Cooperation, Deception and Theory of Mind in a Cyclic Game with Inter-Player Signalling

Jakob Dirk Top*[1], Harmen de Weerd[1], Abisharan Raveenthiran[1], Catholijn Jonker[23], Rineke Verbrugge[1]

[1]Bernoulli Institute, University of Groningen, Groningen, Netherlands
[2]Interactive Intelligence, Delft University of Technology, Delft, Netherlands
[3]LIACS, Leiden University, Leiden, Netherlands
{J.D.Top, Harmen.de.Weerd, L.C.Verbrugge}@rug.nl; Abisharan.R@gmail.com; C.M.Jonker@tudelft.nl

Abstract

The Mod game is an *m*-action non-zero-sum variant of rock-paper-scissors. Actions are laid out on a circle. In each round, players choose an action simultaneously, and gain a point for each player they are one step ahead of, in clockwise direction. A cooperative strategy exists, but is rarely used. In the present article, we facilitate this cooperative strategy, as well as lying, by introducing a signalling phase to the Mod game, where one player signals to the other player which action they will play this round. In this novel Mod-Signal game, players can cooperate by adhering to their signal, but they can also lie and play a different action. We conduct an experiment where human participants play the 24-action Mod-Signal game with an agent and with each other. We find that despite cooperative play being faster and more efficient in terms of gaining points, players still lie and play non-cooperatively in the majority of rounds. Our results show that participants tend to use no more than second-order theory of mind when trying to one-up their opponents. While the Mod game is mainly played as a competitive task, our Mod-Signal game can be used to investigate cooperation and deception in the context of theory of mind.

Keywords: Mod game; cooperation; deception; theory of mind; behavioural research

Introduction

The main contribution of this article is our novel Mod-Signal game, a modified version of the Mod game that allows us to study cooperation and deception in the context of theory of mind. The Mod game is a non-zero sum variant of rock-paper-scissors, where in each round, *n* players simultaneously select from one of *m* spaces arranged on a circle (Frey & Goldstone, 2013). Players gain a point for each player that they are exactly one step ahead of, in clockwise direction.

Frey and Goldstone (2013) show that participants playing the Mod game use *theory of mind* (ToM), the ability to attribute and reason about unobservable mental states of others, such as knowledge, beliefs, and intentions (Premack & Woodruff, 1978). ToM can be used recursively. For example, in the Mod game, if Ann thinks that Bob will play action a, then Ann may want to play action $a+1$ to gain a point. This is zero-order ToM, or ToM-0, as Ann only reasons about Bob's actions, but not about his mental states. However, Bob may think that Ann thinks this, opting to play action $a+2$ instead. In this case, Bob uses first-order ToM, or ToM-1, as Bob makes one perspective switch. Now, suppose Ann thinks that Bob thinks that Ann thinks that Bob will play action a. In this case Ann makes a second-order ToM attribution, as Ann makes two perspective switches. As such, the Mod game can be used to investigate which ToM orders participants use by looking at how far participants 'jump ahead' of each other.

ToM can be used to facilitate cooperation (Etel & Slaughter, 2019; Paal & Bereczkei, 2007), but it can also be used to deceive others. Ann tells a *lie* if Ann says something that she believes to be false with the intention that some addressee, such as Bob, believes that it is true (van Ditmarsch et al., 2020). To successfully lie, Ann would need to use her ToM to reason about the beliefs of Bob. Even though humans lie in many contexts (Rosenbaum et al., 2014), human deception detection is seldom better than chance (Granhag & Vrij, 2005). This contrast stresses the importance of deception research, especially in the context of ToM.

Cooperation and deception are possible in the Mod game as well: Frey (2013, Ch. 3) comments on a cooperative strategy where players "*split into two groups and 'leap-frog' through the strategy circle*".[1] As participants' payouts were based only on the points they gained and not on whether they 'beat' the other player, they could have increased their earnings by cooperating. However, mean point gains were only 32.5% of the maximum that would be obtained by playing cooperatively, showing most participants played competitively.

Similarly, participants can use deception: "*Other... ...participants would 'play dead', select the same choice repeatedly to lure others into trying to earn points off of them, so that they could increment by two. The behaviour was less effective and less common than the clustering and climbing that dominate this treatment of the game*." (Frey, 2013)

In the present article, we propose the new *Mod-Signal* game as a task that can be used to effectively study cooperation and deception in the context of ToM. The Mod-Signal game extends the Mod game by adding a non-binding signalling phase, intended to facilitate cooperative and deceptive strategies and making them more effective. Before each round, one player must select one of the available actions to signal to the other player. We restrict ourselves to two-player games such that each player can signal every other round. The signal can be used to coordinate a cooperative 'leap-frogging' strategy, where players alternate in giving each other a point. Furthermore, in our analysis, the signal serves as a reference

[1]More methods of cooperation are possible, but not reported in Frey (2013). One is trading places every round. In another, one player remains stationary, while another jumps back and forth.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

space that makes it easier to investigate our participants' behaviour and determine which ToM order they are using.[2]

Signals are *cheap talk* (Farrell, 1987): They are cost-free and non-binding. Because of the latter, participants can also use signals for deception. For example, Ann could lie that she will play a in the hopes that Bob will play $a+1$, so that she can play $a+2$ to gain a point. In both examples, the intentional aspect of lying is important (van Ditmarsch et al., 2020): Ann intends that Bob believes that Ann will play a certain action. Talwar et al. (2007) show that lying and ToM are closely related: Children who fail to understand second-order false beliefs also fail to successfully maintain a lie.

Using this novel task, we run a behavioural experiment to investigate three research questions about cooperation, deception, and ToM, which are as follows:

How often do participants use deception, and at which stage of play do they deceive? Krockow et al. (2016) show that participants primarily cooperate in repeated centipede games. In repeated rounds of the prisoner's dilemma with a known game end, participants primarily cooperate until later rounds where non-cooperation starts increasing (Bruttel et al., 2012). Because of this, we hypothesize that *participants will use cooperative actions more frequently than non-cooperative and deceptive actions, with the latter being more prevalent in later rounds of the game* (**H1**): In the Mod game, cooperation can yield more points than non-cooperation or random play, and cooperation becomes easier when a signal is present. We also predict that *cooperative actions will be faster than non-cooperative and deceptive actions* (**H2**), as leap-frogging is a straightforward procedure, whereas predicting and one-upping the other player requires ToM.

Which theory of mind (ToM) orders do participants use? While game theory prescribes that players should follow some rationally optimal solution, experimental evidence shows us that human participants often fail to apply ToM in strategic games. Adults usually use first- or at most second-order ToM in a wide variety of tasks, such as the P-Beauty Contest (Nagel, 1995), matrix games (Hedden & Zhang, 2002), normal-form games (McKelvey & Palfrey, 1995), hide-and-seek, sender-receiver games, and matching pennies (de Weerd et al., 2018), and epistemic logic puzzles (Zhang et al., 2021). Nonetheless, it has been shown that adult ToM use can be trained to up to ToM-2 by using a stepwise training regimen (Verbrugge et al., 2018). In the Mod game, participants seem to use between first- and second-order ToM on average (Frey & Goldstone, 2013). Therefore, we hypothesize that *the majority of our participants will use zero-, first-, or second-order ToM* (**H3**). As for the distribution of different orders of ToM: *We expect lower orders to see more use* (**H4**), as higher orders of ToM require more cognitive resources (Zawidzki, 2013).

Which norm is established for cooperative play? We hypothesize that *participants will cooperate by having signallers play the signal, and receivers playing* $[signal + 1]$ (**H5**). Per instructions, the signal is used to indicate what the signaller will play, so the only cooperative options for the receiver are $[signal - 1]$ and $[signal + 1]$. Playing $[signal + 1]$ would be in accordance with observed behaviour in the original Mod game (Frey & Goldstone, 2013).

In our current experiment, we let our participants play with both human participants as well as a computational agent. Including human-agent play allows us to obtain more data from our limited participant pool, caused by resource limitations, and to investigate human-agent play for future work (see Discussion section). To ensure the validity of our results, this gives rise to one additional design goal:

Our computational agent is not distinguishable from human play (**H6**). We design the agent to mimic human play (For details, see Methods section). In our Results section, we show that there are no significant differences between humans playing with other humans, and humans playing with the agent, allowing us to aggregate the data.

Methods

Participants

In total, 21 university students (9 female) between the ages of 16 and 25 (mean 20.3) participated in our study. Each participant received a fixed 10 euros as compensation. Participants signed an informed consent form prior to participation.

The Mod-Signal game

Participants played the Mod-Signal game, an adaptation of the Mod game (Frey & Goldstone, 2013) that includes a novel signalling phase. The Mod-Signal game is a two-player game where both players are shown the numbers 1 through 24 arranged on a circle (Figure 2). Each round consists of two phases: A signalling phase and an action phase. In the signalling phase, one player, the *signaller*, must choose a number to signal. We call the player who is not signalling the *receiver*. Both players alternate between being the signaller and receiver each round. Participants were instructed that the signal could be used to indicate what the signaller was going to play, but participants also knew that signallers could flout this rule, i.e., they could lie. After signalling, the signal is shown in red to both players. Then, in the action phase, both players simultaneously select a number to play, after which both numbers are revealed to both players. A player gains a point if they are directly ahead of the other player by exactly one space, in clockwise direction. The number 1 is exactly one space ahead of 24, since 24 + 1 = 1 mod 24.[3]

Experimental design

Three participants were present in a session, which consisted of three blocks. Each block, two participants played the Mod-Signal game with each other (the human-human condition), while the remaining participant played the game with a computer agent (the human-agent condition). Each participant in

[2] See Chapter 3 of Frey's thesis (Frey, 2013) for more details.

[3] x mod 24 repeatedly adds or subtracts 24 from x until the result is in the range $\{0..23\}$. For example, 49 mod 24 $= 1$, -3 mod 24 $=$ 21. 'Mod' in 'Mod(-Signal) game' is shorthand for 'modulo'.

a session played with the other two participants as well as the computer agent in exactly one block. These pairings were randomized. Each block consisted of twenty rounds of the Mod-Signal game. The computer agent was always the signaller in the first round in the human-agent condition. Each participant randomly started as the signaller in exactly one of their two human-human blocks, starting as the receiver in the other one. An example of a session can be found in Figure 1.

All code and data used in our experiment and analysis can be found at `https://github.com/jdtoprug/CogSciModSignalProject/`

The computer agent

In Frey and Goldstone (2013), the distribution of choices of the participants playing the Mod game is approximately uniform. Therefore, we let the agent draw from a uniform random distribution over the 24 choices when selecting a signal. For its actions, our agent has access to two distinct strategies. First, it can cooperate by playing a leap-frogging strategy similar to the one described in Frey (2013). The agent's leap-frogging strategy consists of playing its signal if it is the signaller, and playing the signal plus one if it is the receiver.[4]

Secondly, the agent can use a competitive strategy where it tries to 'one-up' its opponent based on the actions observed in the previous round, similar to the cyclic behaviour and 'one-upping' observed in Frey and Goldstone (2013). In this case, the agent will play $(((o_{-1}-s_{-1})+s_0+c) \bmod 24)+1$, where s_0 is the signal in the current round, s_{-1} the signal in the previous round, and o_{-1} the agent's opponent's choice in the previous round. If the agent is the signaller in the current round, then $c=-1$, otherwise $c=1$. This formula takes the distance between the agent's opponent's cooperative play in the previous round, and adds it, plus one, to the agent's opponent's cooperative play in the current round. Due to a bug in the code, o_{-1} and s_{-1} would be swapped whenever $s_{-1} > o_{-1}$, which was the case in 59 of the 400 rounds the agent played. However, we establish that this issue did not make our agent's behaviour noticeably different from human behaviour.

Our goal is to make agent play indistinguishable from human play (**H6**). We expect that strategy selection in the Mod-Signal game is comparable to a prisoner's dilemma, as mutual cooperation, which can be exploited, should yield more points than mutual competition. Bruttel et al. (2012) show that human participants tend to be more forgiving in a known-horizon repeated prisoner's dilemma than a grim strategy prescribes, so we let our agent use a procedure similar to tit-for-tat. In the first round, it always follows the leap-frogging strategy by playing the action that it signalled. In all other rounds, the probability that it follows the leap-frogging strategy is equal to the proportion of the human player's leap-frogging actions in the last three or fewer rounds. Probabilistic strategy selection makes the agent less predictable, and therefore harder to identify as an agent.

Our agent has a bias towards competitive play: Frey and Goldstone (2013) show that participants mainly play competitively in the regular Mod game. Furthermore, we expect human players to be less incentivized to play cooperatively when payment amounts are fixed. Starting in round 3, the agent always views the human's play of three rounds ago as not leap-frogging. Even if a participant fully adheres to leap-frogging, there is a one-third probability the agent moves to its competitive strategy in a given round.

[4] A receiver can also cooperate by playing the signal minus one, which would be altruistic. Our participants played signal plus one much more frequently than signal minus one, even before they encountered our agent.

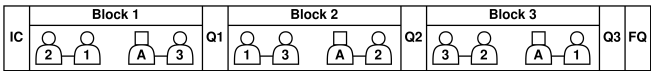

Figure 1: An example of a session. IC, Q1, Q2, Q3, and FQ indicate the informed consent form, questionnaires 1 through 3, and the final questionnaire. For each block it is indicated how players 1 through 3 and the computer agent are paired, where the leftmost player in a pair sends the first signal. The instructions read at the start of each block are not depicted.

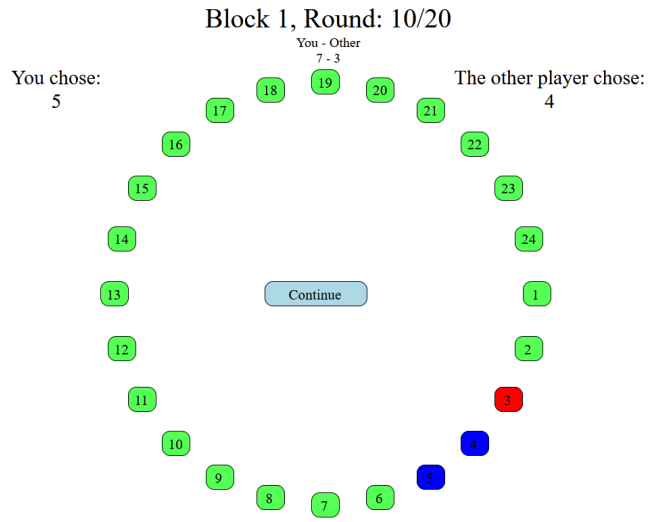

Figure 2: The Mod-Signal game as shown in our experiment.

Procedure and materials

Upon arrival at a session, participants, as well as a confederate, were seated at a computer and instructed to sign an informed consent form. Participants were led to believe the confederate was another participant. To ensure that participants did not know who they were playing with, we separated participants (including the confederate) with banners and instructed them to wear headphones. Participants first read instructions on the Mod-Signal game as well as the experimental procedure. They were instructed they could use the signal to let the other player know which number they intended to choose (as opposed to what number the other player should play). Participants were instructed that they were playing with a different player in each block. In reality, they never

played with the confederate, and played with the agent in one out of three blocks instead. To avoid priming participants towards cooperative or competitive behaviour, instructions used neutral phrasing, such as using 'other player' instead of 'opponent', and 'gain a point' instead of 'win a round'.

Participants then played four practice rounds against a simple computer agent. This agent signaled in the first and third practice round, always played what it signaled, and always played the participant's signal plus one. After finishing the practice trials, participants played three blocks of the ModSignal game consisting of twenty rounds each. At the end of each block, participants filled in a questionnaire, and they received an additional questionnaire at the end of the experiment. At the start of each block, participants were reminded that their goal was to get as many points as possible (as opposed to, e.g., 'beating the other player'), and that they were playing with a different player in each of the three blocks. In line with Veltman et al. (2019), our user interface always showed the block and round number as well as both players' scores for the current block (see Figure 2).[5]

Questionnaires

After each block, participants rated the cooperativeness, competitiveness, and competence of the other player on a 5-point Likert scale. They also described their own and their opponent's strategy, and whether they would want to play against the other player again. At the end of the experiment, participants received an additional questionnaire, which revealed that some blocks may have been played against a computer agent. Here, participants had to rate their confidence of playing with a human in each of the three blocks.

Results

Blocks 2 and 3 of one participant were excluded due to technical issues encountered during the experiment, leaving 20 blocks of 20 rounds each in each of the human-human and human-agent conditions. The remainder of this section will follow the research questions and design goals outlined in our Introduction section. The results of our statistical tests on six metrics of interest can be found in Table 1. These will be explained as we discuss each of our research questions:

Our experimental instructions did not specify how receivers should react to the signal. Nonetheless, $[\text{signal} - 1]$ was played 21 times as receiver, whereas $[\text{signal} + 1]$ was played 170 times as receiver. This suggests $[\text{signal} + 1]$ was the predominant norm for cooperative play (**H5**).

To investigate ToM use, we look at the *choice-minus-signal*, which is a player's choice minus the same round's signal, modulo 24. In our study, this metric takes the role of the *rate* found in Frey and Goldstone (2013), and is an indication of the levels of ToM a participant may have used. The distribution of signals and choice-minus-signal values for the signaller and receiver, for human players only, can be found in Figure 3. The distributions for the human-human and human-agent conditions are very similar, and Table 1 shows no significant differences between humans in the human-human condition, and humans in the human-agent condition. Because of this, we combine the human-human and human-agent conditions in these histograms.

The leftmost graph, choice-minus-signal for signallers, shows peaks of decreasing size at $[\text{signal} + 2n]$, for n an integer with $n \geq 0$. The center graph, choice-minus-signal for receivers, shows similar peaks at $[\text{signal} + 2n + 1]$. These peaks correspond to different levels of ToM. The peaks at 0 and 1 in the leftmost and center graphs correspond to zero-order ToM, as the cooperative strategy requires no perspective-switching. A player can simply play [signal] or [signal $+ 1$], a purely behavioural procedure. de Weerd et al. (2015b) show that such communication norms can indeed be established even without ToM. The leftmost and center graphs of Figure 3 show that participants usually use no more than second-order ToM with lower orders being used more frequently than higher orders (**H3**, **H4**).

We define the *expected action* as the signal when you are the signaller, and as signal $+$ 1 when you are the receiver. Defined as such, 334 of all human plays were expected, and 866 plays were unexpected. We define the *expected difference* as $[[\text{action} - \text{expected action}] \bmod 24]$. Analogous to Frey and Goldstone (2013), your acceleration for round $n \geq 2$ is your expected difference in round n, minus your expected difference in round $n-1$, modulo 24. The rightmost graph of Figure 3 shows the acceleration of all human participants across both conditions, starting in round 2. Once again, for the same reasons, we combine the human-human and human-agent conditions. There are clear peaks at 0, 2, and 22.

As hypothesized (**H2**), playing expected actions is faster than playing unexpected actions: A two-sample t-test on log-transformed participant reaction times shows that expected actions ($M = 8.2, SD = 0.6$) were indeed significantly faster than unexpected actions ($M = 8.6, SD = 0.4$), with $t(20) = -4.4715, p = 0.0002$. However, note that the physical act of playing an expected action is already faster than playing unexpected actions: As a signaller, you simply click the same action twice, whereas as a receiver, you click [signal $+$ 1].

Next, we show that choice-minus-signal values correlate with reaction times. We omit expected plays as they have already shown to be faster, and only look at choice-minus signal values in the range 0 to 6, as this is where the peaks occur. A Spearman rank-correlation test shows that there is a positive rank-correlation (of 0.27) between log-reaction times and choice-minus-signal values for this data, with $S = 23268332$ and $p \approx 10^{-12}$. This shows that the peaks in Figure 3 may indeed correspond to different levels of ToM.

We now investigate the frequency of deception by our participants. Several metrics found in Table 1 are relevant here:

- **Efficiency** - Efficiency is the proportion of rounds where a point was gained out of the total number of rounds where

[5] Showing both players' scores may cause some players to maximize the difference between scores, but Verbrugge et al. (2018) show that participants still follow instructions to be self-interested when given this information.

a point would be gained if perfect cooperation was used. Mean efficiency across both conditions for human participants was 0.15, slightly higher than the mean efficiency of 0.13 found in Frey and Goldstone (2013).[6] Perfect cooperation would result in a point gained in each round, and an efficiency of 1, whereas randomizing would give a point in 1 out of 24 rounds, or an efficiency of approximately 0.04.

- **Honesty** and **trust** - Honesty is the proportion of rounds where the signaller played the signal. We define trust as the proportion of rounds where the receiver played [signal + 1]. The distribution of honesty and trust levels for human participants in all conditions can be found in Figure 4.

Contrary to our hypothesis, participants primarily played non-cooperatively (**H1**). In fact, non-cooperation between humans started early. If we count playing the signal as signaller, [signal + 1] as receiver, and [signal − 1] as receiver, as cooperative actions, then the first non-cooperative action in the human-human condition still always occurred within the first three rounds. Despite our agent being biased towards non-cooperation, it was still more cooperative than our participants: In 16 out of 20 games in the human-agent condition, participants were the first to deviate from cooperative play.

Next, we investigate why our participants play competitively when cooperation can yield more points. In the questionnaires, 18 out of 21 participants use competitive terminology, such as 'win', 'lose', and 'opponent', whereas only three participants acknowledge that both players can gain more points by cooperating.[7] This suggests participants may have viewed the Mod-Signal game as a competitive, and not a mixed-motive game.

Recall that our computational agent should not be distinguishable from human play. In Table 1, it can be seen that, for any of our metrics, there is no significant difference between humans in the human-human condition, humans in the human-agent condition, and agents in the human-agent condition. In the final questionnaire, participants had to give a percentage rating of their confidence of playing with a human partner in each of the three blocks. This data was divided into ratings for human-human blocks and ratings for human-agent blocks, and can be found in the **Confidence other is human** row in Table 1. The results suggest that our participants could not distinguish between our agent and other participants (**H6**). However, it must be noted that *efficiency*, discussed above, approaches significance, which is in line with the observation that the agent was more cooperative. Despite its bug, our agent was better at gaining points than human players were.

Discussion and future work

The most frequent method of cooperation we observed in the Mod-Signal game is a leap-frogging method where you play the signal as signaller, and [signal + 1] as receiver. We call this method 'expected play'. In our Results section we show that expected play, which yields more points, is faster than unexpected play. This matches the finding that mutual honesty is the best policy in negotiation games such as Colored Trails (Brok, 2023). Despite this, participants mainly use non-cooperative actions. Participants may have viewed the Mod-Signal game as a competitive game, as they mainly described the game in competitive terms in the questionnaires. A possible cause might be that participants were only given a fixed payment amount and that both players' scores were always visible. Without an incentive to play cooperatively, participants may favor the more 'interesting' strategy.

Perhaps our most informative results are the histograms found in Figure 3. In the Mod-Signal game, the signal serves as a salient anchoring position for both players, whereas in the Mod game, it is difficult to determine what participants used as an anchor when taking actions. As there is no signal, participants could use both their own and the other player's previous action as an anchor. Compared to the rate and acceleration graphs in Figure 3 of Frey and Goldstone (2013), our histograms in Figure 3 show clear and easily interpretable spikes. Like the rates in the graphs of Frey and Goldstone (2013), it can be seen that choice-minus-signal is positive, though the choice-minus-signal values we observe are closer to zero than the rates in Frey and Goldstone (2013). This is probably due to the Mod-Signal game facilitating cooperation, compared to the Mod game. A signaller's choice-minus-signal is zero when cooperating, whereas a receiver's choice-minus-signal, when cooperating, is plus or minus one.

The first histogram in Figure 3 shows that signallers frequently deceived the other player by playing [signal + 2] or [signal + 4]. However, note that it may be difficult to distinguish between deception and competitive play. After several rounds of deception, both players may reach a common ground where it is assumed that the signaller will not play the signal, and that both players will try to one-up each other.

The spikes in our choice-minus-signal graphs show different levels of theory of mind (ToM), with each successive spike being smaller. Our results suggest that few participants use more than two orders of ToM. This corresponds to the well-established finding that people have difficulty with increasingly higher levels of ToM (see, e.g., Camerer et al., 2004; Nagel, 1995). Similar to Frey and Goldstone (2013), our participants' acceleration is centered around zero, as seen in Figure 3: Participants mainly keep using the same difference to the cooperative play that they used in the previous rounds, sometimes adjusting by moving two steps ahead or two steps backwards. This may suggest that participants sometimes adjust their ToM order.

Our current results suggest we should replicate our experiment with several modifications: Our agent had a bug, which should be fixed. Our low number of participants should be increased to ensure our findings stay (in)significant. For example, the difference in *efficiency* in Table 1 might be significant with more participants. To facilitate cooperation, participants should be paid based on the points they gain, as was

[6] Once again based on Figure 2 of Frey and Goldstone (2013).

[7] Decided by two annotators with 0.88 initial agreement.

Table 1: In each row, 'metric' is the dependent variable under consideration. The column 'test' is the statistical test that was used to compare the means (M) of this variable across three conditions: HH is human data in the human-human condition, HA is human data in the human-agent condition, and A is agent data in the human-agent condition. For Chi-Squared tests, M is a proportion instead of a mean. If the 'A' column is empty, HH and HA consist of *all* data for their respective conditions. The columns 'test statistic' and 'p' (for 'p-value') show the results of each statistical test. Because there are no significant differences between humans in the human-human condition and humans in the human-agent condition, we combine both conditions when investigating human play.

metric	test	HH		HA		A		test statistic	p
		M	SD	M	SD	M	SD		
Confidence other is human	T-Test	65.8%	13.9%	64.4%	26.4%	-	-	t(18)=0.18	0.86
Efficiency	Chi-Squared	0.16	-	0.14	-	0.20	-	χ^2=5.12, df=2	0.08
Honesty	Chi-Squared	0.26	-	0.30	-	0.29	-	χ^2=1.17, df=2	0.56
Trust	Chi-Squared	0.30	-	0.26	-	0.30	-	χ^2=0.90, df=2	0.64

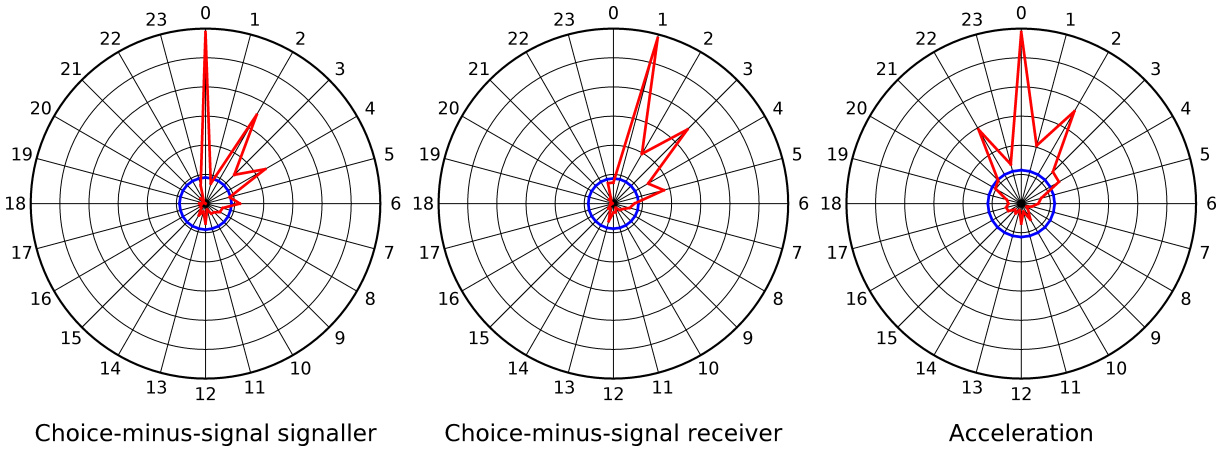

Figure 3: Histograms over choice-minus-signal and acceleration values in all blocks of the Mod-Signal game. In each graph, the blue curve shows expected random behaviour, while the red curve shows the behaviour of our human participants.

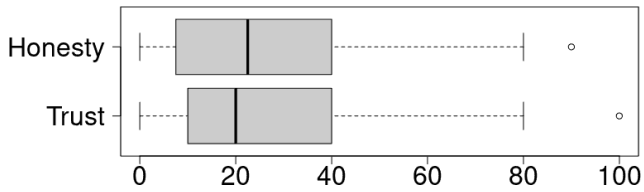

Figure 4: Boxplots of honesty and trust levels of human participants for all trials, as a percentage.

done in Frey and Goldstone (2013). To prevent wealth effects, payments can be based on a single, randomly-chosen block only (Azrieli et al., 2018). Lastly, to ensure human-agent blocks cannot influence human-human blocks and vice versa, we should separate these conditions.

In future work, we intend to develop computational models of the human behavioural data we obtained in our experiment as well as training agents with which participants can play the Mod-Signal game, similar to the models presented in Devaine et al. (2014), de Weerd et al. (2014), de Weerd et al. (2015a), Verbrugge et al. (2018) and Veltman et al. (2019). Computational models give more insight into mental processes of our participants and hence into cooperation, deception, and ToM, while training agents could help improve participants' ToM, or could nudge participants towards cooperative play. Our agent and its results may inform the design of such models and agents. For example, like our agent, humans might only look at the previous round when deciding on an action for the current round.

Conclusion

In this article, we introduce a novel task called the Mod-Signal game. While the Mod game is mainly played as a competitive game, the Mod-Signal game allows for the investigation of cooperation and deception in the context of ToM, as its novel signalling phase can facilitate both cooperation and deception. We find that participants mainly play non-cooperatively, despite cooperation being faster and yielding more points, and we find that participants usually use no more than two ToM steps. The Mod-Signal game and variants thereof can be employed in future behavioural research, and our data and agent may be used to develop computational models of human play in the Mod-Signal game, which can aid in our understanding of cooperation, deception, and ToM.

Acknowledgements

This research was funded by the project 'Hybrid Intelligence: Augmenting Human Intellect', a 10-year Gravitation programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022. We would like to thank Maxime Bos, BSc for her contribution to developing the experiment's and agent's code, and we would like to thank our anonymous reviewers for their helpful comments and suggestions.

References

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, *126*(4), 1472–1503.

Brok, S. (2023). *The Influence of Lying in a Negotiation Setting: Colored Trails* [Master's thesis, University of Groningen].

Bruttel, L. V., Güth, W., & Kamecke, U. (2012). Finitely repeated prisoners' dilemma experiments without a commonly known end. *International Journal of Game Theory*, *41*, 23–47.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

de Weerd, H. A., Broers, E., & Verbrugge, R. (2015a). Savvy software agents can encourage the use of second-order theory of mind by negotiators. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 542–547.

de Weerd, H. A., Diepgrond, D., & Verbrugge, R. (2018). Estimating the use of higher-order theory of mind using computational agents. *The BE Journal of Theoretical Economics*, *18*(2),

de Weerd, H. A., Verbrugge, R., & Verheij, H. B. (2014). Theory of mind in the Mod game: An agent-based model of strategic reasoning. In A. Herzig & E. Lorini (Eds.), *Proceedings of the European Conference on Social Intelligence (ECSI-2014), CEUR Workshop Proceedings* (pp. 128–136, Vol. 1283).

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

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLOS Computational Biology*, *10*(12), e1003992.

Etel, E., & Slaughter, V. (2019). Theory of mind and peer cooperation in two play contexts. *Journal of Applied Developmental Psychology*, *60*, 87–95.

Farrell, J. (1987). Cheap talk, coordination, and entry. *The RAND Journal of Economics*, *18*(1), 34–39.

Frey, S. (2013). *Complex Collective Dynamics in Human Higher-Level Reasoning. A Study over Multiple Methods* [Doctoral dissertation, Indiana University].

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

Granhag, P. A., & Vrij, A. (2005). Deception detection. In N. Brewer & K. D. Williams (Eds.), *Psychology and Law: An Empirical Perspective* (pp. 43–92). The Guilford Press.

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

Krockow, E. M., Colman, A. M., & Pulford, B. D. (2016). Cooperation in repeated interactions: A systematic review of centipede game experiments, 1992–2016. *European Review of Social Psychology*, *27*(1), 231–282.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*(1), 6–38.

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

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences*, *43*(3), 541–551.

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

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, *45*, 181–196.

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

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

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

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

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.

Zhang, C., Ham, H., & Holliday, W. H. (2021). Does Amy know Ben knows you know your cards? A computational model of higher-order epistemic reasoning. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2588–2594.