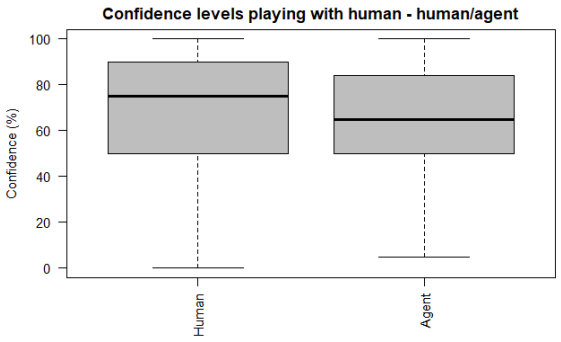

Figure 4.8: Confidence levels of the participants for the blocks that they played with a human and for the blocks that they played with the agent

4.4.2 End of block ratings

In the questionnaires given at the end of every block the participants were asked to rate the other player's competitiveness and cooperativeness in that block on a scale from 1 to 5. The competitiveness (M = 3.93, SD = 1.05) and cooperativeness (M = 2.51, SD = 1.26) of the participants corre-

spond to the mean honesty and trust levels that were found in section 4.2.1. The competence ratings of the participants were also measured. Similar to the confidence levels mentioned in section 4.4.1, the competence ratings show the same order effect, where the competence ratings given by the participants in block 2 (M = 3.55, SD = 1.28) are a lot lower than the competence ratings in block 1 (M = 3.95, SD = 1.07) and block 3 (M = 4, SD = 0.97). The lower competence levels in block 2 could explain the lower confidence levels that were found, however then the question still arises why the competence level was a lot lower in block 2.
The participants were also asked whether they would like to play again with a participant, the answers were stored in discrete values of -1, 0 and 1 meaning 'no', 'neutral' and 'yes' respectively. The play again ratings (M = 0.25, SD = 0.96) that the participants gave to their co-players show that despite the low honesty and trust levels, the participants are still willing to play again with their co-players. The P-beauty contest in the final questionnaire unfortunately did not result in any interesting results. The questionnaires used in the experiment can all be found in appendix D.

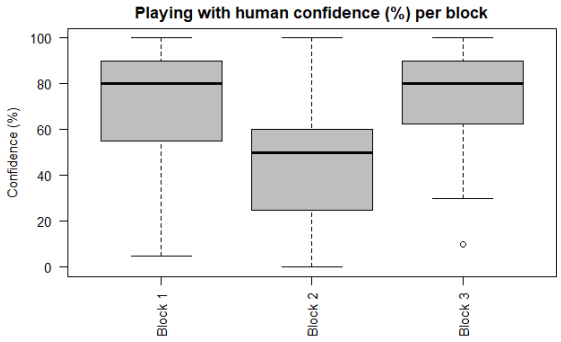

Figure 4.9: Confidence levels of the participants that they played with a human for each block

4.5 Reaction times

The reaction times of the participants were also analysed, a two-paired t-test found that the signalers reacted faster when playing honestly (M = 8.043579, SD = 0.71) compared to playing dishonestly (M = 8.562491, SD = 0.39), $t = 3.82$ and $p = 0.001$. In figure 4.10. The reaction times for the

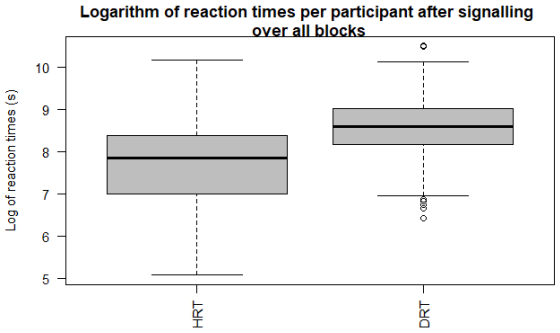

Figure 4.10: Logarithmic reaction times of signalers when playing honestly and dishonestly

responders showed similar results but to a lesser extent. The graph for the responders can be found in appendix B.

5 Discussion & Conclusion

The results bring to light some interesting patterns in the behaviour of the participants. In section 4.1 the signal distribution shows that the participants do have a preference for certain numbers when signalling. Although, there is no strategy in the signalling phase as it is a simple task of picking a random number to signal to the co-player. There are high preferences for the numbers that lie exactly on the cardinal points of the circle (1, 7, 13 and 19). There also seems to be a preference for the numbers that are on the lower half of the circle compared to the numbers on the top of half of the circle. This also explains why number 19 is picked less than the other cardinal points. In the choice distribution the preference for these numbers is not as noticeable. This is to be expected as not every signal will be followed by an honest choice. The dishonest choices made by participants cause the choice distribution more evenly distributed compared to the signal distribution.
The choice - signal difference and response - signal difference graphs show that the participants utilised the concept of ToM in their decision making. The graphs can also be used to get an indication of which order of ToM reasoning is used by looking at the distance of the number and the signal.

The results analysis shows that there seems to be no difference between the behaviour of the participants when playing with a human or the agent. It also showed that our participants seemed to mostly play competitive as observed from the user scores and the honesty and trust levels. The mean honesty and trust levels were always below 30% which indicates that the participants played competitively, this also validates some of the findings in Bos (2022). The low honesty and trust levels also relate to the scores of the participants which ranged from 1-5. The participant scores give an indication of the strategy used while playing, where a score around 10 indicates cooperative play, while lower scores relate more towards competitive behaviour. Randomly choosing would lead to a mean score of 0.83 ($\frac{1}{24} \times 20$) per game.
From the questionnaire data it was found that the participants had a hard time distinguishing in which blocks they played with a human. As the confidence levels were approximately the same as seen in figure 4.8. This could also explain why there are no behavioral differences when playing with the agent or a human. As the participants most likely were not thinking about whether they were playing with an agent or not during the experiment. Only after the final questionnaire were they actively thinking about whether they played with an agent in each block.

5.1 Conclusion

The results show that humans tend to play competitively despite there being a higher payoff of playing cooperatively. The participants scored 3.05 on average while fully cooperative play results in a score of 10. It is interesting that the participants played so competitively as the experiment was carefully designed to not prime any cooperative or competitive behaviour. The results also found cooperative play to be significantly faster than competitive play, so time also did not seem to play a role in the behaviour of the participants. From the questionnaire data and talking to the participants after the experiment, it was found that a lot of participants seem to have considered the Mod-Signal Game as a competitive game as they used a lot of terminology that is generally used within competitive play, such as 'won', 'lost' and 'opponent', while this is not mentioned anywhere in the instructions of the experiment. The participants also mentioned that they found cooperative play to be boring, while they enjoyed the thought process that went into competitive play.
The results also show that there are no significant differences in the behaviour of the participants when playing with a human and when unknowingly playing with an agent. Several measured variables were taken into account, but no interesting differences were found in the behaviour. Despite these results, it can not be concluded that the behaviour of the participants does not differ when playing with a human and when unknowingly playing with an agent. This is because different behaviour patterns can still lead to the same results. To be able to confirm that behaviour does not differ, extensive analysis needs to be performed on the strategies used by the participants when playing with a human and playing with the agent.
This thesis has also introduced human-human play in the Mod-Signal Game, which has not been done before. The implementation used in this thesis can be used in future research to study ToM in mixed-motive settings. The Mod-Signal Game should especially be used for this as the signalling phase and the large action space make it easier to interpret strategies and the Theory of Mind orders of the participants.

5.2 Improvements

Over the course of conducting this experiment several flaws have occurred that might have influenced the results. The first flaw that I made is that in the first 3 sessions I forgot to tell the participants to put on their headphones. This could have influenced the results, as the mouse clicks of the participants were more audible. Participants can use the mouse clicks to determine whether they are playing with a human or not, by determining whether the clicks are followed by their game screen updating. Although the participants had also used this strategy in the sessions that the the headphones were put on, it could have been harder to do with the headphones on as the participants might have also felt more discouraged to listen to the mouse clicks as the headphones give an indication that they are not supposed to hear anything.
In general it would be better to have a setup where the mouse clicks are not audible to the partici-

pants. This would prevent the human confidence level question from being answered by only listening to mouse clicks.
The second flaw that happened is that setup of the experiment was not consistent over the course of conducting the experiment. Due to unfortunate circumstances I was not able to conduct all sessions in the same room. The sessions were conducted in two different rooms. The layout of the rooms were slightly different and so the experiment setup also needed to change slightly.
The last flaw is that in one of the experiment sessions, the agent malfunctioned for one of the participants. Therefore, the data for this participant had to be partially excluded from the results analysis.

5.3 Future research

In this project I did not have the time to do extensive analysis on the behaviour of the participants. Although the statistical analysis showed that there are no significant differences between the human and agent conditions for e.g. honesty and trust. That does not necessarily mean that the behaviour of the participants when playing with a human is similar to their behaviour when playing with an agent. For future research it would be interesting to extract the different strategies that were used by the participants. It is interesting to look at the strategies that humans use when playing with each other as well as to look at how these strategies compare to the strategies they use when playing with an agent. Additionally, it would be interesting to measure the ToM orders that were used by the participants and determine the order of ToM that humans have when playing with each other as well as comparing these ToM orders to the ToM orders they have when playing with an agent.
The results showed that the participants in this experiment preferred to play competitively despite the higher payoff with cooperative play as well as the detailed experiment design to prevent any priming of cooperative or competitive behaviour. It seemed that the participants played competitively due to them not being completely aware of the higher payoff in cooperative play as well as competitive play being more enjoyable. In future research cooperative play could be primed to the participants, we could then see see whether the participants still show similar competitiveness and honesty and trust levels as in this experiment. If this does not seem to work, alternatively participants could be primed with cooperative play by basing their payments on the scores that they get in the games, as in this experiment participants were not rewarded for the scores they achieved in the games.
Research on the Mod-Signal Game has already been performed within a human-agent context by Bos (2022). In this thesis, human-human play in the Mod-Signal Game is introduced for the first time to study the behaviour of humans when playing with other humans. This new implementation can be used as a tool in future research to study ToM in mixed-motive settings. It can be used and altered to study ToM in many different ways. An example would be to have 2 participants play against different agents in the first block. The agents differ in which ToM order they utilise in their decision making, where one agent uses a lower ToM order and one uses a higher order of ToM. It is interesting to see whether playing against different ToM order agents affects their performance when playing with each other and to see how they adapt their strategy throughout the experiment.

6 Acknowledgements

First, I would like to thank my supervisors mr. J.D. Top and dr. H.A. De Weerd for guiding me through this project. They assisted me a lot in the preparation of this experiment and in writing this thesis.
Second, I would also like to thank Tim Eckhardt and Denny Verbeek, who helped me by acting as the confederate in the experiment.
Last, I would like to thank the University of Groningen for providing me with funds to compensate the participants as well as providing me with a room and supplies that I needed to conduct the experiment.

References

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., ... Seli, P. (2023). Humans versus ai: whether and why

we prefer human-created compared to ai-created artwork. *Cognitive Research: Principles and Implications*, *8*(1), 1–22.

Bos, M. (2022). *Cooperation and exploitation in a modified mod game* (Bachelor's Thesis).

Frey, S., & Goldstone, R. L. (2013). Cyclic game dynamics driven by iterated reasoning. *PloS one*, *8*(2), e56416.

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., & Gerjets, P. (2021). Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text? In *International conference on human-computer interaction* (pp. 520–527).

McGloin, R., Hull, K. S., & Christensen, J. L. (2016). The social implications of casual online gaming: Examining the effects of competitive setting and performance outcome on player perceptions. *Computers in Human Behavior*, *59*, 173–181.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313–1326.

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

Turing, A. M. (2009). *Computing machinery and intelligence*. Springer.

Veltman, K., de Weerd, H., & Verbrugge, R. (2017). Socially smart software agents entice people to use higher-order theory of mind in the mod game. In *The 29th benelux conference on artificial intelligence* (pp. 253–267).

Veltman, K., de Weerd, H., & Verbrugge, R. (2019). Training the use of theory of mind using artificial agents. *Journal on Multimodal User Interfaces*, *13*(1), 3–18.

A Participant scores

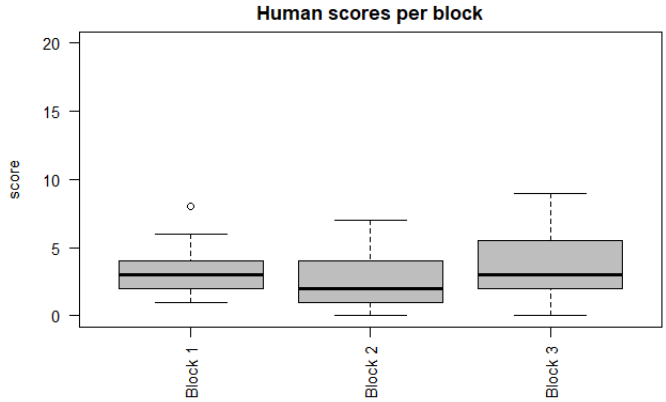

Figure A.1: Participant scores per block

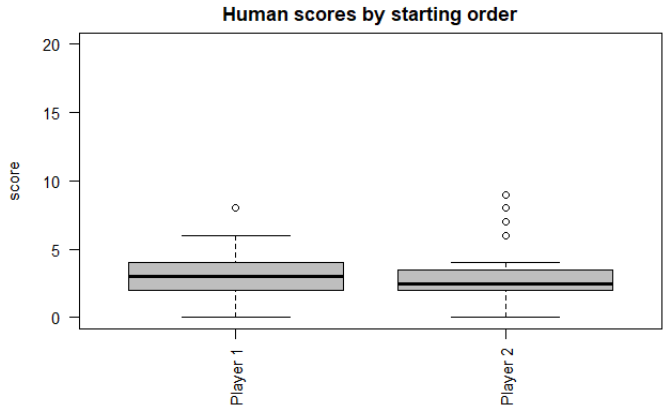

Figure A.2: Participant scores based on which player started as the signaler in round 1 (player 1 is the signaler in round1)

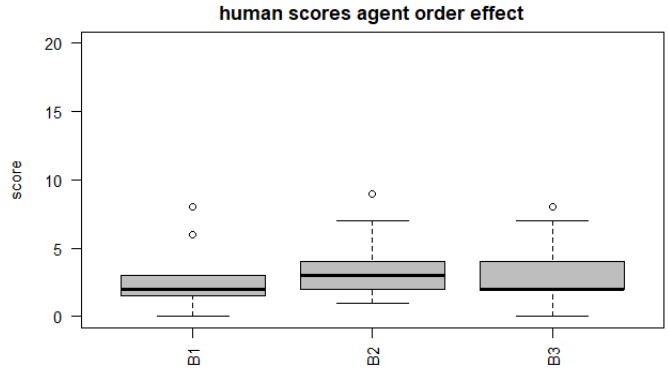

Figure A.3: Participant scores based on in which round they played with the agent

B Reaction times

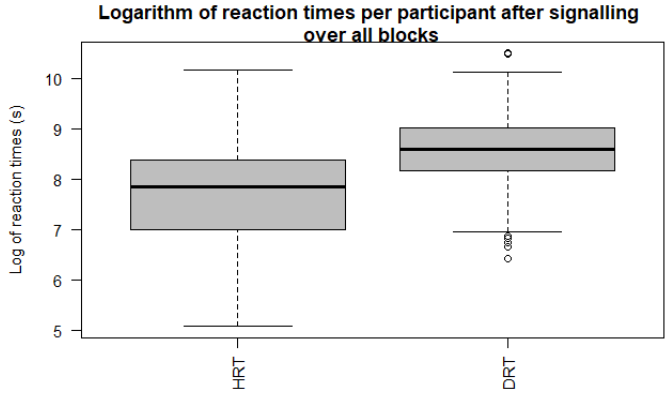

Figure B.1: Reaction times of responders when trusting and not trusting the signal

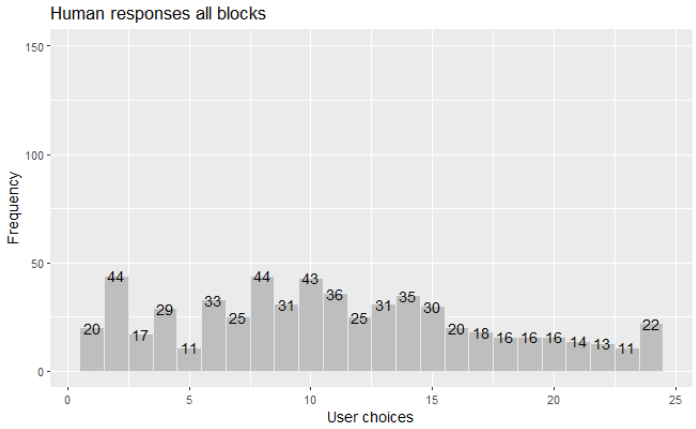

Figure C.1: Response distribution of human responders over all blocks

C Signal and choice distribution

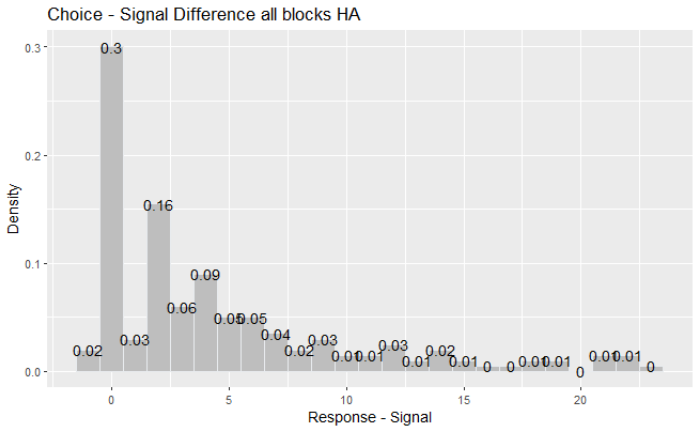

Figure C.2: Distribution of choice - signal difference when playing with the agent (rounded to two digits)

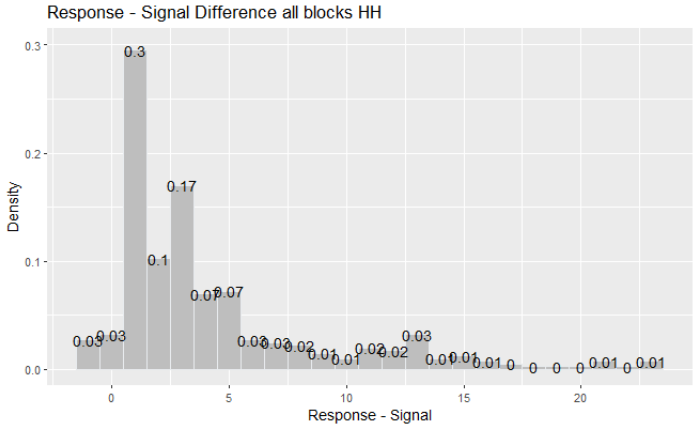

Figure C.3: Distribution of response - signal difference when playing with a human (rounded to two digits)

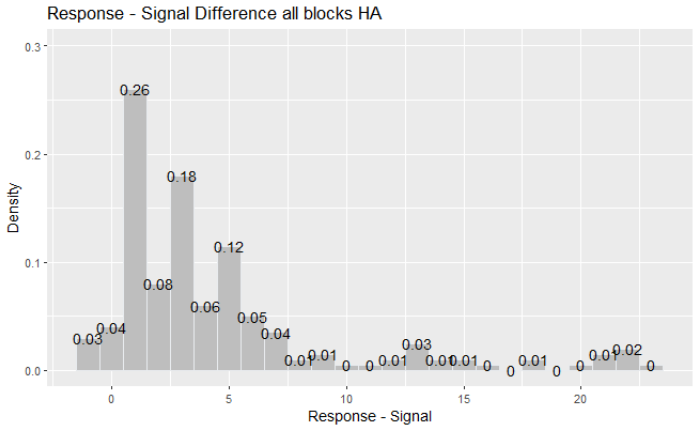

Figure C.4: Distribution of response - signal difference when playing with the agent (rounded to two digits)

D Documents

INFORMED CONSENT

I
(name participant)

hereby consent to be a participant in the current research performed by
(name researcher)
Abi Raveenthiran

My email address is:

This is required for your payment. If you already have an account for the RUG payment system, write down the email that you used to sign up for it. Otherwise your student email is highly preferred, but a private email will also work.

I have agreed to take part in the study entitled
Human behaviour in the Mod-Signal Game
and I understand that my participation is entirely voluntary. I understand that my responses will be kept strictly confidential and anonymous. I have the option to withdraw from this study at any time, without penalty, and I also have the right to request that my responses not be used.

The following points have been explained to me:

1. The goal of this study is
to learn about human behaviour in a simple two-player game
Participation in this study should help advance our understanding of
human strategic behaviour

2. I shall be asked to
play 20 rounds of the Mod-Signal Game 3 times and fill in questionnaires

3. The current study will last approximately **30-45** minutes. At the end of the study, the researcher will explain to me in more detail what the research was about.

4. My responses will be treated confidentially and my anonymity will be ensured. Hence, my responses cannot be identifiable and linked back to me as an individual.

5. The researcher will answer any questions I might have regarding this research, now or later in the course of the study.

Date:	Signature researcher:

Date:	Signature participant:

Figure D.1: Informed consent form

Questionnaire Block 1

Question 1
Briefly describe the strategy that you used in this block.

Question 2
What did you think of the other player's strategy? Briefly describe their strategy as well.

Question 3
How would you rate the cooperativeness of the other player?

1 2 3 4 5
uncooperative neutral cooperative

Question 4
How would you rate the competitiveness of the other player?

1 2 3 4 5
uncompetitive neutral competitive

Question 5
How competent was the other player at the Mod-Signal Game?

1 2 3 4 5
incompetent neutral competent

Question 6
Would you want to play with this player again? Explain briefly.

Figure D.2: Questionnaire asked after every block

General Questionnaire

Question 1

What is your age, sex and what are you studying?

Question 2

Were the instructions provided at the start sufficient enough to understand the game? If not, were the trials enough to get a proper understanding of the game?

Question 3

What was your mood like today?

Question 4

Did you have any prior knowledge of the Mod Game before the start of the experiment?

Don't forget to answer the questions on the other side of the paper!

Figure D.3: Questionnaire asked at the end of the experiment part 1)

Question 5
Did you know any of the other participants that you played with in the session before the start of the experiment?

Question 6
Some of our participants have played the Mod-Game with a computer partner rather than a human partner in some of the blocks.
For each block give a percentage on how confident you are that you played with a human partner in that block and briefly explain why. (e.g. Block 1: 100%, you are completely certain that you played with a human partner)
Block 1:

Block 2:

Block 3:

P-beauty contest
All participants will do this contest. The winner gets a Tony's Chocolonely chocolate bar of their choice! Mark the number that you think will be the average of the numbers chosen by all participants, divided by 3 and multiplied by 2.(take ⅔ of the number you think will be the average of the numbers chosen by all participants, e.g. if you think 90 will be the average of the numbers chosen by all participants, you mark 60)

Figure D.4: Questionnaire asked at the end of the experiment part 2)