Interacting with Man or Machine: When Do Humans Reason Better?*

Ralph-C. Bayer† Ludovic Renou‡

December 8, 2023

Abstract

The resolution of complex problems is widely seen as the next challenge for hybrid human-AI teams. This paper uses experiments to assess whether there is a difference in the quality of human reasoning depending on whether they interact with humans or algorithms. For this purpose, we design an interactive reasoning task and compare the performance of humans when paired with other humans and AI. Varying the difficulty of the task (i.e. steps of counterfactual reasoning required), we find that for simple tasks subjects perform much better if they play with other humans, while the opposite is true for difficult problems. Additional experiments, in which subjects play with human experts, show that the differences are driven by the knowledge that AI reasons correctly rather than that it is non-human.

*We would like to thank three anonymous referees, the editor and an associate editor for extremely helpful comments and suggestions, which improved this paper considerably.

†School of Economics and Public Policy, University of Adelaide, Nexus 10, Adelaide 5005, Australia. Phone: +61 (0)8 8303 4666. ralph.bayer@adelaide.edu.au

‡School of Economics and Finance, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom. l.renou@qmul.ac.uk

"Organizations that use machines merely to displace workers through automation will miss the full potential of AI. Such a strategy is misguided from the get-go. Tomorrow's leaders will instead be those who embrace collaborative intelligence [...]" (Wilson and Daugehrty, 2018, p. 123).

1 Introduction

An important purpose of AI is to help humans make better decisions. Traders often make their decisions after interacting with AI-based investment research tools combining their "hot cognition" with AI's "cold cognition" (Buczynski et al., 2021). In our everyday life we encounter more and more AI-based digital assistants (like Siri, Google Home, or Alexa). Information systems engineers have accepted the challenge to provide answers to the question of how users and AI can best interact according to the task at hand (Maedche et al., 2019).

The arguably biggest challenge for AI designers is the design of systems, where humans and AI collaboratively solve complex problems. Memmert and Bittner (2022) review the emerging literature with respect to the direction taken. The predominant research questions addressed by experiments with game-like environments in AI, systems science, and engineering are the determinants of human acceptance of AI and its perceived likeability, humanness, and trustworthiness (Tabrez et al., 2019; Ashktorab et al., 2020; Liang et al., 2019; Wang et al., 2016; Merritt and McGee, 2012). Only in some studies where there is an objective performance measure is joint performance evaluated (Geraghty et al., 2020; Gao et al., 2020). The important question of whether human reasoning quality differs depending on whether they interact with humans or with AI has only received limited attention. Knowing the factors that determine whether interacting with humans or AI leads to better reasoning is important for the decision of which problems should be solved by groups of humans and which by mixed human-AI teams.

The mentioned studies are not suitable for answering this important question. The main contribution of this paper is the design of experiments that provide an answer to this question. For this, we exploit the methodological advantages of experimental economics. While the experiments described above are very rich and sophisticated with respect to the game environment and the AI algorithms used, they suffer from lack of control. The complexity of the environment makes it virtually impossible to measure the quality of reasoning. Furthermore, human participants

are not incentivized and deception is used regularly.[1] Our novel experiments are designed to cleanly measure the reasoning performance of our subjects conditional on the difficulty of the task and on being paired with humans or algorithms. To achieve this, we need a well-defined group task that allows a clean measure of individual reasoning quality.

We use a logical puzzle commonly referred to as the Red-Hat Puzzle or the Dirty-Faces Game (Littlewood, 1953), which has recently been used to test iteration abilities in humans (Weber, 2001; Bayer and Renou, 2016b,a). We develop two versions of this puzzle, one where a subject plays with other humans, and one where a subject plays with algorithms. The structure of the game and the decision interface for the two versions are identical. Both conditions present the puzzle framed as a game with either other humans or infallible AI players. The structure of the Red-Hat Puzzle allows us to vary the difficulty of the puzzles as measured by the steps of counterfactual reasoning required. We employ four different difficulty levels in both conditions. We find that humans reason better when paired with humans in easy tasks (i.e., two steps of iterative reasoning required), while interacting with algorithms yields better reasoning in difficult tasks (four steps required). We conjecture that the human mode of reasoning differs depending on with whom they play.

The two main differences between the two treatments are a) participants facing humans or machines and b) that the humans' opponents might make mistakes while the algorithms are always correct. Therefore, our initial experiments cannot cleanly attribute the observed treatment effect to either of these two differences. In an additional experiment, we pair our subjects with expert humans, where it is common knowledge that they are able to solve the puzzles. The reasoning quality in this expert condition is virtually identical to the reasoning quality in the AI condition, while it is significantly different from the behavior in the human condition. We conclude that knowing that you are playing with humans who also have to figure out how to solve the puzzle and therefore are in the same boat as yourself activates a different cognition mode.

We conjecture that subjects who are paired with humans in the same situation are building mental models of their own situation and the situation that other

[1] In most of these studies participants always interact with the AI, and treatment variations consist of either being correctly informed of this fact or being wrongly told that their partner was human.

humans face (see Johnson-Laird, 2006, for an introduction to mental-model theory). For simple tasks, this is a very successful mode of reasoning, as it is easy to put yourself in the shoes of the other humans. For difficult tasks, the number of mental models a human needs to hold in her memory becomes prohibitive. When playing with algorithms or experts who know the underlying algorithm, humans do not build mental models but try to understand the pattern in the algorithmic behavior. This is more difficult in general, but once it is accomplished, it still works in difficult situations.

2 Human-AI interaction experiments in business, economics, and psychology

There is a quickly growing literature in organizational decision making that develops frameworks for how to best organize human-AI interaction (e.g, Shrestha et al., 2019; Murray et al., 2021; Kellogg et al., 2020). In the experimental business literature, the determinants of distrust in algorithmic advice are a very prominent research topic. Studies carried out in experimental psychology in general find that humans are to a certain degree algorithm-averse (Dietvorst and Bharti, 2020; Dietvorst et al., 2015; Efendić et al., 2020; Hertz and Wiese, 2019). This is also the case in a variety of different business domains such as accounting (Cao et al., 2021; Commerford et al., 2021), management (Chen et al., 2021), or finance (Filiz et al., 2022).[2] Factors that have been found to mitigate algorithm aversion are time pressure (Jung and Seiter, 2021), feedback on poor past performance when ignoring AI advice (Dargnies et al., 2022), the possibility to modify the algorithmic advice (Kawaguchi, 2021), tournament incentives and AI advice framed to contain human expertise (Greiner et al., 2022).

There is also a large body of literature at the intersection of economics and finance, where trading algorithms are used in market experiments. The recent literature is reviewed in Bao et al. (2022). Several studies compare humans to algorithms and assess their relative performance (e.g., Das et al., 2001; Gjerstad, 2007; Luca and Cliff, 2011; Akiyama et al., 2017; Tai et al., 2018; Hanaki et al., 2018; Aldrich and Vargas, 2019; Peng et al., 2020; Ashktorab et al., 2020). In general, algorithms perform better, while humans better adjust to crashes or periods of high volatility.

[2] An exception in finance is Germann and Merkle (2019), who do not find algorithm-aversion.

Our study is related to this literature, since performance in a market is related to reasoning performance. The main difference is that we study cooperative problem solving, while market environments are competitive.

The use of computer players in experiments has a long tradition in experimental economics. March (2021) surveys the literature and concludes that computer players are used predominantly as methodological tools.[3] Direct investigations of the impact of the human-AI interaction on behavior and outcomes are rare. March concludes that it is still possible to draw a few conclusions about human-AI interactions from this literature. The two insights most relevant to our paper are that humans behave differently when paired with computer players and that their cognition changes when interacting with computer players. Our experiments further strengthen these points. Our setting ensures that there is no incentive to behave differently for our participants if paired with humans or AI, regardless of what they believe about how their partners play. Hence, our finding that behavior differs across playing with humans and AI, shows that there is a direct link between the changed mode of reasoning and behavior.

3 The Red-Hat Puzzle and its experimental implementation

Picture the following situation. There are n players in a room wearing white or red hats. Each player sees the hat color of all the other players but not his own. Everyone's task is to find out one's own hat color. It is commonly known that at least one person has a red hat. A moderator comes along tells players that at least on person wears a red hat and asks the players about their hat color. Players can answer "I have a red hat with certainty", "I have a white hat with certainty" or "I can't possibly know." Once everyone has answered, the responses of all players are made public. Once everyone has learned the previous responses, the moderator asks again. This continues until everyone has deduced his hat color.

Although at first sight it seems unclear how people could deduce their hat color, some counterfactual reasoning soon reveals that it is possible. Start with the case where there is only one person with a red hat. Then this person sees no other red

[3]The typical methodological purpose of the use of computer players are the removal of social preferences, the elimination of strategic uncertainty, noise reduction, or the inducement of types of players.

hats and therefore should be able to infer, taking one step of reasoning, that she must have a red hat. She announces "I have a red hat with certainty." All other players see one red hat, ex ante cannot possibly know their hat color, and should answer accordingly. In the next round, though, they can infer their hat color by looking at what the person with the red hat has said the first time round. The two steps of counterfactual reasoning for this are as follows: "This person said that she has a red hat. She could only have inferred this if she did not see any other red hats. Therefore, she did not see any red hats and I must have a white hat."

Now suppose that we have a situation with two red hats. The players with the red hats see one red hat each, the others see two. In the first round of answers, nobody can possibly know their hat color. Then in round two, the players with red hats can know their hat color. The reasoning (using two steps) is as follows: "The person I see wearing a red hat would have inferred having a red hat if she had not seen anyone else with a red hat, but she announced that she did not know her hat color. So she must have seen another red hat. As I don't see anybody else with a red hat I must have one." Consequently, in the second period the two players with red hats will announce "I have a red hat with certainty" while the others cannot possibly know. In the next round, though, the white-hatted players can deduce their hat color by making three steps of counterfactual reasoning: "the two red-hatted players inferred that they have red hats, realizing that the other player could not infer their hat color in the first round. But in order to infer that, they could not have seen a third person wearing a red hat. So I must have a white hat."

This logic extends to more and more complex situations with more and more red hats. In general, it is always possible to solve the puzzle. If m is the number of red hats a person sees, then the number of counterfactual reasoning steps necessary to solve the puzzle is $m + 1$. There is one complication, which we exploit in our design. An individual has to rely on the other players' announcements stemming from correct reasoning in all but the simplest case (where they only see white hats). Playing this game with other humans or algorithms causes strategic uncertainty of different kinds. Playing with humans might help improve decision-making, as it provides a cue to put oneself in the shoes of others, which is one way of solving the puzzle. Playing with computers (or experts who are always correct), on the other hand, might provide a cue to think about a rule or algorithm, which is another way of solving the puzzle.

For our experiments, we chose to restrict the number of hat carriers to four.

Then, from the point of view of a player, there are seven logically distinct situations. The situations differ in how many red hats the player sees and if she wears a red hat herself or not. Table 1 lists the situations where the level of difficulty, measured in necessary steps of reasoning, is equal to the number of red hats seen plus one.

Red hats seen	0	1	1	2	2	3	3
Reasoning steps required	1	2	2	3	3	4	4
Having a red hat	yes	no	yes	no	yes	no	yes

Table 1: The seven distinct situations

4 Experiment 1: AI and human condition

In what follows, we explain the two conditions that are designed to test the impact of playing with humans or AI on the quality of reasoning. In both conditions, subjects face strategic uncertainty. If paired with humans they cannot be sure how they will behave. If paired with AI, strategic uncertainty is only removed if the humans perfectly understand the structure of the puzzle. However, as we will show, there is always a unique logically correct choice for any history regardless of how it has been reached. Therefore, differences in beliefs about the game of the opponent are irrelevant for someone who fully understands the structure of the problem. The way a person reasons is likely to depend on the source of strategic uncertainty, though. If playing with humans, a participant is likely to be prompted to think about what the other humans see and how they decide what to answer in each round. Instead, if paired with infallible AI players, a participant is likely to be prompted to determine the algorithm the AI players are using. Moreover, we are careful to make sure that the conditions have as few differences as possible and that there is no incentive in the human condition to not follow the logically correct solution path.

4.1 The AI condition

In the AI condition, subjects were paired with three computer players. Subjects were informed in the instructions that at any stage of the game only one of the following three answers was logically correct: "I have a red hat with certainty", "I have a white hat with certainty" or "I can't possibly know." Furthermore, subjects were informed that computers will always choose the logically correct answer. Subjects saw the hat

colors of the three computers on the computer screen. Their own hat color was shown as a question mark. In a pregame test subjects had to answer control questions that showed if they understood the screen and its immediate implications for the information the computers have. After each round of announcements, the subject sees all previous own and the computers' announcements. Each subject played only one randomly determined puzzle out of the seven distinct puzzles. All subjects who participated in the AI condition entered a draw for two prizes worth 300 Australian dollars (approximately $US 200) each if they solved their puzzle correctly. Note that subjects were told that solving a puzzle correctly meant choosing the logically correct answer until the correct color of their hat was determined.

4.2 The human condition

For the human condition our aim is to produce an environment as close as possible to the computer environment, where all players are human. In particular, we are interested in an environment that generates the same incentives as the AI condition, which are robust to other-regarding preferences. In other words, we are looking for an environment where subjects have an incentive to follow the logical path of reporting in order to determine their hat color. We replace all the computer players with humans. Now four humans are playing in one group. Instructions and test questions are identical, up to the few changes necessary to accommodate groups consisting entirely of human subjects. The only difference on the decision screen was that the word "computer" was replaced by "human." Again, we stressed in the instructions that there is a logically correct solution to the puzzle. We add the information that a group has the necessary number of announcement rounds to find all their hat colors. Our aim is to give the same incentives to subjects in the human condition as in the AI condition. For this purpose, we have to adjust the payment rules. In the human condition, a subject in a group gets a ticket for the lottery of two prizes of AUD 300 if she **and the other three subjects** of the group correctly determine their hat color.

There are multiple reasons for making the payment contingent on the whole group's success rather than on individual behavior alone. One rationale is related to the possibility that a subject makes a mistake early on, which "spoils" the game for the others. Say a person who sees two red hats in the first round already wrongly announces "I have a white hat." In this case the other players observe a clear contradiction and realize that someone made a mistake. Hence, it is not possible

anymore for these players to logically deduce what their hat color is. Consequently, it is not possible to individually determine if this person is correct and to decide if this person deserves a ticket or not if an individual payment rule is used. Under the collective payment, in this case no one gets a ticket. We stopped the game whenever someone made a mistake, and other players cannot logically deduce anything from the other players' answers. Note that not all mistakes lead to logical inconsistencies. To see this, assume that a participant sees one other red hat and correctly reports "I cannot possibly know" in period one. The person with a red hat should declare "I have a red hat", or "I cannot possibly know" depending on whether or not she sees another red hat. Both answers, regardless of whether the correct one was chosen, still allow the other player to draw conclusions. Therefore, it is not necessary to stop the game in this situation, even if a mistake was made. Stopping the game for any mistake regardless of whether it caused a logical inconsistency or not would have exacerbated the problem that some participants' decisions are not observed, which causes a selection problem (see Section 4.4 for details).

An additional benefit of the collective payment is that distributional motives (like maximizing the payoff difference to other players or maximizing the total group profit) that could lead to different behavior do not change incentives, since all players in the group have the same payoff by construction. This design provides an environment with incentives that are as similar to the AI condition as possible. However, it does not remove strategic uncertainty. A subject who fully understands how the puzzle should be solved by all players cannot be sure that the other players will follow the path to the solution. However, despite the strategic uncertainty, there is no incentive for a subject to deviate knowingly from the solution path.[4] This enables us to compare the reasoning quality of humans in situations that differ only with respect to the partner being human or AI, as long as the reasoning effort is not different between treatments.

One possible concern with the joint payment rule is that subjects might exert less effort to solve the puzzle than when they are solely responsible for payment. We are quite confident that the high prize and the intrinsic motivation to solve the puzzle together are enough to induce maximum effort under both conditions. A comparison of the time subjects take across the two conditions will provide evidence

[4] Any deviation from the logical path is weakly dominated by following the path. Any positive probability placed on the other players following the logical path makes following the path as well the unique best response.

for this hypothesis.

In conclusion of this section, we briefly want to comment on an alternative design that we considered but decided against. One could have told the subjects that the game is stopped as soon as a player deviates from the logically correct path. This would have removed strategic uncertainty in the interim. Subjects would have known that their human partners have not made a mistake up to the period in which they are in. In this case, the only strategic uncertainty points to the future. Then giving a ticket to all individuals who had not made a mistake up to the point where the game stops also gives rise to a dominant strategy equilibrium. We decided against this design because social preferences would come into play. Making a mistake early might not be too bad for some players anymore, since it gives the other three group members a lottery ticket. With social preferences that include a social efficiency motive (Charness and Rabin, 2002; Andreoni and Miller, 2002), many types of play, including deliberately making a mistake early, become rationalizable.

4.3 Experimental procedures

This study follows the standard rules for economic experiments: no deception, fully scripted instructions, and monetary incentives (Hertwig and Ortmann, 2001). We recruited our participants using ORSEE (Greiner, 2015) from our subject pool, which contains undergraduate and postgraduate students from different disciplines and from different universities in Adelaide together with some non-students. The experiments were conducted at AdLab, the Adelaide Laboratory for Experimental Economics, prizes were drawn among the eligible subjects once all sessions were completed, and the prize monies were paid out in private.[5] The actual treatments were programmed in z-Tree (Fischbacher, 2007). The situations subjects could face differed with respect to the number of counterfactual reasoning steps that are necessary to determine the color of their own hat. One situation requires one step of reasoning, while in two situations each, two, three, and four steps are required. Table 2 reports the number of subjects in situations with one, two, three and four steps of reasoning in the two conditions. [6]

[5]The data for Experiment 1 stem from a broader long-running research project on cognitive abilities, counterfactual reasoning and behavior in strategic situations and were collected between November 2009 and August 2010 but remained unused for a prolonged period of time.

[6]Note that in the human condition, subjects in the same group can have different levels of difficulty, as the difficulty is determined by the number of red hats a subject sees, which differs whenever two subjects have different hat colors.

	COMPUTER 5 sessions	**HUMAN** 12 sessions
1 step	21	15
2 steps	40	69
3 steps	34	69
4 steps	34	75
Total	129	228

Table 2: Number of participants in the different conditions

4.4 Sample selection in the human condition

If a player in the group makes a mistake early, which stops the game if it creates a logical inconsistency, then we do not observe if the other players are able to solve the puzzle. This is unavoidable due to the structure of the puzzle and does not originate from the collective payment rule. Similarly, some mistakes that are not ending the game (such as delaying the announcement of a deducible hat color) either do not allow the other players to draw valid inferences or even imply a different logical course of action. In anticipation, we ran more sessions in the Human treatment. Hence, an easy solution to the limited observability could be to just drop all subjects that cannot be classified as correct or wrong. Unfortunately, this procedure would lead to biased results. To see this, imagine the following scenario: A player with a white hat sees two red hats and makes a mistake in the first round to announce "I have a red hat." This player was clearly wrong and will be counted as such. Also, suppose that the other players all announce "I cannot possibly know" in the first round, which is logically correct. These three other players' actions would not enter the analysis, since the game ends after the mistake of the other subject. The information that these subjects have completed the first round of answers without mistakes is lost. Just dropping these observations leads to a downward biased estimate of the success rate for all difficulty levels that require more than one step of counterfactual reasoning.

In what follows, we demonstrate this with the help of an example. Suppose the true distribution of abilities is such that for a given puzzle 3 out of 4 people can solve the puzzle but the other quarter of subjects always makes a mistake in the first round. The true probability of individual success is 3/4. If someone makes

a mistake early, then we cannot observe whether the others are able to solve the puzzle to the end or not. The average number of subjects we observe finishing the puzzle correctly per group is $4 * (3/4)^4 \approx 1.27$, since correct answers would only be observed in groups with four subjects that are able to solve the puzzle. In all other cases we would only observe the subjects that are making mistakes, while those that are not making mistakes in the first round would be missing observations. Taking into account that the number of errors in a group is binomially distributed, we can calculate the number of expected observed errors per group, which is $4 * 1/4 = 1$. By just dropping all the missing observations, our estimated success probability would be the number of successes observed divided by the number of successes and failures observed. With the sample size going to infinity, our estimate would converge to the expected number of observed successes per group divided by the sum of expected successes and expected failures. In the example, this yields an estimated individual success probability of $1.27/(1.27 + 1) \approx 0.56$, which is clearly lower than the true success probability of 3/4.

4.5 Results

It will be necessary to correct for the selection bias in the human condition detailed above. We assign the value "correct" to a subject if she solved her puzzle correctly. The value "wrong" is assigned if the subject made a mistake. Finally, all subjects who ended up in a situation where they had no possibility to infer their hat colors due to mistakes of others are assigned the value "not observed." Table 3 shows the results of this classification (together with play in the AI condition). A first look already reveals some interesting facts. Compared to the AI condition, subjects in the human condition seem to do very well in the two-step puzzles, since 32 of 42 observed subjects (69.5%) solved the puzzle correctly, while a third of all the subjects were not observed, which means that they were still on the correct path when their game ended prematurely. In contrast, not a single person solved the puzzle with four steps correctly, while still a fifth of subject did in the AI condition.

A more formal analysis needs to correct for the describe selection problem. For this, we employ a sample-selection probit approach in the tradition of Heckman (1979) in order to correct for the selection bias. These models contain two equations, a selection equation that estimates the determinants for a subject being observed, and a second equation that estimates the determinants of the variable of interest. In the equation of interest, the results from the selection equation are used to correct

	HUMAN				AI			
	1 step	*2 steps*	*3 steps*	*4 steps*	*1 step*	*2 steps*	*3 steps*	*4 steps*
correct	15	32	2	0	21	21	6	7
	100.0%	46.4%	2.9%	0.0%	100%	52.5%	17.7%	20.6%
wrong	0	14	29	34	0	19	28	27
	0.0%	20.3%	42.0%	45.3%	0%	47.5%	82.3%	79.4%
unobs.	–	23	38	41	–	–	–	–
	–	33.3%	55.1%	54.7%	–	–	–	–
Total	15	69	69	75	21	40	34	34
	100%	100%	100%	100%	100%	100%	100%	100%

Table 3: Play in the human and AI condition

for the non-random sample of observations. In our case, the variable of interest (solving a puzzle correctly) is dichotomous. Therefore, we use probit models on both stages by following the procedure first proposed in Van de Ven and Van Praag (1981).

In order to obtain identification independently of the functional form of the probit, we require at least one variable in the selection equation that has no direct influence on the performance of subjects. The intuition for this is as follows: The results of the selection equation are used to form an additional variable in the equation of interest to control for subjects' different likelihoods of being observed. For the additional variable to contain any useful information for the equation of interest beyond the functional form of the probit regression it was generated with, it needs to contain at least one relevant variable that is not already contained in the equation of interest. Recall that a subject is not selected to be observed whenever other players make mistakes early. We use a dummy for the number of male players that a person plays with as an important variable, since males tend to make fewer mistakes in our puzzles. Furthermore, we use a dummy variable for the hat color of the individual, since this is what the other players see of a subject. Similarly, we include a set of dummies that indicates in which position (left, left-middle, right-middle, or right) a subject's answers were presented on-screen to the other subjects. Together, these variables appropriately identify the selection equation. The coefficients of the selection equation can be found in Appendix A. As a robustness check, we will in a later Section compare what we observe in the human condition with what we would

have observed if the play in the AI condition had occurred in groups of four.

It is straightforward to estimate the success rate of subjects in the AI condition for the different number of steps required. The simplest way is just to take the fraction of subjects that correctly solved the puzzles. A slightly more sophisticated approach uses probit regression. In such a regression, the dichotomous dependent variable takes the value of one if a puzzle was solved successfully and zero otherwise. Individual characteristics such as mathematical background, gender, course of study, and age can be controlled for. The most important independent variable is the number of reasoning steps required for a puzzle. Then, using the regression results, average predicted success rates for the different levels of difficulty and their standard errors can be calculated. As mentioned above, we have to use a two-step selection model in the human condition to correct for the selection error in order to arrive at comparable estimates.

Table 4 shows the results (average marginal effects) of the probit regression (AI condition) and the second-stage probit of the sample selection regression (human condition).

If subjects play with computers, on average the probability that subjects can solve a puzzle with three or four steps is lower by about 30 percentage points than that for solving a puzzle requiring two steps. The success rates for puzzles with three or four steps are not significantly different. We did not obtain a coefficient for one step, as everyone solved that problem. When playing with computers, medical students tend to perform worse than other students. Their performance is significantly worse than that of subjects classified as "others" ($p < 0.05$, Wald test, one-sided) and as law students ($p < 0.076$). If subjects play with humans, then all one-step problems are solved correctly. Similarly, when playing with humans subjects are less likely to solve problems with three steps than with two steps. This time, we do not obtain an estimate of the impact of four-step puzzles on likelihood, since remarkably **all 34** observed subjects did not solve the level four puzzles when playing with humans. More evidence that reasoning differed considerably across the two conditions is that medical students performed far better than other students ($p < 0.02$, one-sided Wald tests versus "Engineering/Science, "Law" and "arts / economics / business"). Recall that in the AI condition, medical students performed poorly. Finally, as observed in other studies, we observe a gender effect that is robust between conditions.

We are primarily interested in the performance of humans paired with humans

	AI Probit	HUMAN 2nd stage Probit
Prob{*correct* = 1}		
Step dummies (*2 steps* is base)		
3 steps	−0.310**	−0.524***
	(0.099)	(0.062)
4 steps	−0.291**	
	(0.098)	
Course dummies (*Arts/Econ/Business* is base)		
Engineering/Science	0.155	−0.006
	(0.106)	(0.010)
Law	0.301	−0.092
	(0.246)	(0.105)
Medicine	−0.062	0.247*
	(0.132)	(0.103)
Other	0.335	−0.036
	(0.201)	(0.167)
Male	0.199*	0.190**
	(0.083)	(0.064)
Controls (age, maths, control ok) not significant		
$Log\mathcal{L}$	−51.376	−117.353
N	108	138

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Average marginal effects on success probabilities

versus those who faced computers. For this purpose we calculate the predictive margins (averaged over the sample) for the different levels of difficulty from the two regressions (probit for the AI condition and sample-selection probit for the human condition). Figure 1 shows the predicted success probabilities (with error bars that represent the 95 percent confidence interval) for the two treatments. This gives rise to the central result of this paper.

Finding 1 *Subjects have higher success rates in human groups if two steps of reasoning are necessary, while the success rate is higher playing with computers if four steps are required.*

In order to establish the finding above, we constructed tests of proportions across conditions for the different steps using the estimated success rates and their standard errors from the estimation. The difference is highly significant for two steps (z-test, $p < 0.01$, two-sided), but not significant for three steps ($p > 0.42$, two-sided). For four steps, frequentist statistical tests are not very sensible as the predicted probability in the human condition has no variation.[7] Instead we used a Bayesian approach, where we assume a flat prior for the true success probability of subjects in both treatments and then update our prior according to the observed data. The resulting posterior probability that the success rate in the human condition is at least as high as in the AI condition is smaller than 0.004.[8]

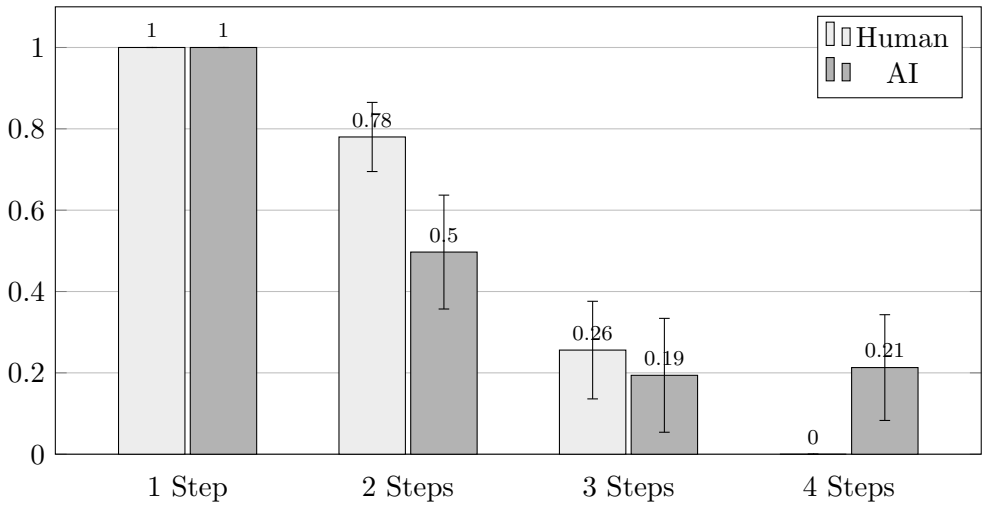


Figure 1: Success rates across difficulty levels in the human and AI conditions

Comparing the estimated success rates clearly shows that it matters if subjects play with computers or humans. Before we can discuss the mechanism behind this result, we need to make sure that the differences are not an artifact of the payment rules inducing different levels of effort in the two conditions. A priori, it is unlikely that effort differences drive performance differences in tasks like ours. Camerer

[7] Taking the frequentist approach seriously would yield a p-value of zero, as the estimated success probability in the human condition is zero without error, and we observed at least one success in the AI condition.

[8] Starting with a uniform prior for the success probability, the posterior after z successes in N trials is given by the beta distribution with parameters $z+1$ and $N-z+1$. Denoting the posterior densities as f_{AI} and f_H, the probability that the success rate in the human condition is at least as high as in the AI condition is $\int_0^1 f_{AI}(p)(1-F_H(p))\,dp$.

and Hogarth (1999) show in their meta-analysis that switching from hypothetical to incentivized experiments often improves performance and efforts, while the incentive size has a small impact or no impact at all. Bonner et al. (2000) show in another meta-analysis that incentive effects are particularly rare in problem-solving and reasoning tasks. The predominant explanation is that intrinsic motivation induces subjects to think as hard as they can for as long as they can sustain it.

A good proxy for the effective strength of incentives is the time that subjects take to think about a decision. Note that we did neither have a time limit for decisions nor displayed the time taken. Hence, subjects could think as long as they wanted before each announcement. Table 5 reports the average and (in parentheses) the median decision times for each problem and decision stage. The time that subjects took to think about their decision is extremely similar in the two treatments. We observe large differences in reasoning quality in problems with two steps, while the average time taken in the two treatments is within a second at both decision stages. In the human condition, not a single subject solved a four-step puzzle, while in the AI condition about every fifth subject solved the puzzle correctly. Still, for the first three decisions, the average thinking times are extremely similar. In the final and therefore crucial decision stage, the average decision time in the human condition is **longer than** in the AI condition, which is inconsistent with subjects exerting less effort in the human condition.

decision	1 step		2 steps		3 steps		4 steps	
	human	AI	human	AI	human	AI	human	AI
1st	15.5 (7.0)	14.2 (11.0)	17.8 (13.0)	18.7 (16.0)	20.6 (14.0)	18.6 (14.5)	22.3 (17.0)	23.1 (17.0)
2nd			36.3 (26.5)	35.5 (29.0)	49.2 (40)	54.4 (43.5)	48.3 (35)	44.5 (34)
3rd					50.7 (25)	90.9 (69)	57.6 (41.5)	58.12 (28)
4th							90.2 (80)	65.9 (38)

Table 5: Average (median) decision times by treatment, difficulty and decision

For more formal evidence, we tested whether the distribution of thinking times for each difficulty and the answer round differed between AI and the human condition. Although we did not correct for multiple hypotheses testing, we did not

receive a single significant result on any of the ten tests. P-values ranged from 0.16 and 0.97 for Kolmogorof-Smirnov tests and from 0.11 to 0.72 for the slightly more powerful Epps-Singleton tests. We conclude that efforts are extremely unlikely to drive the better (worse) performance when playing with humans in two (four) step problems.

5 Experiment 2: Expert condition

Having established that humans reason differently when paired with other humans still leaves room for two different potential drivers of the differences. It is possible that just knowing that one plays with humans changes cognition. Alternatively, mere knowledge of playing with humans might not be sufficient. It might be necessary that cognition is only changed if a player knows that she faces other humans, who are in exactly the same position as themselves. In other words, the question is whether the driver of behavioral differences is the source of strategic uncertainty (AI vs. human) or the kind of strategic uncertainty (group members know the same amount vs. group members know how to solve the puzzle). An additional treatment, which we call the expert condition, in which subjects are paired with human experts who know how to solve the puzzle, allows us to discriminate between these two drivers.

We recruited an additional 144 subjects in eight sessions, who all played one Red-Hat puzzle (21 with a difficulty of one, 59 with a difficulty of two, 34 with a difficulty of three, and 30 with a difficulty of four reasoning steps). The sessions took place in August and September of 2023. All subjects were paired with three experts. It was made clear in the instructions (see Appendix 7) that the experts knew how to solve the Red-Hat puzzle. This was made credible by the experts being introduced with their credentials (like having a Ph.D. in economics) and by reminding subjects that economic experiments never use deception. The only difference the subjects experienced on their screens was that the other players were now named "Expert" instead of "Computer" or "Group Member".

5.1 Results

We ran the same probit regression as in the AI condition (see Appendix 9) and calculated the average predicted success probabilities for the different difficulty levels. Table 6 reports these success probabilities together with those estimated earlier for the AI and human condition. Inspection shows that the probabilities in the expert

condition are extremely similar to those in the AI condition. Z-tests reveal that they are not significantly different ($p > 0.55$, $p > 0.77$ and $p > 0.36$ for two, three, and four steps, respectively). The success probabilities differ significantly from those in the human condition for two steps and four steps. For two steps, playing with experts leads to lower success rates ($p < 0.001$). For four steps, the opposite is true. The posterior probability that the underlying success rate is at least as high in the human condition than in expert condition (if we start with flat priors for both conditions) is smaller than 0.02.

	Expert	**AI**	**Human**
1 step	1.00 (–)	1.00 (–)	1.00 (–)
2 steps	0.43 (0.06)	0.50 (0.07)	0.78 (0.04)
3 steps	0.21 (0.06)	0.19 (0.07)	0.25 (0.06)
4 steps	0.12 (0.05)	0.21 (0.06)	0.00 (–)

Table 6: Estimated success probabilities with standard errors in parentheses

Finding 2 *Subjects' success rates in the expert condition do not differ significantly from those in the AI condition, but are lower for two steps and greater for four steps than in the human condition.*

This indicates that the kind of strategic uncertainty rather than the source of strategic uncertainty drives differences in reasoning quality.

6 Robustness check

Before we interpret our results, we briefly present a robustness check we conducted. Recall that in the human condition we had to use a statistical procedure (i.e. a Heckman selection model) to correct for potential selection bias. As a robustness check that the observed behavioral differences are not an artifact of sample selection, we take the behavior of subjects in the two treatments without sample selection and simulate what we would have observed if the same sample selection issue had existed.

The sample selection issue in the human condition arises, as a subject's behavior might not be observed, as another subject in her group might make a mistake before she either makes a mistake herself or solves her puzzle. In the AI and expert conditions, this does not happen, as the other group members (computers or experts) do not make mistakes. For both treatments without sample selection issues, we randomly form 20,000 groups of observed human play and determine what would have happened if they had played together.[9] Comparing the simulation results in the AI and expert conditions with the observed results in the human condition can show if the observed behavior in the human condition can be generated by a combination of individual behavior as observed in the other treatments and selection.

Table 7 Shows the simulated and real fractions of observed subjects and the percentage of observed subjects who solved their puzzle correctly. We see that our result that playing with humans increases the performance in problems with two steps of reasoning is highly robust. The fraction of observed subjects that solve the puzzle is about 25 percentage points higher in the human condition than in the simulated groups with AI or experts.

		1 step	2 steps	3 steps	4 steps
Human	observed	100%	66.7%	44.9%	45.3%
	correct if obs	*100%*	*69.6%*	*6.5%*	*0.0%*
AI	observed	100%	65.1%	48.2%	45.1%
	correct if obs	*100%*	*44.7%*	*15.9%*	*8.2%*
Expert	observed	100%	77.0%	39.8%	56.9%
	correct if obs	*100%*	*45.1%*	*10.6%*	*8.8%*

Table 7: Outcomes of simulated groups in the AI and expert conditions and actual outcomes of human groups.

The other main result from above was that in the human condition, nobody solved the puzzles with four steps, while in the AI and the expert condition still a significant number of subjects managed to do that. Our simulations show that the fraction of observed subjects in level four problems who solve them is still positive, but below the fraction for all subjects. This implies that the result that playing with humans yields worse outcomes for four-step problems is potentially less robust.

[9] In the human condition there are four possible different group compositions differing by the distribution of hats. We simulated 5000 groups each.

Finding a statistical test that establishes differences between the data from the Human condition and the simulated data from the other treatments is tricky. A reasonable approach is to treat the success fractions in the simulations as underlying probabilities. It is then possible to test whether the observed successes and failures are likely to have been generated by such probabilities. The appropriate test for this is a two-sided binomial test.

For two steps, the Human data are very unlikely to have been generated by the success probabilities resulting from the simulations ($p < 0.001$ for both AI and Expert). This confirms our first main finding.

For three steps we cannot reject the hypotesis that the Human condition data were generated from the success probabilities obtained from the simulations ($p > 0.216$ AI; $p > 0.767$ Expert).

For four steps, where we did not observe a single success in the Human condition, we receive borderline significant results ($p < 0.111$ AI; $p < 0.071$ Expert).

Given these borderline test results, we investigate further. It might be the case that the way the subjects were paired in the human condition was just a bad draw. Maybe there exist quite a few possible bad draws when randomly pairing the subjects in the other two conditions, which also yield zero observed successes. In order to test the likelihood of this occurring, we ran further simulations. These further, more detailed simulations are designed to take into account correlations within groups, which are ignored by the tests conducted above.

In the human condition, we had 30 groups involving subjects who required four steps of reasoning to solve their puzzles. In 15 of these groups there were three red hats, and the person without a red hat required four steps. In the other 15 groups there were four red hats and all four subjects required four steps of reasoning. In the simulation we randomly filled these 30 groups with subjects from the two other conditions that faced exactly the same situation. We repeated this procedure 5,0000 times for both the expert and the AI condition, and counted how many sets of thirty groups would produce zero correctly solved level four problems. It is very unlikely that behavior under the expert condition would have led to the observation of zero successes in four-step problems. The fraction of simulated sessions with zero observed successes was 0.03. Observing zero successes was more likely in the AI condition, with the fraction being 0.13. At least in the AI condition the simulated fraction of zeros confirms the binomial test result. In order to further investigate whether introducing selection in the AI condition could have plausibly produced the

outcomes observed for four-step problems in the human condition, we compared the fraction of observed subjects between runs with zero and runs with positive success rate. The average fraction of observed subjects in the simulation runs with zero successes is 39.7% and therefore lower than the fraction of observed subjects in the human condition (45.3%), while the simulation fraction in runs where successes are observed is 46.1% and closely matches the observed rate. Hence, it is implausible that the behavior in the AI condition together with grouping could have led to both zero successes and a selection rate of 45% as in the human condition.

7 Discussion

In the Wason Selection Task (Wason, 1966), a famous logic puzzle in psychology, humans perform better if the task is framed in the sense of social relations and norms rather than abstractly. The abstract version of the task goes like this: "*You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show* 3*,* 8*, red and brown. Which card(s) must you turn over in order to test the truth of the proposition that if a card shows an even number on one face, then its opposite face is red?*"

Less than 10 percent of subjects get it right and turn over the 8 and the brown card. However, if one changes the context, then the number of people getting it right increases considerably. An observation related to our results is that most people get the task right if the context is about social rules and people like "People only drink alcohol if they are older than 21 years" (Cosmides, 1989). Canessa et al. (2005) show that the performance difference can be explained by the recruitment of additional parts of the brain (in the right hemisphere) for social exchange tasks. If this effect extends to our puzzle with humans, then ceteris paribus the Red-Hat Puzzle should be easier if played with humans.

Mental Model Theory, which is sometimes used to explain performance differences in different versions of the Wason Selection Task, defines reasoning as "the ability of humans to construct models from perception, description and knowledge to formulate novel but parsimonious conclusions from these models, and to grasp the force of counterexamples to these conclusions" (Johnson-Laird, 2006, p 249). This theory is particularly good at explaining the variations in frequencies of correct deductions, inductions, and abductions associated with different reasoning tasks. In

the Red Hat Puzzle, it is easy to construct the two necessary own models, which contain the hat colors of the other player, and the premise "there is at least one red hat" and the two possible own hat colors (one for each model). For the case where the subject sees only white hats, nothing else is necessary. It is easy to see for a subject that the model containing a white hat for the subject is inconsistent with the premise that there is at least one red hat. The conclusion "I have a red hat" is an immediate and simple deduction.

All puzzles where a subject sees at least one red hat require a subject not only to construct her own models, but also models about the models the other players might construct. This is hard. As humans have problems holding many different models in mind. It is also reasonable to expect that constructing a model of the model someone else might construct is easier when the other person is known to be similar to oneself. In other words, it should be easier to construct a model of the models constructed by a similar human than by a computer or an expert with a vastly better knowledge of the task. There are some studies (Baron-Cohen et al., 1985; Leslie and Thaiss, 1992, e.g.) showing that autistic children have difficulty with this kind of model construction, while they are able to solve equivalent tasks about maps and pictures. Moreover, making sense of another player's announcement requires an induction on an explanation (sometimes called an abduction). Humans are typically better at making sensible abductions if they have experience with the object (Johnson-Laird, 2006, chapter 14). In our puzzles with humans, a subject has to make an abduction on a human having faced the same problem as oneself, which should be more familiar than making sense of what a computer or robot-like expert has announced. In other words, putting oneself into the shoes of a similar human should be easier than putting oneself into the shoes of a computer or of an other human who differs significantly in knowledge from oneself. This should be the case as long as a problem is simple enough to hold the required models in ones mind.

Once the problem becomes harder, the strategic uncertainty becomes a real problem if subjects indeed use mental models. The number of models of the own mind and the models of the minds of others quickly surpasses the working memory of humans. In the case of puzzles with difficulty four, the number of mental models that a subject needs to hold in the working memory to solve the puzzle is so large that the authors were not able to enumerate them all. Therefore, one would not expect anyone using mental models to solve such a puzzle in the human condition.

This is what we observe. This observation allows for a model-theoretic explanation for why strategic uncertainty is so damaging in difficult puzzles, regardless of the fact that nobody has an incentive to move away from the logically correct solution path.

In contrast, playing with a computer might activate a different way of thinking. Instead of trying to put oneself in the shoes of the other players (i.e. building mental models of the situation as seen by other players), humans might search for a rule or algorithm. The correct algorithm for the red hat puzzle is as follows: If you do not see any red hats, announce "I have a red hat." Otherwise anounce m times "I cannot possibly know," where m is the number of red hats you see. Then observe the last announcement of the player(s) who have a red hat, announce "I have a white hat," if they declared red, and "I have a red hat" otherwise. The algorithm is difficult to discover, but works for any level difficulty. The fact that we observe no drop or only a modest drop in the success rate between difficulty three and four in AI and expert condition suggests that successful subjects in the more difficult puzzles in these conditions might have used algorithm-based reasoning. The absence of a drop in success rates between difficulty three and four problems is confirmed within subjects by a related experiment with AI (Bayer and Renou, 2016b), where subjects play the seven situations.

8 Conclusion

Our results are of importance for theorists and AI designers alike. A general insight from our study is that knowing how people solve problems interacting with humans cannot easily be extrapolated to how good they will be at solving problems with artificial intelligence. AI designers should keep this in mind. Theorists interested in developing formal behavioural models can learn that bounded rationality depends not only on the difficulty of a task but also on who we interact with.

Our specific results on the comparison of reasoning quality conditional on who humans interact with is less general. To achieve clean identification, we used a very specific task and a very stylized version of AI. The Red-Hat Puzzle is solvable by round-based reasoning, where previous reasoning steps of the team members are used as input. Therefore, the reasoning performances of individuals are complements, and all members of the group have to reason correctly for a successful solution. We cannot be sure that our findings generalize to tasks where individual performances

are substitutes and the good performance of a team member can compensate for the poor performance of another. It is possible that the change in cognition depending on whether someone interacts with humans or an algorithm is specific to tasks with complementarities.

In our experiments, AI consisted of a deterministic algorithm that always chooses the logically correct option. In reality, most AI tools behave probabilistically and sometimes make mistakes. Behavior in the AI condition is based on participants knowing that their counterparts do not make mistakes. This knowledge eliminates doubts of participants about the ability of computer players. Consequently, our AI condition gives the subjects the best possible environment to succeed. This has important implications for the generalizability of our results to the interaction between humans and AI in real life. Our result that interacting with humans yields better reasoning in relatively easy problems is likely to generalize as reasoning in the AI condition was inferior despite the environment being more favorable than in reality. Our second result is less likely to generalize. The relatively good performance of participants in difficult problems when paired with computer might disappear once they suspect that the computers make mistakes.

Finally, our results have implications for the methodology in experimental economics and finance. The standard technique in experimentation of replacing some human players by algorithms (e.g. computer buyers or traders) can be problematic, as this might change the cognition of human subjects, even if the computer players behave exactly as humans would.

References

Akiyama, E., N. Hanaki, and R. Ishikawa (2017, oct). It is not just confusion! strategic uncertainty in an experimental asset market. *The Economic Journal 127*(605), F563–F580.

Aldrich, E. M. and K. L. Vargas (2019, mar). Experiments in high-frequency trading: comparing two market institutions. *Experimental Economics 23*(2), 322–352.

Andreoni, J. and J. Miller (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica 70*(2), 737–753.

Ashktorab, Z., Q. V. Liao, C. Dugan, J. Johnson, Q. Pan, W. Zhang, S. Kumaravel,

and M. Campbell (2020, oct). Human-AI collaboration in a cooperative game setting. Volume 4, pp. 1–20. Association for Computing Machinery (ACM).

Bao, T., E. N. an Tibor Neugebauer, and Y. E. Riyanto (2022). Algorithmic trading in experimental markets with human traders: A literature survey. In S. Füllbrunn and E. Haruvy (Eds.), *Handbook of Experimental Finance*, pp. 302–322. Edgar Elgar.

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

Bayer, R.-C. and L. Renou (2016a, oct). Logical abilities and behavior in strategic-form games. *Journal of Economic Psychology 56*, 39–59.

Bayer, R. C. and L. Renou (2016b, jun). Logical omniscience at the laboratory. *Journal of Behavioral and Experimental Economics 64*, 41–49.

Bonner, S. E., R. Hastie, G. B. Sprinkle, and S. M. Young (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research 12*(210179775), 19–64.

Buczynski, W., F. Cuzzolin, and B. Sahakian (2021, apr). A review of machine learning experiments in equity investment decision-making: why most published research findings do not live up to their promise in real life. *International Journal of Data Science and Analytics 11*(3), 221–242.

Camerer, C. F. and R. M. Hogarth (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty 19*, 7–42.

Canessa, N., A. Gorini, S. F. Cappa, M. Piattelli-Palmarini, M. Danna, F. Fazio, and D. Perani (2005). The effect of social content on deductive reasoning: An fmri study. *Human Brain Mapping 26*(1), 30–43.

Cao, T., R.-R. Duh, H.-T. Tan, and T. Xu (2021, jul). Enhancing auditors' reliance on data analytics under inspection risk using fixed and growth mindsets. *The Accounting Review 97*(3), 131–153.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

Chen, C. X., R. Hudgins, and W. F. Wright (2021, jun). The effect of advice valence on the perceived credibility of data analytics. *Journal of Management Accounting Research 34*(2), 97–116.

Commerford, B. P., S. A. Dennis, J. R. Joe, and J. W. Ulla (2021, dec). Man versus machine: Complex estimates and auditor reliance on artificial intelligence. *Journal of Accounting Research 60*(1), 171–201.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition 31*(3), 187 – 276.

Dargnies, M.-P., R. Hakimov, and D. F. Kübler (2022). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *SSRN Electronic Journal*.

Das, R., J. E. Hanson, J. O. Kephart, and G. Tesauro (2001). Agent-human interactions in the continuous double auction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, San Francisco, CA, USA, pp. 1169–1176. Morgan Kaufmann Publishers Inc.

Dietvorst, B. J. and S. Bharti (2020, sep). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science 31*(10), 1302–1314.

Dietvorst, B. J., J. P. Simmons, and C. Massey (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General 144*(1), 114–126.

Efendić, E., P. P. V. de Calseyde, and A. M. Evans (2020, mar). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes 157*, 103–114.

Filiz, I., J. R. Judek, M. Lorenz, and M. Spiwoks (2022, aug). Algorithm aversion as an obstacle in the establishment of robo advisors. *Journal of Risk and Financial Management 15*(8), 353.

Fischbacher, U. (2007). Z-tree - Zurich toolbox for readymade economic experiments. *Experimental Economics 10*(2), 171–178.

Gao, X., R. Gong, Y. Zhao, S. Wang, T. Shu, and S.-C. Zhu (2020). Joint mind modeling for explanation generation in complex human-robot collaborative tasks. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1119–1126.

Geraghty, R., J. Hale, S. Sen, and T. S. Kroecker (2020). Fun-agent: A 2020 HUMAINE competition entrant. In *Proceedings of the 1st International Workshop on Multimodal Conversational AI*, MuCAI20, New York, NY, USA, pp. 15–21. Association for Computing Machinery.

Germann, M. and C. Merkle (2019). Algorithm aversion in financial investing. *SSRN Electronic Journal*.

Gjerstad, S. (2007, may). The competitive market paradox. *Journal of Economic Dynamics and Control 31*(5), 1753–1780.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125.

Greiner, B., P. Grünwald, T. Lindner, G. Leitner, and M. Wierensperger (2022). Incentives, framing, and trust in algorithmic advice: An experimental study. *mimeo*.

Hanaki, N., E. Akiyama, and R. Ishikawa (2018, mar). Behavioral uncertainty and the dynamics of traders' confidence in their price forecasts. *Journal of Economic Dynamics and Control 88*, 121–136.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), pp. 153–161.

Hertwig, R. and A. Ortmann (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences 24*(3), 383–403.

Hertz, N. and E. Wiese (2019, sep). Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied 25*(3), 386–395.

Johnson-Laird, P. (2006). *How we reason*. Oxford University Press.

Jung, M. and M. Seiter (2021, sep). Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study. *Journal of Management Control 32*(4), 495–516.

Kawaguchi, K. (2021, mar). When will workers follow an algorithm? a field experiment with a retail business. *Management Science 67*(3), 1670–1695.

Kellogg, K. C., M. A. Valentine, and A. Christin (2020, jan). Algorithms at work: The new contested terrain of control. *Academy of Management Annals 14*(1), 366–410.

Leslie, A. M. and L. Thaiss (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition 43*(3), 225 – 251.

Liang, C., J. Proft, E. Andersen, and R. A. Knepper (2019, may). Implicit communication of actionable information in human-AI teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Littlewood, J. E. (1953). *A Mathematician's Miscellany*. London: Meuthen & Co. Ltd.

Luca, M. D. and D. Cliff (2011). Agent-human interactions in the continuous double auction, redux - using the OpEx lab-in-a-box to explore ZIP and GDX. In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*. SciTePress - Science and and Technology Publications.

Maedche, A., C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner (2019, jun). AI-based digital assistants. *Business & Information Systems Engineering 61*(4), 535–544.

March, C. (2021, dec). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology 87*, 102426.

Memmert, L. and E. Bittner (2022). Complex problem solving through human-AI collaboration: Literature review on research contexts. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences.

Merritt, T. and K. McGee (2012). Protecting artificial team-mates: More seems like less. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA, pp. 2793–2802. Association for Computing Machinery.

Murray, A., J. Rhymer, and D. G. Sirmon (2021, jul). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review 46*(3), 552–571.

Peng, Y., J. Shachat, L. Wei, and S. S. Zhang (2020). Speed traps: Algorithmic trader performance under alternative market structures. *ESI Working Papers* (20-39).

Shrestha, Y. R., S. M. Ben-Menahem, and G. von Krogh (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review 61*(4), 66–83.

Tabrez, A., S. Agrawal, and B. Hayes (2019). Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 249–257.

Tai, C.-C., S.-H. Chen, and L.-X. Yang (2018, jun). Cognitive ability and earnings performance: Evidence from double auction market experiments. *Journal of Economic Dynamics and Control 91*, 409–440.

Van de Ven, W. P. M. M. and B. M. S. Van Praag (1981). The demand for deductibles in private health insurance : A probit model with sample selection. *Journal of Econometrics 17*(2), 229 – 252.

Wang, N., D. V. Pynadath, and S. G. Hill (2016). The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, Richland, SC, pp. 997–1005. International Foundation for Autonomous Agents and Multiagent Systems.

Wason, P. C. (1966). Reasoning. In B. M. Voss (Ed.), *New horizons in psychology*, Volume 1, Chapter 6, pp. 135–151. Penguin.

Weber, R. A. (2001). Behavior and learning in the dirty faces game. *Experimental Economics 4*(3), 229–242.

Wilson, H. J. and P. R. Daugehrty (2018, July). Collaborative intelligence: Humans and ai are joining forces. *Harvard Business Review*, 114–123.

A Selection equation in the human condition

	probit coefficients
Selection equation	
Red hat	-0.829^{**}
	(0.280)
Position in the group, 1 is the base	
2	-0.336
	(0.336)
3	-0.493
	(0.363)
4	-0.891^{*}
	(0.420)
Male	-0.256
	(0.209)
Number of males as group members, the base is 0	
1	0.575
	(0.389)
2	0.316
	(0.379)
3	0.651
	(0.379)
Constant	1.087
	(0.588)

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: The selection equation

B Probit regression in the expert condition

Table 9: Average marginal effects in the expert condition

	Expert condition
Prob$\{correct = 1\}$	
Step dummies (*2 steps* is base)	
3 steps	-0.228^{**}
	(0.0877)
4 steps	-0.309^{***}
	(0.0792)
Male	-0.0393
	(0.0755)
Step dummies (*Arts/Econ/business* is base)	
Engeneering/Science	0.296^{*}
	(0.126)
Law	-0.00847
	(0.104)
Medicine	0.284
	(0.250)
Other	-0.0961
	(0.123)
Controls (age, maths, control ok) not significant	
$Log\mathcal{L}$	-60.7079
N	123

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

C Screenshots - not for publication

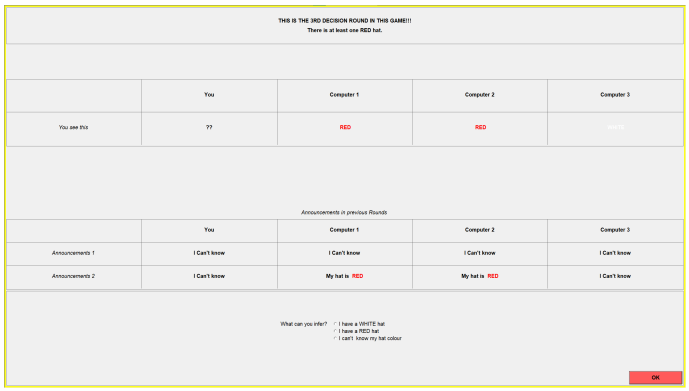


Figure 2: Screenshot from the AI condition

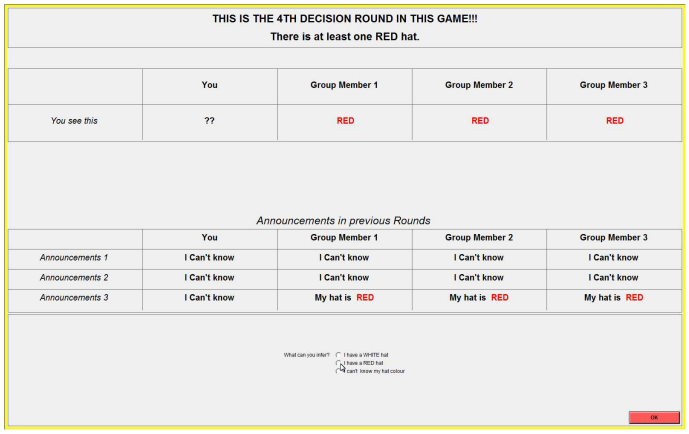


Figure 3: Screenshot from the human condition

D Instructions - not for publication

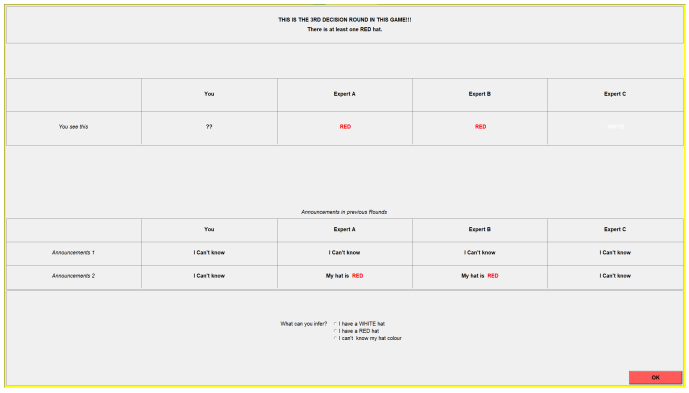


Figure 4: Screenshot from the human expert condition

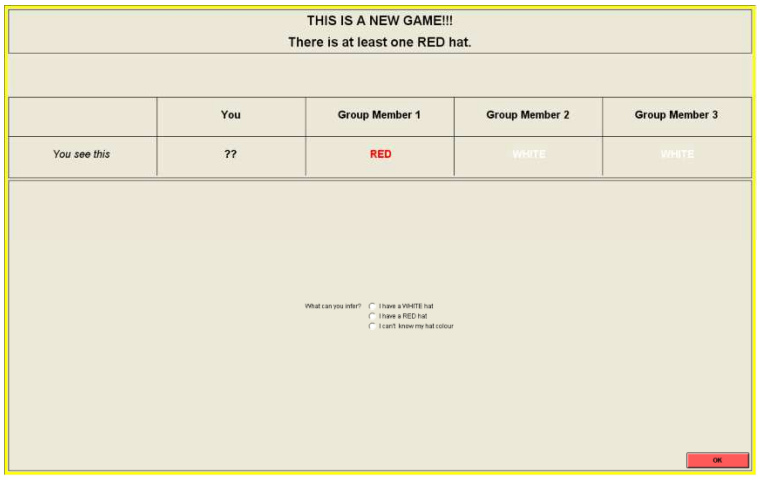
Figure 5: Instructions for the human condition

Instructions

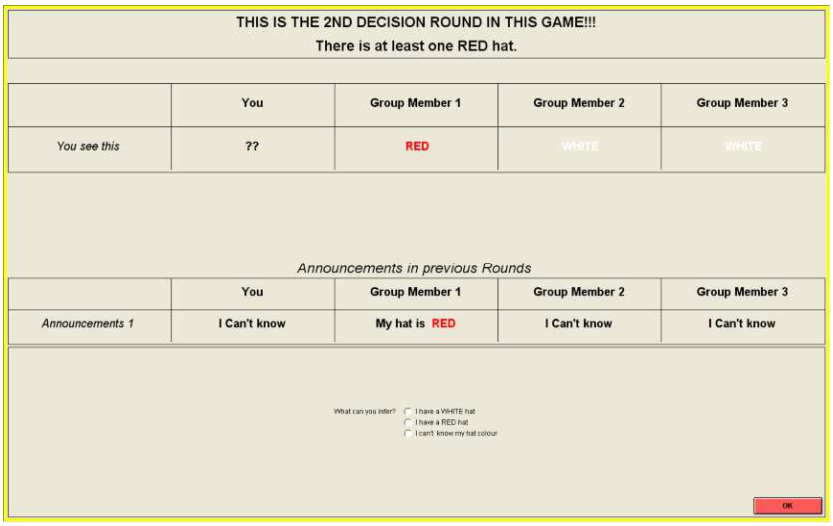
Thank you for your participation in this experiment. If you read these instructions carefully and act upon them, you can earn real money.

You are not allowed to communicate with other participants during the course of the experiment. If you do not follow this rule you may be excluded from the experiment.

Your task
Your task in this experiment is to determine the colour (red or white) of your hat. You will be paired with 3 other players. You will be able to see the colour of the hats of the other players, but **not** the colour of your own hat. The other players in your group are in an equivalent situation. They observe your hat colour and the hat colours of the other group members, but not the colour of their own hat. However, everybody knows (you and all other group members) that *atleast one player has a red hat*.The picture below shows a typical situation:



You observe in this case that one of the other players has a red hat while the other two have white hats. The question marks "??" indicate that you do not know your hat colour.
You are asked to decide what you can infer from the information you are given. The game may end after your initial decision. If you announced "I can't know my hat colour" the game may or may not continue. In the case where the game continues, you will be given the additional information on what the other players have answered from their observation in the round before. Recall that the other players face the same problem as you do. They can see the hats of all the others but not their own. Therefore, in the above situation, Group member 1 knows that the hats of Group members 2 and 3 are white, and also knows your hat colour. However, group member 1 does not know her/his own hat colour.



Above you can see a possible screen for your second decision. You again have to decide what you can infer about your hat colour. However, now you have the additional information about what the other group members announced in the decision round before. After you have made another decision, the game may end or continue. If the game continues, you will again be given the additional information of what the other group members inferred from the previous round. The other group members will get the same feedback.
Once you have decided on your hat colour and announced "RED" or "WHITE" you will not be asked to make further announcements. If a group member has not participated in the previous round, as he/she has announced a hat colour before, then this group member's previous decision will be indicated by "--".
Note that the game will always go on for long enough such that a group can get to the correct answer provided that all group members make logically consistent announcements.

Payment
You will play one of these games. If everyone in the group correctly inferred her/his hat colour then you will be put in the draw for a prize of AUD 300. If any group member gets the hat colour wrong or the game ends before everyone has decided on a hat colour then you will not participate in the draw. The draw will be conducted later this year, when the whole series ofexperiments has been conducted.

Introductory questions
Before you start the actual game, we will ask you some questions about the game. These questions will be designed to test if you understand the instructions. Please make sure to read the instruction very carefully, as failing to answer the pre-game questions correctly may lead to exclusion from the experiment.

Questions
Do you have any questions? If yes please raise your hand and we will come and answer them in private.

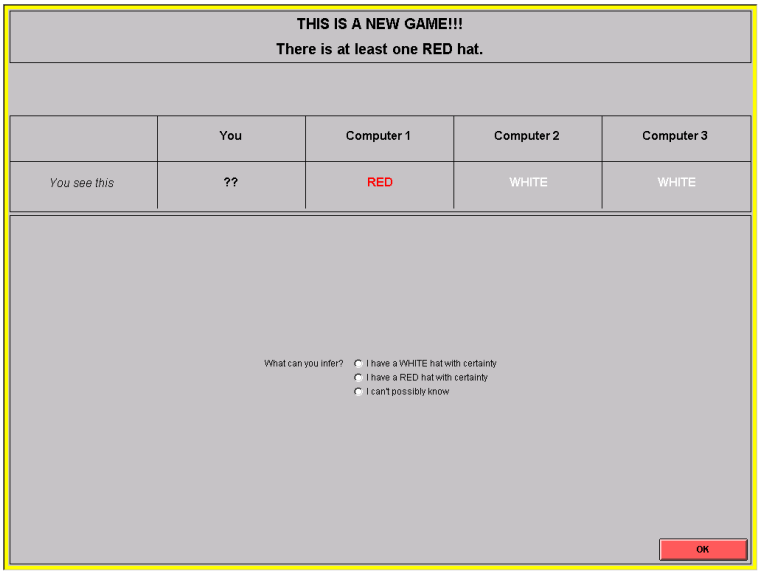
Figure 6: Instructions for the AI condition

Instructions

Thank you for your participation in this experiment. If you read these instructions carefully and act upon them, you can earn real money.

You are not allowed to communicate with other participants during the course of the experiment. If you do not follow this rule you may be excluded from the experiment.

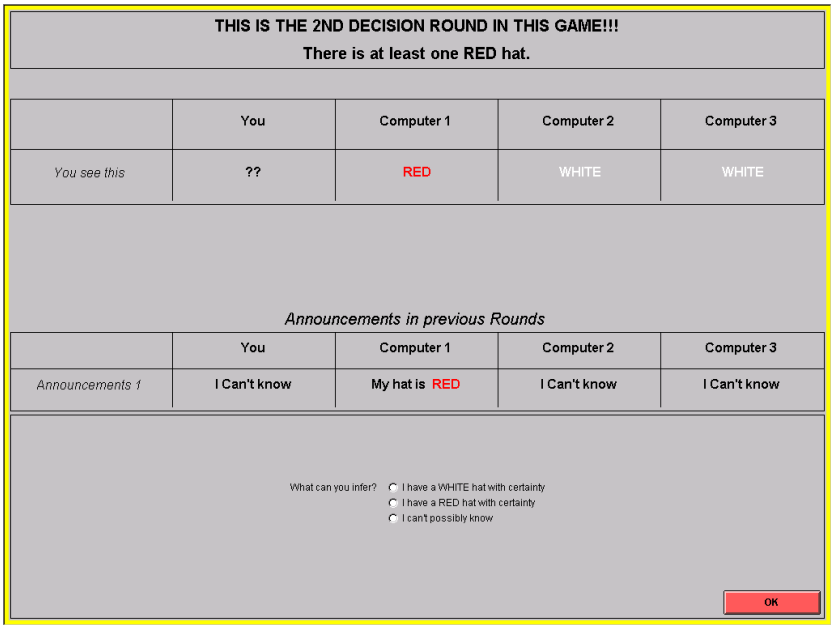
Your task
Your task in this experiment is to determine the colour (red or white) of your hat. You will be paired with 3 computer-players. You will be able to see the colour of the hats of the computer-players, but **not** the colour of your own hat. The computer-players are in a similar situation. They observe your hat colour and the hat colours of their fellow computer-players, but not the colour of their own hat. However, everybody knows (you and the computer-players) that *at least one player has a red hat*. The picture below shows a typical situation:



You observe in this case that one of the computer-players has a red hat while the other two have white hats. The question marks "??" indicate that you do not know your hat colour.
You are asked to decide what you can infer from the information you are given. Possible answers are: "I have a WHITE hat with certainty", "I have a RED hat with certainty", and "I can't possibly know". One of these answers is correct, the two others are wrong. Note that answering "I can't possibly know" is wrong whenever it is possible to correctly infer the hat colour from the information given. Similarly ticking "I have a WHITE hat with certainty" or "I have a RED hat with certainty" is only correct if it is actually possible to logically infer that your hat colour is white or red.

The game may end after your initial decision. If the game continues, you will be given the additional information of what the computer-players have inferred from their observation.

Recall that the computer-players face the same problem as you do. They can see the hats of all the others but not their own. Therefore, in the above situation, Computer 1 knows that the hats of computers' 2 and 3 are white, and also knows your hat colour. However, it does not know its own hat colour. Consequently, the computers also have a logically correct answer to the question: what can you (Computer) infer about your hat colour? The computers ALWAYS choose the logically CORRECT answer.



Above you can see a possible screen for your second decision. You again have to decide what you can infer about your hat's colour. However, now you have the additional information about what the computers (correctly) announced in the decision round before. After you have made another decision, the game may end or continue. If the game continues, you will again be given the additional information of what the computers inferred from the previous round. This process will go on until you either correctly inferred your hat colour or until you made a mistake.

Payment

You will play one of these games. If you solve your puzzle and correctly determine your hat colour then you will be put in the draw for a prize of AUD 300. The draw will be conducted later this year, when the whole series of experiments has been conducted.

Introductory questions

Before you start the actual game we will ask you some questions about the game. These questions will be designed to test if you understand the instructions. Please make sure to read the instruction very carefully, as failing to answer the pre-game questions correctly may lead to exclusion from the experiment.

Questions

Do you have any questions? If yes please raise your hand and we will come and answer them in private.

Figure 7: Instructions for the human expert condition

Instructions

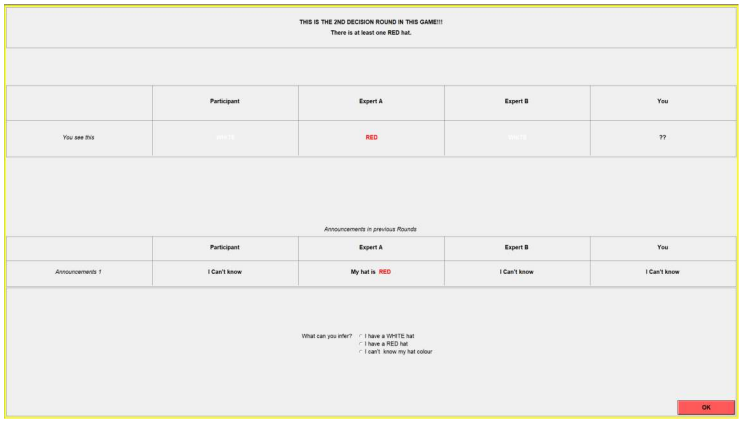
Thank you for your participation in this experiment. If you read these instructions carefully and act upon them, you can earn real money.

You are not allowed to communicate with other participants during the course of the experiment. If you do not follow this rule you may be excluded from the experiment.

Your task

Your task in this experiment is to determine the colour (red or white) of your hat. You will be paired with 3 other players which are Experts. These Experts are experienced with this decision task and fully understand how it works. You will be able to see the colour of the hats of the other players, but **not** the colour of your own hat. The experts in your group are in an equivalent situation. They observe your hat colour and the hat colours of the other group members, but not the colour of their own hat. However, everybody knows (you and all other group members) that *at least one player has a red hat*. The picture below shows a typical situation:

You observe in this case that one of the experts has a red hat, while the other two have white hats. The question marks "??" indicate that you do not know your hat colour.
You are asked to decide what you can infer from the information you are given. There is one logically correct answer in every possible situation. Recall that the experts face the same problem as you do. They can see the hats of all the others but not their own. Therefore, in the above situation, Expert A knows that the hats of Experts B and C are white. Expert A also knows your hat colour. However, Expert A does not know her/his own hat colour.



Above you can see a possible screen for your second decision. You again have to decide what you can infer about your hat colour. However, now you have the additional information about what the Experts announced in the decision round before. After you have made another decision, the game may end or continue. If the game continues, you will again be given the additional information of what the other group members inferred from the previous round. The Experts will get the same feedback.

Once you have decided on your hat colour and announced "RED" or "WHITE" you will not be asked to make further announcements. If a group member has not participated in the previous round, as he/she has announced a hat colour before, then this group member's previous decision will be indicated by "---".

Note that the game will always go on for long enough such that a group can get to the correct answer provided that all group members make logically consistent announcements. Also note that the Experts will make the logically correct announcement at any stage. Also note that all participants face the same starting situation, and hence all participants play with the same three Experts.

Payment

You will play one of these games. If you make the correct announcement at all stages and therefore correctly deduce your hat colour, then you will be put in the draw for a prize of AUD 300. If you make an incorrect announcement at any stage, then you will not participate in the draw. The draw will be conducted later this year, when the whole series of experiments has been conducted.

Introductory questions

Before you start the actual game, we will ask you some questions about the game. These questions will be designed to test if you understand the instructions. Please make sure to read the instruction very carefully, as failing to answer the pre-game questions correctly may lead to exclusion from the experiment.

Questions

Do you have any questions? If yes please raise your hand and we will come and answer them in private.