ETH Library

The Robotic Herd: Using Human-Bot Interactions to Explore Irrational Herding

Working Paper

Author(s):
Verginer, Luca; Vaccario, Giacomo; Ronzani, Piero

Publication date:
2023-09-21

Permanent link:
https://doi.org/10.3929/ethz-b-000645338

Rights / license:
Creative Commons Attribution-NoDerivatives 4.0 International

Originally published in:
SocArXiv, https://doi.org/10.31235/osf.io/ajusq

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

The Robotic Herd: Using Human-Bot Interactions to Explore Irrational Herding

Luca Verginer[1], Giacomo Vaccario[1] and Piero Ronzani[2]
[1]Chair of Systems Design, ETH Zurich, Switzerland;
[2]International Security and Development Center, Berlin, Germany

Abstract

We explore human herding in a strategic setting where humans interact with automated entities (bots) and study the shift in the behavior and beliefs of humans when they are aware of interacting with bots. The strategic setting is an online minority game, where 1,997 participants are rewarded for following the minority strategy. This setting permits distinguishing between irrational herding and rational self-interest—a fundamental challenge in understanding herding in strategic contexts. Moreover, participants were divided into two groups: one informed of playing against bots (informed condition) and the other unaware (uninformed condition). Our findings revealed that while informed participants adjusted their beliefs about bots' behavior, their actual decisions remained largely unaffected. In both conditions, 30% of participants followed the majority, contrary to theoretical expectations of no herding. This study underscores the persistence of herding behavior in human decision-making, even when participants are aware of interacting with automated entities. The insights provide profound implications for understanding human behavior on digital platforms where interactions with bots are common.

1 Introduction

Herding, a phenomenon studied in ethology, psychology, and economics, refers to the convergent social behaviour where individuals align their actions without explicit coordination (Raafat et al., 2009; Gavriilidis et al., 2016). This behaviour involves making choices based primarily on popularity rather than expected utility. Examples of intentional exploitation of herding include claques (individuals hired to applaud or heckle at theatre performances) and shill bidders (individuals inflating auction prices by placing fake bids in cahoots with the seller).

However, this inclination towards collective alignment is not always rooted in manipulative intent. Herding can be rational when the group has better information than the individual (Couzin et al., 2005). Consider the simple act of choosing a restaurant in an unfamiliar city: the sight of a bustling establishment, with patrons waiting outside, can be a compelling indicator of the restaurant's quality. Here, the individual assumes that the collective knowledge of the crowd surpasses their own limited information, leading them to align with the popular choice. Yet, this rational basis for herding becomes murkier in more complex scenarios, such as stock markets during economic bubbles. Investors, driven more by a fear of missing out and market sentiment (Indārs et al., 2019; Yarovaya et al., 2021) than by objective analysis, invest in popular stocks (Barber and Odean, 2007). This behaviour, although rooted in the perceived wisdom of the majority, can lead to inflated prices and result in devastating financial downturns when bubbles inevitably burst (Lux and Marchesi, 1999).

Stock market bubbles are examples of *irrational* herding, a fascinating phenomenon caused by flawed human decision-making. This phenomenon manifests when the desire to fit in or avoid missing out overwhelms logical judgment. It can be attributed to psychological factors such as social proof and cognitive biases, which lead individuals to rely on readily available information, in this case, the collective actions of others, rather than conducting their comprehensive analysis. Therefore, there is stark difference between rational and irrational herding: while the former is observationally equivalent to rational behavior, the latter showcases how human heuristic guiding decision-making can be costly.

Both rational and irrational herding have a long history in behavioural research, reaching back to the conformity experiments by Asch (1956). Depending on the alleged mechanism driving this collective phenomenon, it is known under different names, such as social proof (Cialdini, 2006), conformity (Asch, 1956), information cascades (Anderson and Holt, 1997) and social learning Burton-Chellew et al. (2017). This conformity is considered positive as it reduces conflicts (Becker et al., 2019) and increases accuracy (Mannes et al., 2014) but may (Lorenz et al., 2011) or may not undermine the wisdom of crowds (Davis-Stober et al., 2014). This observation raises societal concern to understand when herding is beneficial for the emergence of cooperation and its sustainability.

One of the inherent challenges in studying herding lies in the complex feedback loop between individuals and groups. An individual's decisions influence the collective group behaviour, and in turn, this group behaviour impacts individual choices. This intricate interplay makes it challenging to discern the relative importance of individual versus collective influences.

This challenge in studying collective behaviour is not unique to humans. Ethologists encounter similar issues when studying animal behaviour and have developed methods to isolate the impact of the group on the individual. For example, Oscar et al. (2023); Sridhar et al. (2021) let animals, such as zebrafish and locusts, interact with computer-simulated conspecifics. By manipulating the behaviours of these virtual counterparts, researchers measure their impact on the live animals, shedding light on the link between individual and collective behaviour.

Building on this approach, our study uses scripted automated player—*bots* for short—to investigate how group behaviour influences individual choices and remove one side of the feedback loop: the individual on the group. To test for cooperative herding behaviour, we conduct a large-scale online experiment where participants played a minority game with bots. Moreover, by selectively informing half the participants about the automated nature of their opponents, we study the effect of perceived opponent nature (i.e., bot or human) on herding tendencies and the beliefs about their opponents. We refer to these two conditions as C1 (informed) and C2 (not informed). Participants in the not informed condition were debriefed after the experiment was concluded.

Using bots to study prosocial behaviour is an “interesting yet underutilized resource” (Nielsen et al., 2022). Indeed, humans attribute different degrees of anthropomorphism when interacting with robots (Jauch et al., 2022; van der Woerdt and Haselager, 2019). Moreover, Sandoval et al. (2015) investigating strategic interactions of humans playing against bots in ultimatum and prisoner’s dilemma games, find that cooperation with bots is lower, while reciprocity is unaffected. Also, Crandall et al. (2018) showed that in two-player repeated games (e.g., prisoner’s dilemma) bots interacting with humans achieve human-level cooperation. However, the effectiveness of using bots to increase cooperation is still debated (Oliveira et al., 2021). This leads us to our first research question: Is there excessive cooperation in the minority game and if so, does it change when opponents are known to be bots?

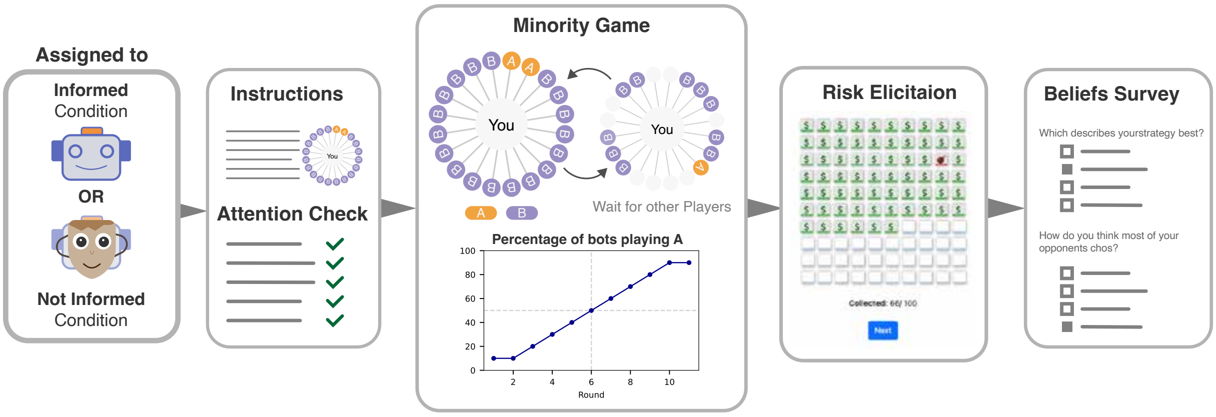

Figure 1: A flowchart of the five key stages of the experimental design. Panel 1: Random assignment of participants to either the informed condition (opponents referred to as "Bots") or the not informed condition (opponents referred to as "Players"). Panel 2: introduction and attention checks; Panel 3: a wheel of choices, denoting A for cooperation and B for defection, showing the choices of opponents (referred to as “Bots” in the informed condition, “Players” in the not informed condition), and the proportion of bot players choosing A; Panel 4: the Bomb Risk Elicitation Task (BRET) to assess risk propensity; Panel 5: the survey on participants’ beliefs about decisions and opponents. Note: The only difference between the conditions is the replacement of the word “Bots” with “Players”.

In our experiment participants engage in an iterated minority game or anti-coordination game. In this game, participants must choose whether to cooperate (C) or defect (D), observing the popularity of each choice in the preceding round. In the minority game, alignment with the majority is not the rational strategy for maximizing payoffs. This property is pivotal: If participants follow the majority, it is indicative of herding rather than rational self-interest. Hence, it allows us to distinguish costly herding from rational choice. This leads us to the second research question: Is there irrational herding, and does it change when opponents are known to be bots?

We further nuanced the game by introducing a 'bonus' system to mimic complexities found in real-world herding situations. Participants face a prisoner's dilemma game, where defection is the dominant strategy. They are also informed about a bonus—awarded only to cooperators for every defector—which transforms the game into a minority game. This layer of complexity is deliberate, challenging participants to discern the game's mechanics and encouraging them to derive cues from the choices of others.

Each participant plays the iterated minority game against 20 other players, unaware of the pre-set total of 11 rounds, see Fig. 1. The other players are scripted bots whose programmed decisions are independent of participants' choices. These bots gradually increase their cooperation, starting at 10% and culminating at 90% by the final round. Midway in round six, the bots' strategies are evenly split. Up to this switching point, cooperation is the optimal strategy; afterwards, defection is. We chose a high number of opponents, as Asch (1956); Bond (2005); Pereda et al. (2019) found that herding increases with group size. We also use dynamic visualizations to make it seem like real-time competition, showing bot decisions after participants choose (see middle of Fig. 1).

Post-gameplay, participants complete a risk elicitation task to control for potential confounding factors related to risk attitudes. We finally gather insights into their beliefs about their decisions and their opponents. A comprehensive breakdown of the experimental setting is detailed in the methods section.

Our study has two objectives. First, we probe how group dynamics, simulated through bots, influence individual choices. Second, we explore the nuances of human-bot interactions, investigating how awareness of competing against bots may change behaviour and beliefs.

Based on the above research questions, we pre-registered the following three hypotheses (Ronzani et al., 2022):

H1 Excessive Cooperation: Players will often play the sub-optimal strategy cooperate (C) even when the majority of the bots play C. We test this hypothesis separately in C1 and C2.

H2 Herding: The proportion of players following the majority is significantly larger than 0. In other words, there are players playing D (and C) when the majority of bots play D (and C). We test this hypothesis separately in C1 and C2.

H3 Human-bot interaction: players in C1 play more C than players in C2.

Assessing these hypotheses serves two purpose: it enriches our understanding of both human herding and human-bot interactions. Our results hold relevance for the architecture and governance of online platforms in a world becoming progressively digital and powered by artificial

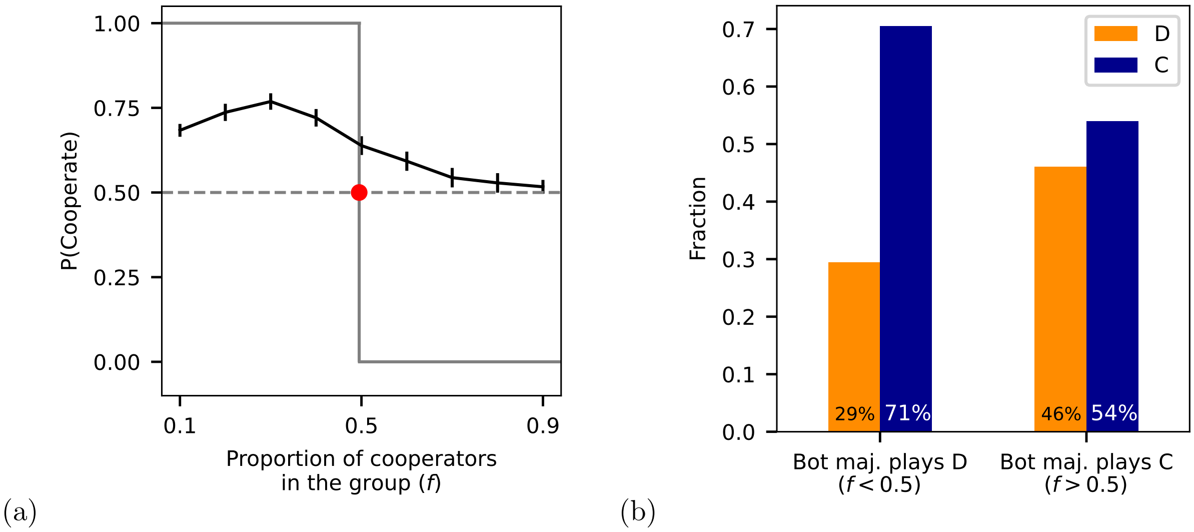

Figure 2: (a) The graph illustrates the likelihood of cooperation based on the percentage of bots that chose to cooperate in that round, represented by a black line. The error bars provide a 99% confidence interval. The grey line represents the ideal rational choice. The red point at P(C)=0.5 shows the anticipated response from both perfectly rational and boundedly rational agents when cooperation and defection yield identical payoffs (the switching point). The difference ($\Delta > 0$) between the black line and the red dot indicates a prevalent trend towards over-cooperation. (b) The graph showcases the fraction of players choosing C or D in response to the majority bot behavior. It reveals instances of irrational herding, where players make costly decisions to follow the majority bot choice, even when it is suboptimal in this minority game.

intelligence. For instance, bots are ubiquitous on platforms like Twitter (now known as X) and Reddit, wielding significant influence over public dialogue and individual user conduct. Therefore, the implications of our study may inform discussions and policy-making aimed at ensuring ethical and transparent human-bot interactions on digital platforms.

2 Results

2.1 Excessive Cooperation

In the context of the minority game—where selecting the less popular option results in greater rewards—standard economic theory suggests that players choose the less common strategy for the highest expected return. Over time, perfectly rational players are expected to transition from cooperation (C) to defection (D) when cooperation becomes more widespread. We use f to represent the proportion of bots that choose to cooperate.

Fig. 2 depicts the best response function for this game as a step function based on f: $P(C|f < 0.5) = 1$, that is, the majority chooses C, and $P = 0$ in other cases. However, it has long been recognized, since Simon (1955)'s groundbreaking work, that human decision-making is not perfectly rational. Bounded rationality offers a more flexible perspective, considering individual cognitive limitations and available information when making decisions.

Accordingly, in the bounded rationality framework, the probability of playing C transitions smoothly as a function of popularity rather than a step function. In a minority game, there exists a concept known as the 'switching point'. This is where options C and D are equally popular. This is illustrated by the red dot in Fig. 2. When the game reaches the switching point, we expect participants to choose C and D with equal probability.

If, instead, the probability of playing C exceeds $f = 0.5$ when there is no majority, it points towards excessive cooperation. This would imply that, overall, participants favour cooperation more than would be expected in a perfect or bounded rational scenario. Moreover, when there is a large majority playing C, according to bounded rationality, we would expect $P(C)$ to be significantly below 0.5. However, if it is significantly higher, the subjects might follow the majority decision, i.e., herding.

Our experiment provides evidence in line with these predictions. Confirming our first pre-registered hypothesis, we find that the average probability of cooperating at the switching point ($f = 0.5$) is 63.9%, significantly higher than the expected 50%, implying a preference for C over D. With a 99% confidence level, we can assert that the observed proportion is above 50%.

Also, in line with our predictions and confirming our second pre-registered hypothesis, a large fraction of the players follow the majority (herding) even if it is irrational (see Fig. 2 (b)). Precisely, when the majority of bots chose D, players still chose D 29% of the time. Likewise, cooperation persists even when it becomes the more costly choice. Despite rational or bounded rational expectations of no cooperation (0%), on average, 51.7% of participants still opted for cooperation, even when faced with a considerable majority of cooperators ($f = 0.9$). This pattern suggests that, once established, excessive cooperation can be maintained over short periods.

2.2 Risk Aversion and Decision-making

A plausible explanation for the excessive cooperation we observed could be risk aversion: participants, seeking to minimize potential losses or variability in outcomes, chose C as it offers lower variance in expected payoffs. If this were the case, the choice for C would not indicate cooperative intent but rather a strategy to mitigate risk.

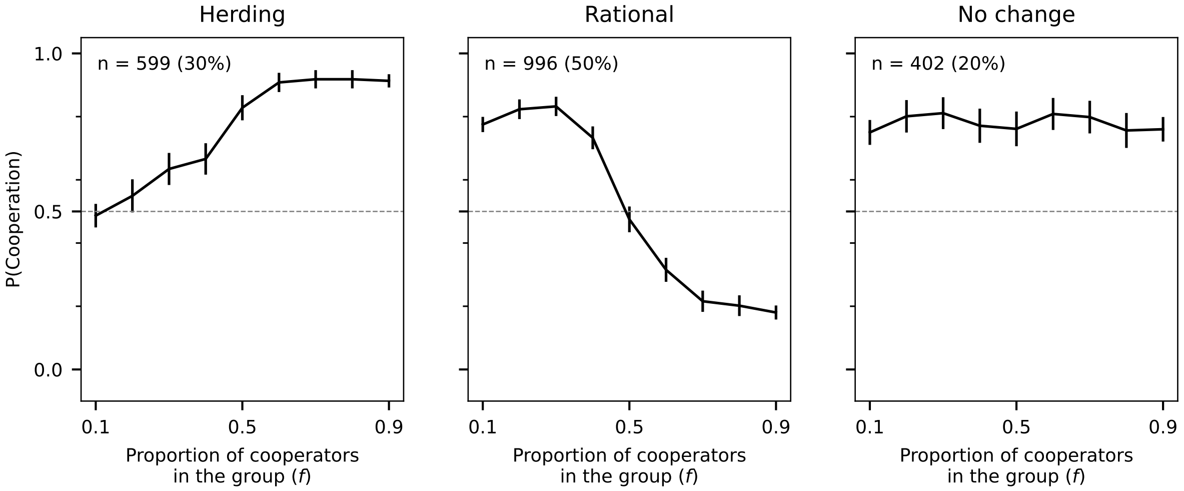

Figure 3: Illustrated here is the probability of cooperation, $P(C)$, plotted as a function of the number of cooperating bots from the preceding round. The left panel depicts the average $P(C)$ for scenarios where $P(C|f < 0.5) < P(C|f > 0.5)$, applicable to 30% of the participants. In the centre, the average $P(C)$ is shown for cases where $P(C|f < 0.5) > P(C|f > 0.5)$, representing 50% of the participants. On the right, the average $P(C)$ is displayed for situations where $P(C|f < 0.5) = P(C|f > 0.5)$ corresponds to the remaining 20% of the participants.

However, our data counters this interpretation. When examining participants' risk propensities (refer to the methods section for measurement details), we found no substantial correlation between risk aversion and the likelihood of choosing C. The correlation was slightly negative, as evidenced in Table 5. This suggests an intriguing conclusion: Participants with higher risk propensity were marginally more inclined to choose C.

In summary, while the choice of C can be seen as a risk-mitigating strategy, our findings indicate that risk aversion does not drive this behaviour. As observed in our experiment, excessive cooperation cannot be solely explained by participants' risk propensities.

2.3 Breakdown of strategy profiles

We propose 'herding' as a potential mechanism to explain the observed costly cooperation, i.e., cooperation even when the majority plays C. To test this, we look at the propensity to cooperate when the minority of other participants (bots) plays C, defined as $P(C|f < 0.5)$, and the propensity to cooperate when the majority plays C, defined as $P(C|f > 0.5)$. In other words, we examine the propensity to cooperate before and after the switching point ($f = 0.5$), the moment when the majority of other participants (in this case, bots) switch from C to D.

This enables us to deduce whether the participant follows the majority, indicating a "herding" behaviour, or the minority, suggesting a "rational" decision-making process.

To distinguish between herding and rational decision-making, we categorized participants into three types:

- **Herding**: Participants follow the *majority* more after the switching point, $P(C|f < 0.5) > P(C|f > 0.5)$
- **Rational**: Participants follow the *minority* more after the switching point, $P(C|f < 0.5) < P(C|f > 0.5)$.
- **No change**: Participant choose C irrespective of the proportion of cooperators, $P(C|f < 0.5) = P(C|f > 0.5)$.

Fig. 3 illustrates the distribution of participant types. Of 1,997 participants, 599 (or 30%) followed the majority, indicating herding behaviour. In contrast, 996 participants (or 50%) made what we classify as a "rational" choice, while the remaining 402 participants (or 20%) showed no discernible influence from the majority's choice.

Excessive cooperation is most noticeable at the switching point ($f = 0.5$), where herding participants cooperate with a probability of 82.8% and increase to 91.3% when the majority reaches $f = 0.9$.

On the other hand, Fig. 3 (b) illustrates that players identified as rational do not engage in excessive cooperation at the switching point. Their propensity to cooperate at the switching point of 47.5% is not statistically significantly different from 50% at the 99% confidence level. Therefore, we conclude that rational players are indifferent between the two options when no majority exists.

Turning to Fig. 3 (c), we see that participants who maintain a consistent cooperation probability throughout the game, no change, tend to cooperate more often. This observation corroborates prior findings on human pro-social behaviours. It also suggests that herding is not the sole factor contributing to excessive cooperation.

2.4 Differences across Bot and Not-Bot conditions

As part of our pre-registered analysis, we tested whether knowledge about the nature of opponents affects participants' decisions. In one condition, we omitted the information that participants were playing against bots, while in the other, we explicitly informed participants about it (see Sect. 4 for more details). Despite participants' awareness of interacting with bots, their decisions appear to be unaffected.

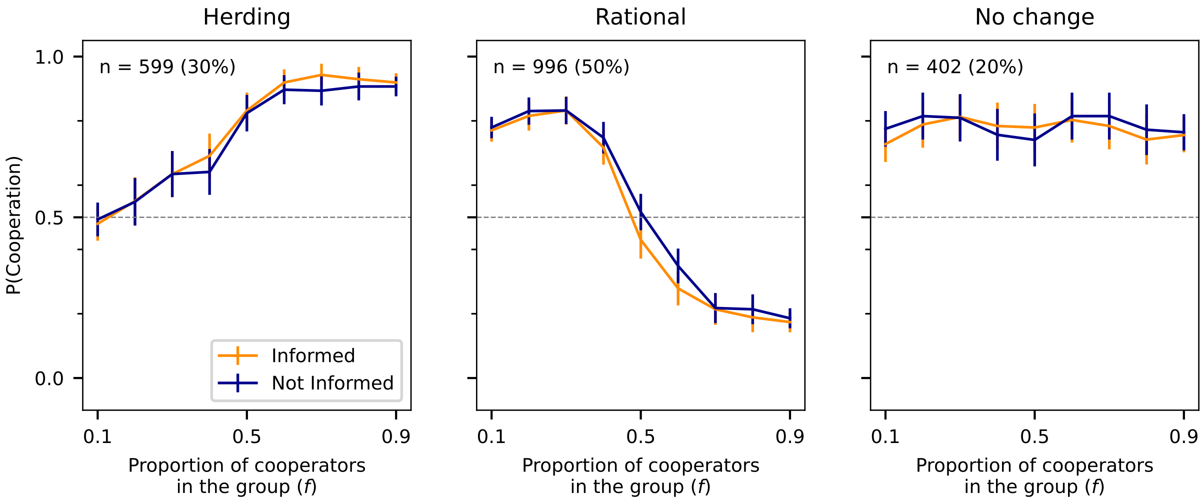

Figure 4: Probability of choosing C across conditions for different levels of bot cooperation (f): comparison between participants informed and uninformed about opponents being bots.

This null result is illustrated in Fig. 4, where we plot the probability of choosing C as a function of cooperators in the group. In all three plots, the 99% confidence intervals overlap across conditions providing a negative answer to our third hypothesis. On average, the difference across conditions was around 1% (See appendix Figure 12). This null result is also suggested by a Bayesian test of association yielding Bayes' factor below 1 (see Sect. 4 for further details).

It is worth noting that our experiment was sufficiently powered to detect a 10% effect size at a significance level of 0.01 (α) and a power of 0.95 ($1 - \beta$). This effect size, in our view and as stated in the preregistration, represents a meaningful level for detecting substantive behavioral differences between human-to-human and human-to-bot interactions.

Given this null result, we explore one other dimensions that may differ between the conditions. Specifically, we consider response time as a metric of attention and effort. Examining response time—the time it takes participants to complete a round—we found no credible difference across conditions. This suggests that attention and effort remained consistent whether participants knew they were playing against bots or humans (see Fig. 11 in the Appendix).

This null result might raise the concern that participants ignored the information about the nature of their opponents. As shown in the following sections, we ruled out this possibility by showing significant manipulation effects on beliefs.

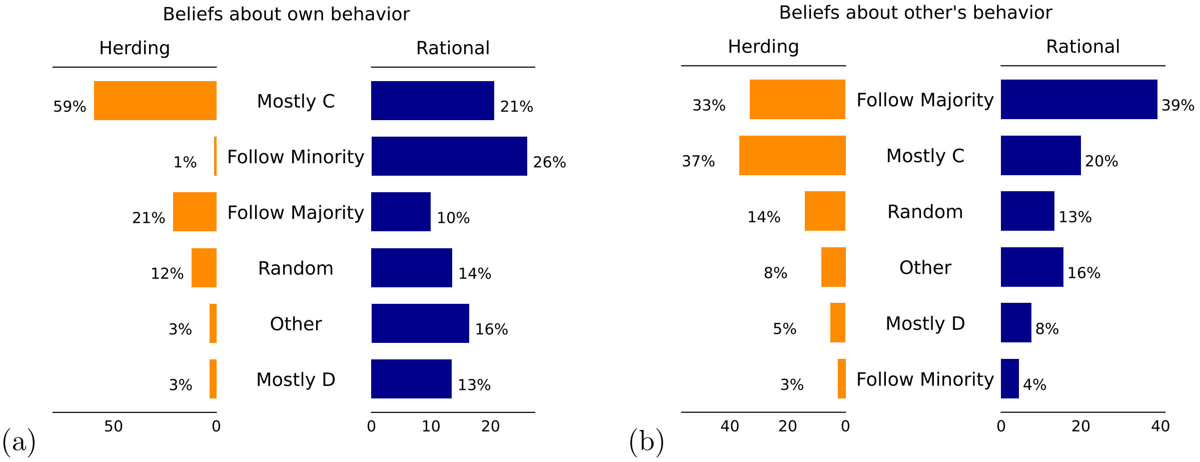

Figure 5: Results from participant surveys for those classified as herders (represented in orange) and rational participants (represented in blue). (a) Participants' responses to the question, "Which describes your strategy best?" (b) Participants' responses to the question, "How do you think most of your opponents chose?"

2.5 Beliefs about own and opponents' behaviour

We explore whether participants were aware of their own behaviour. In Fig. 5, we present responses to questions probing participants' perceived strategy and their beliefs regarding opponents' decisions.

Responses from the post-experiment survey reveal that 41% of all participants believed that they chose "Mostly C", suggesting that their actual behaviour — cooperating more than expected — aligns with their beliefs of their own actions. Moreover, 34% of participants believed that their opponents, who were bots, primarily "Follow(ed) the Majority", while a mere 4% thought they were following the minority. This points to participants' understanding of their opponent's strategies.

The second most common response was "Mostly C" (26%), although this is inaccurate as bots played C and D equally. However, "Mostly C" is a correct descriptor for the latter half of the game, suggesting that this belief may stem from a recency bias.

When analyzing "herding" participants — those who followed the majority — we find that 80% reported that they either "mostly played C" (59%) or "followed the majority" (21%) in their decision-making process. These reported beliefs align with their observed choices and a recency bias.

Congruence between beliefs and behaviour is also evident in participants who followed the minority — the rational strategy. The most common self-reported behaviours among these participants

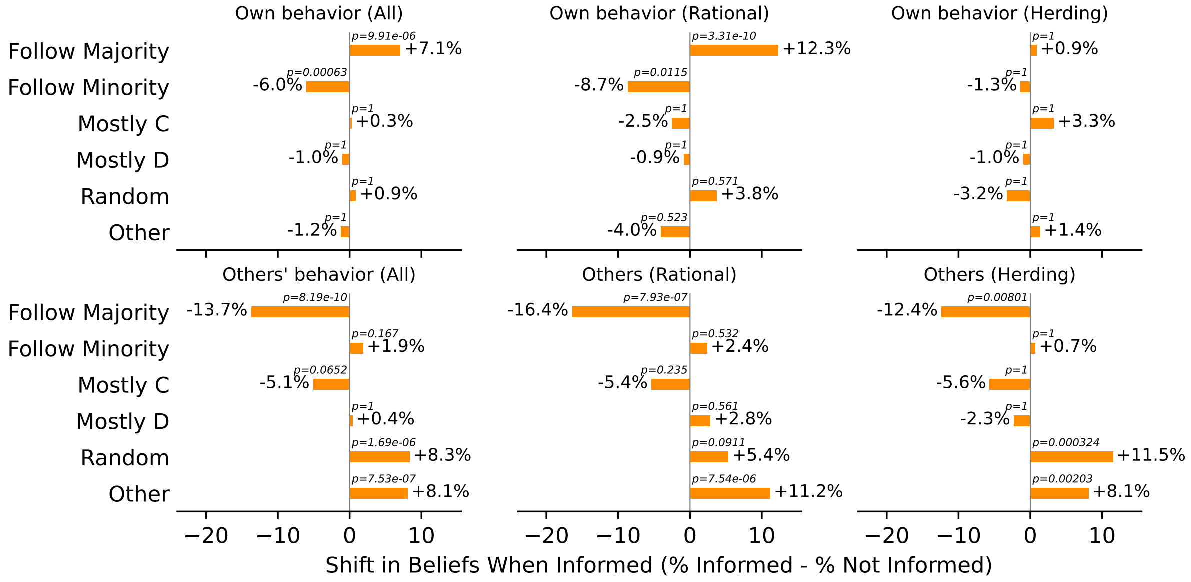

Figure 6: Multi-panel figure displaying the absolute changes in participants' strategy choices when informed they're playing against bots. Each bar represents the difference in percentage choosing each survey answer when informed versus uninformed. Positive (negative) values indicate an increase (decrease) in choice frequency of a specific answer. The top row examines beliefs of own actions, and the bottom row explores beliefs of opponents' actions. Columns from left to right represent shifts for all, rational, and herding participants, respectively. P-values above bars are Fisher's Exact Test results, adjusted for multiple comparisons (Bonferroni). For raw percentages and further details, see Figure 9 in the Appendix.

were "following the minority" and "mostly played C". Specifically, 26% reported "Follow Minority", a figure that increases to 31% when excluding participants whose behaviour aligns with bounded rationality. Furthermore, 21% of participants reported "Mostly C", a response that aligns with the observed excessive cooperation among relatively rational participants.

An intriguing divergence in participants' responses emerges when we consider whether participants were informed about their opponents being bots. In Fig. 6, we show the change in response to the final questions regarding the participants' beliefs about their own and their opponents' behaviour.

Rational players exhibited a significant shift in their beliefs about their own behaviour (top middle panel of Fig. 6). They believed they were more likely to follow the majority when informed about their opponents' nature (i.e., bot condition). One plausible explanation for this observation is that informed, rational participants believed their opponents to be more sophisticated as they were bots. Hence, when participants were asked about their behaviour, they answered that they were following the majority of their sophisticated opponents. Since the actual behaviour is statistically indistinguishable across conditions (see middle panel of Fig. 4), we observe that rational participants were less coherent when informed. For the other participants (i.e., herding and no-change), no significant shift in the beliefs of their own behaviour is observed.

Finally, participants' beliefs about their opponents' strategies did shift across conditions. Specifically, informed participants moved away from believing that opponents primarily "followed the majority", instead attributing more "random" or "other" strategies to them. This shift in beliefs is true across all strategy types (i.e., rational, herding and no-change). Despite this shift in beliefs, we find no significant change in decisions across conditions.

3 Discussion

We investigated herding — the phenomenon of following the crowd — within the framework of strategic games. We designed an online experiment where participants played a minority game with automated players (bots), allowing us to test for cooperative and herding behaviors. Our experimental design has a twist: We informed half of the participants that their opponents were bots, while the other half remained uninformed. This manipulation allowed us to explore the impact of knowledge about their opponents' nature on decision-making.

Contrary to theoretical predictions from game theory, the results reveal a robust preference for cooperation. Specifically, we observed that 64% cooperated even as 90% of their opponents cooperated, a costly decision in a minority game. In line with our pre-registered hypothesis, we found a high prevalence of herding behaviour (30% of the participants) despite a theoretical prediction of 0%. These results indicate a strong inclination towards conforming to the majority. Surprisingly, and against our expectations, this behaviour remained consistent even when participants were explicitly informed about their opponents' non-human nature.

Despite their awareness of interacting with bots, we found no credible evidence to suggest that participants changed their behaviour. Neither their cooperation nor herding behavior were affected. This consistency underscores the robustness of these behavioural traits.

An intriguing aspect of our findings centres on the shift in beliefs. When participants knew they were playing against bots, they changed their beliefs about their opponents' behaviour. Instead of thinking their opponents would follow the majority, participants assigned "random" or "sophisticated" strategies to the bots. Moreover, a specific type of participants (the more rational ones) also changed beliefs about their own behavior.

Despite these shifts in beliefs, there is little evidence that the participants' decision-making was altered. Their tendencies to cooperate and follow the herd (majority) did not change. In other words, altering the opponents' nature changes beliefs but not behaviour. Hence, the asymmetric effect highlights a fascinating difference between beliefs and behavior in humans.

Our findings have potentially significant implications for designing and regulating online platforms in our increasingly digital and AI-driven era. For instance, bots commonly operate on social media platforms like Twitter (now X), influencing public discourse and user behaviour. Our study suggests that awareness of interacting with bots does not alter the ingrained human tendency to follow the majority. This suggests that bots can still manipulate decisions even if marked as such.

Moreover, the rising integration of AI into our daily lives makes our findings even more relevant. Our research indicates that human decision-making does not significantly change in response to the perceived nature of an opponent, even if it is artificial. This could suggest that AI has the potential to induce similar human responses as those triggered by interactions with other humans. Also, it presents intriguing opportunities for AI design, particularly where cooperative decision-making is involved.

However, these are complex hypotheses requiring further investigation. Future research should explore how awareness of an opponent's artificial nature might impact human behaviour across diverse contexts, such as varying social media environments, or when the human-AI interactions are more sophisticated. The persistence of herding behaviour, even when faced with non-human opponents, underscores the urgent need for a nuanced understanding of human-AI interactions in strategic games and broader online interactions.

While the experiment provides valuable insights, it's also crucial to consider its limitations. Though carefully designed, the experimental setting is a simplified model of real-world social interactions. Online experiments, such as ours, offer anonymity that influences the applicability of the findings to real-world scenarios where social norms play a more substantial role.

Furthermore, our research primarily focused on how group behaviour influences individual decision-making without considering how individual decisions might, in turn, affect the group's behaviour. Understanding this interplay is an exciting avenue for future research.

In conclusion, our study offers a novel perspective on the robustness of human cooperative and herding tendencies within the context of strategic games, even when the nature of the opponents is known to be artificial. These findings contribute to a deeper understanding of the complex mechanisms driving human decision-making. They also underscore the need to further explore this phenomenon, particularly in online environments, where humans and bots interact regularly. Ultimately, the insights gained from this research could prove instrumental in informing the design and regulation of our increasingly AI-driven digital landscapes.

4 Methods

All studies were approved by ETH Zurich Ethics Commission approval number 2022-N-37. Participants in the study gave their informed consent beforehand.

4.1 Participants

Participants were recruited through Prolific and paid an average rate of £9.57 per hour. Participants' final earnings depended upon their choices within the experiment. All the participants were at least 18 years old and were residents of the UK at the time of the data collection. The sample provided by the recruitment platform was balanced on sex. Our target sample size was 2,000 participants, guaranteeing a power of 95% or higher to test the pre-registered hypotheses, see the pre-registration on OSF (Ronzani et al., 2022) and section "Sample Size and Power Analysis" below. After exclusion criteria and removing dropouts, the resulting sample is 1,997.

4.2 Experiment, setup

Our experiment adopts a 2 by 1 design, reflecting two conditions wherein the manipulated factor is the participants' awareness of their interactions with bots. Participants were randomly assigned to:

- **Bot Condition**: Participants are explicitly informed of their engagement with bots.
- **Human Condition**: The interaction with bots is concealed, referring to them simply as "players."

Note, the only practical difference across the two treatment conditions was the change of the word "Bot" to "Player", no other aspect of the game was altered.

Videos and screenshots of the game, as shown to participants, are available in the Supplementary Materials.

Instructions and comprehension checks During the initial phase of the experiment, all participants are presented with instructions and asked to complete comprehension questions regarding the task's rules and the calculation of payoffs. In the event of errors in the comprehension checks, participants are asked to check instructions and provide revised answers. They can advance to the next phase only when all the answers are correct.

Minority game After completing the comprehension checks, participants are told they are placed in an ongoing game where they know they will play multiple rounds. However, they are unaware of the exact number of rounds (set to 11). In each round, players observe the strategy played by other participants in the previous round, along with the payoff matrix. Participants choose between Strategy A and Strategy B for each round, corresponding to C "collaborate" or D "defect". We decided to use the labels A and B instead of C and D to avoid attributing any positive or negative connotations to the strategies the player can choose, where C "collaborate" may be perceived as a virtuous choice by some participants.

Participants played this game simultaneously against 20 other bot players and the payoffs are determined according to the following matrix, as presented to the participants:

$$\begin{array}{cccc} & \text{A} & \text{B} & \text{Bonus} \\ \text{A} & 6 & 0 & +4 \\ \text{B} & 10 & 0 & +0 \end{array}$$

In this matrix, the rows represent the player's strategy, and the columns represent the strategies of the 20 bot players. The 'Bonus' column indicates an additional payoff for a player choosing A, for every bot choosing B.

From a game theoretical perspective, this matrix is equivalent to the following simplified form:

$$\begin{array}{ccc} & \text{A} & \text{B} \\ \text{A} & 6 & 4 \\ \text{B} & 10 & 0 \end{array}$$

The entries in the matrix represent the payoffs to the player for each combination of strategies. The representation shown to the participants with the 'Bonus' column was used to make the strategic nature of the game more opaque and requires more effort to understand, thereby encouraging participants to get cues from the decisions of others. In each round, players observe the strategy played by other participants in the previous round, along with the payoff matrix, and choose between Strategy A and Strategy B.

Testing for H1 In the first pre-registered hypothesis, we aim to test whether excessive cooperation exists. To do this, we check the average probability of cooperating at the switching point ($f = 0.5$). We find that this probability is 63.9% and is higher than the expected 50%. To test whether this difference is significant, we show that the standard error (indicated via error bars) at a 99% confidence level does not include 50% (see Figure 2).

Testing for H2 The second pre-registered hypothesis is about irrational herding, i.e., about the proportion of participants following the majority even if it is costly. To check for this, we perform the following analysis. For each participant i, we compute two quantities: 1) the fraction of times the participant chooses C when the majority of bots played D ($P_i(C|f < 0.5)$), and 2) the fraction of times the participant chose C when the majority of bots played C ($P_i(C|f > 0.5)$). Then, we verify that the fraction of participants following the majority ($|\{i : P_i(C|f < 0.5) < P_i(C|f > 0.5)\}|/N$ where N is the sample size) is different from 0. This fraction is our estimate for the tendency to herd in the population. We find that this fraction is around 30%, implying that about 600 participants out of 2000 followed the majority even if it is costly.

Testing for H3 This section presents empirical tests for our third hypothesis (H3) and its sub-hypotheses (as listed in the pre-registration). Specifically, we test the following sub-hypotheses:

- **H3.1**: Players who are not informed about the nature of their opponents (C1) cooperate more than informed players (C2) ($\beta_2 < 0$).
- **H3.2**: The effect observed in H3.1 intensifies as the proportion of bots playing C increases. In other words, players in C2 will cooperate less when the fraction of bots playing C increases ($\beta_3 < 0$) compared to players in C1.
- **H3.3**: The proportion of herding players in C1 is higher than in C2.

To conduct these tests, we employ both regression analysis and a χ^2 test in accordance with our pre-registration.

The dependent variable in the regression analysis is the participants' decision to cooperate or not, represented as 1 or 0. The independent variables include the proportion of bots cooperating (referred to as f in the main text), ranging from $\frac{2}{20}$ to $\frac{18}{20}$, and whether the participants were 'Informed' or 'Not informed' about their opponents nature. Formally, the regression model is expressed as:

$$\begin{aligned}\text{Cooperate} = \beta_0 + \beta_1 \times \text{Prop. Cooperators} + \beta_2 \times \text{Not Informed} + \\ + \beta_3 \times (\text{Prop. Cooperators} \times \text{Not Informed}) + \text{controls}\end{aligned}$$

If $\beta_2 < 0$, we find support for H3.1, suggesting that players in C1 are more inclined to cooperate than those in C2. If $\beta_3 < 0$, H3.2 is supported, indicating that players in C2 are even less likely to cooperate when more bots cooperate.

	Random Effects Model	Pooled Model
(Intercept)	0.615^{***}	0.615^{***}
	(0.029)	(0.019)
Prop. Cooperators	-0.256^{***}	-0.256^{***}
	(0.015)	(0.016)
Not Informed	0.022	0.022
	(0.014)	(0.013)
Prop. Cooperators **and** Not Informed	−0.040	−0.040
	(0.020)	(0.022)
% Boxes collected	0.032	0.032^{*}
	(0.023)	(0.014)
Prior Approvals	0.021^{***}	0.021^{***}
	(0.004)	(0.003)
Negotiation experience: Na	0.009	0.009
	(0.019)	(0.012)
Negotiation experience: Yes	0.009	0.009
	(0.011)	(0.007)
R^2	0.033	0.031
Adj. R^2	0.033	0.030
Num. obs.	21967	21967

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Estimation Results for Factors Influencing Cooperative Behavior. The table presents logistic regression estimates from two models to identify determinants of the outcome variable 'Cooperate.' Both models include an interaction term between 'Prop. Cooperators' and 'Not Informed,' while controlling for 'Prior Approvals,' 'Negotiation Experience,' and 'Boxes collected'. The 'Random Effects Model' utilizes a random-effects logistic regression approach, and the 'Pooled Model' is based on a pooled logistic regression framework. The analyses are conducted on panel data, indexed by 'ParticipantID' and 'Round.'

Table 1 reports the results of the regression analysis. Neither β_2 nor β_3 are statistically different from zero. As such, we fail to reject the null hypotheses for H3.1 and H3.2, concluding that there is no significant difference in cooperative behaviors across the conditions.

To investigate H3.3, as stated in our pre-registration, we apply a χ^2 test to compare the proportions of cooperators and herding players across conditions. Figure 12 in the Appendix presents the differences and the associated p-values at each round. All p-values exceed 0.19, leading us to fail to reject the null hypothesis for H3.3 as well.

For a detailed explanation of the χ^2 test and Bayes factor t-test methods employed, please refer to the appendix section titled "Chi-Square and Bayes Factor Analysis for Comparing Cooperation Across Conditions".

Sample Size and Power Analysis We aimed for a total sample size of 2,000 participants, with an equal distribution between control and treatment groups. This size was chosen to reliably detect a minimum effect size of 10% with a significance level (α) of 0.01 and a statistical power of 0.95, thereby minimizing the risks of Type I and II errors.

The target effect size was based on the scenario where the probability of choosing C ($P(C)$) in the control group is 0.5, as this represents the most challenging condition to detect an effect. To validate our choice, we employed Stata's `power twoproportions` command. The power analysis confirmed that a sample size of 2,000 participants provides sufficient power to detect an effect size of at least 9.5% under the specified conditions, aligning with our pre-registration commitments. Figure 13 in Appendix graphically presents these findings. Note that when the $P(C)$ of the control group deviates from 0.5, the sensitivity for detecting smaller effect sizes increases.

Risk attitude To assess risk attitude, we employed the Bomb Risk Elicitation Task (BRET) (Crosetto and Filippin, 2013). In this task, participants decide how many boxes to collect out of 100, with one of the boxes containing a bomb. Earnings increase proportionally with the number of boxes accumulated, but if the box with the bomb is also collected, earnings become zero. The BRET requires minimal numeracy skills, avoids data truncation, and enables precise estimation of risk aversion and risk-seeking behaviour (Crosetto and Filippin, 2013). We used the number of boxes collected in BRET to estimate a Constant Relative Risk Aversion (CRRA) parameter that quantifies the individual's level of risk aversion.

Beliefs Once the minority game and the BRET are completed, participants are asked to express their beliefs regarding the strategy they employed during the minority game and the strategy used by the bots (or other players in the human condition). Participants choose between the following options, answering the question: "Which describes your strategy best?". With the following options: "I mostly played A (i.e., bonus choice)", "I mostly played B (i.e., no bonus choice)", "I followed the minority", "I followed the majority", "I randomly choose between A and B" or "I followed a different strategy".

Moreover, they are asked about their beliefs about their opponents: "How do you think most of your opponents chose?" with the analogous answers: "They mostly played A", "They mostly played B", "They followed the minority", "They followed the majority", "They randomly choose between A and B" or "They followed a different strategy".

Demographic Characteristics Table 2 provides an overview of the demographic characteristics of the 1,997 participants in the study. The data includes the total number of approvals and rejections, the approval rate, the age distribution, and the gender balance (coded as 1 for female). The participants' approval rate reflect how often their submissions were included in experiments, as they provided quality answers; this rate is high at 99.43%. The age of the participants ranges from 18 to 84, with a mean age of 39.13, and the sample is balanced with respect to gender.

	Count	mean	std	min	25%	50%	75%	max
Total approvals	1997	590.67	482.68	1	182	499	874	3072
Total rejections	1997	3.46	4.16	0	1	2	5	48
Approval rate	1997	99.43	1.02	94	99	100	100	100
Age	1997	39.13	13.13	18	29	37	48	84
Female	1997	0.50	0.50	0	0	1	1	1

Table 2: Descriptive statistics of participants

Acknowledgements

We express our gratitude to those who significantly contributed to our research. Prof. Frank Schweitzer deserves our thanks for his initial feedback on our research design and for providing the necessary funding. We extend our appreciation to the B.E.E. unit at IMT Lucca, with special mention to Prof. Ennio Bilancini, and to the C.E.E.L. research unit at the University of Trento, particularly Prof. Matteo Ploner, for their valuable input. We also thank Chiara D'archangelo, Ali Seyhun Saral, and Sergiu Burlacu for their assistance, as well as Maria Almudena Claassen for her continuous support. Lastly, we want to acknowledge the members of the Chairs of Systems Design for their constructive feedback. These combined efforts have been instrumental in the realization of this study.

Author Contributions

Conceptualization: L.V., G.V., P.R.; Methodology: L.V., G.V., P.R.; Software: L.V.; Formal Analysis: L.V., G.V., P.R.; Data Curation: L.V.; Writing — Original Draft: L.V., G.V., P.R.; Writing — Review & Editing: L.V., G.V., P.R.; Visualization: L.V., G.V., P.R.;

Competing Interests

The authors declare no competing interests.

References

Anderson, L. R. and C. A. Holt (1997). Information cascades in the laboratory. *The American Economic Review 87*(5), 847–862.

Asch, S. E. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological Monographs: General and Applied 70*(9), 1–70.

Barber, B. M. and T. Odean (2007, December). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies 21*(2), 785–818.

Becker, J., E. Porter, and D. Centola (2019, May). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences 116*(22), 10717–10722.

Bond, R. (2005, October). Group size and conformity. *Group Processes & Intergroup Relations 8*(4), 331–354.

Burton-Chellew, M. N., C. El Mouden, and S. A. West (2017, April). Social learning and the demise of costly cooperation in humans. *Proceedings of the Royal Society B: Biological Sciences 284*(1853), 20170067.

Cialdini, R. A. (2006). *Influence: The Psychology of Persuasion, Revised Edition.* New York: William Morrow.

Couzin, I. D., J. Krause, N. R. Franks, and S. A. Levin (2005, February). Effective leadership and decision-making in animal groups on the move. *Nature 433*(7025), 513–516.

Crandall, J. W., M. Oudah, Tennom, F. Ishowo-Oloko, S. Abdallah, J.-F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, and I. Rahwan (2018, January). Cooperating with machines. *Nature Communications 9*(1).

Crosetto, P. and A. Filippin (2013). The "bomb" risk elicitation task. *Journal of risk and uncertainty 47*, 31–65.

Davis-Stober, C., D. Budescu, J. Dana, and S. Broomell (2014). When is a crowd wise?

Gavriilidis, K., V. Kallinterakis, and I. Tsalavoutas (2016, December). Investor mood, herding and the ramadan effect. *Journal of Economic Behavior & Organization 132*, 23–38.

Indārs, E. R., A. Savin, and Á. Lublóy (2019, March). Herding behaviour in an emerging market: Evidence from the moscow exchange. *Emerging Markets Review 38*, 468–487.

Jauch, M., S. C. Rudert, and R. Greifeneder (2022, October). Social pain by non-social agents: Exclusion hurts and provokes punishment even if the excluding source is a computer. *Acta Psychologica 230*, 103753.

Lorenz, J., H. Rauhut, F. Schweitzer, and D. Helbing (2011, May). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences 108*(22), 9020–9025.

Lux, T. and M. Marchesi (1999, February). Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature 397*(6719), 498–500.

Mannes, A. E., J. B. Soll, and R. P. Larrick (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology 107*(2), 276–299.

Nielsen, Y. A., S. Pfattheicher, and M. Keijsers (2022, February). Prosocial behavior toward machines. *Current Opinion in Psychology 43*, 260–265.

Oliveira, R., P. Arriaga, F. P. Santos, S. Mascarenhas, and A. Paiva (2021, January). Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior 114*, 106547.

Oscar, L., L. Li, D. Gorbonos, I. D. Couzin, and N. S. Gov (2023, May). A simple cognitive model explains movement decisions in zebrafish while following leaders. *Physical Biology 20*(4), 045002.

Pereda, M., V. Capraro, and A. Sánchez (2019, April). Group size effects and critical mass in public goods games. *Scientific Reports 9*(1).

Raafat, R. M., N. Chater, and C. Frith (2009, October). Herding in humans. *Trends in Cognitive Sciences 13*(10), 420–428.

Ronzani, P., L. Verginer, and G. Vaccario (2022). The robotic herd: Using human-bot interaction to estimate the transition from rational choice to irrational herding. *Open Science Framework*.

Sandoval, E. B., J. Brandstetter, M. Obaid, and C. Bartneck (2015, December). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics 8*(2), 303–317.

Simon, H. A. (1955, February). A behavioral model of rational choice. *The Quarterly Journal of Economics 69*(1), 99.

Sridhar, V. H., L. Li, D. Gorbonos, M. Nagy, B. R. Schell, T. Sorochkin, N. S. Gov, and I. D. Couzin (2021, December). The geometry of decision-making in individuals and collectives. *Proceedings of the National Academy of Sciences 118*(50).

van der Woerdt, S. and P. Haselager (2019, August). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology 54*, 93–100.

Yarovaya, L., R. Matkovskyy, and A. Jalan (2021, November). The effects of a black swan event (COVID-19) on herding behavior in cryptocurrency markets. *Journal of International Financial Markets, Institutions and Money 75*, 101321.

A Appendix

A.1 Strict Strategy Definitions

To further explore participants' choices and highlight their heterogeneity, we introduce "strict" and "relative" variants of the previous classes.

- **Strict Herding**: Participants consistently follow the majority, indicated by $P(C|f < 0.5) < 0.5$ and $P(C|f > 0.5) > 0.5$

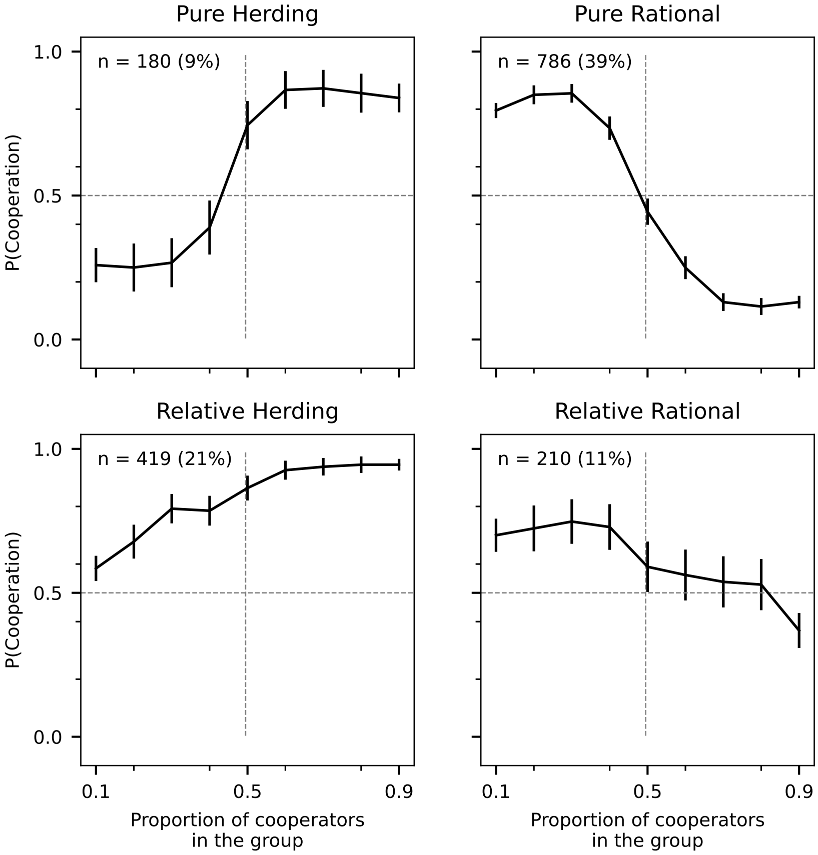

Figure 7: The probability of cooperating as a function of the number of cooperating bots in the previous round. On the top-left, we consider the average $P(C)$ for the strict herders. On the bottom-left, we consider the average $P(C)$ for the remaining herders. On the top-right, we consider the average $P(C)$ for the participants following the bounded rational strategy. On the bottom-right, we consider the average $P(C)$ for the remaining participants following the minority.

- **Relative Herding**: Participants who follow more the majority after the switching point, but do not meet the criteria for 'Strict Herding',
- **Strict Rational**: Participants who consistently follow the minority, indicated by $P(C|f < 0.5) > 0.5$ and $P(C|f > 0.5) < 0.5$
- **Relative Rational**: Participants follow more the minority after the switching point, but do not meet the criteria for 'Strict Rational'.

Fig. 7 discusses the relative proportion of these four variants. Out of 1,997 participants, 180 (or 9%) can be classified as strict herding participants and demonstrate an average cooperation probability of 30% prior to the switching point. Cooperation increases to 86% when the vast majority cooperates (see top-left panel in Fig. 7).

Looking at strict rational participants, we see a textbook example of bounded rationality (see top-right panel in Fig. 7). Initially, participants cooperate with a probability of approximately 80%. However, this probability sharply declines to less than 20% when the majority of bots begins to cooperate.

We observe that relative rational participants after the switching point still cooperate a lot (see bottom-right panel in Fig. 7). This means that they also contribute to excessive cooperation. Moreover, we observe that relative herding participants cooperate a lot before the switching point (see bottom-left panel in Fig. 7). This suggest that they start out with a high propensity to cooperate and follow the cooperation trend of the bots.

A.2 Beliefs about Own and Other's Behaviour

Understanding participants' self-perception and their views on their opponents' strategies is critical for interpreting the results of the minority game. In this section, we present a detailed breakdown of participants' answers to questions regarding their own and their opponents' perceived behavior.

Table 3 illustrates how participants perceived their own choices. From the table, it can be seen that participants often categorized their strategies in various ways, such as following the minority, the majority, or even playing randomly.

Table 4 provides an overview of how participants perceived the strategies of their opponents. This understanding is critical as it can give insights into how players anticipated and reacted to other's actions in the game.

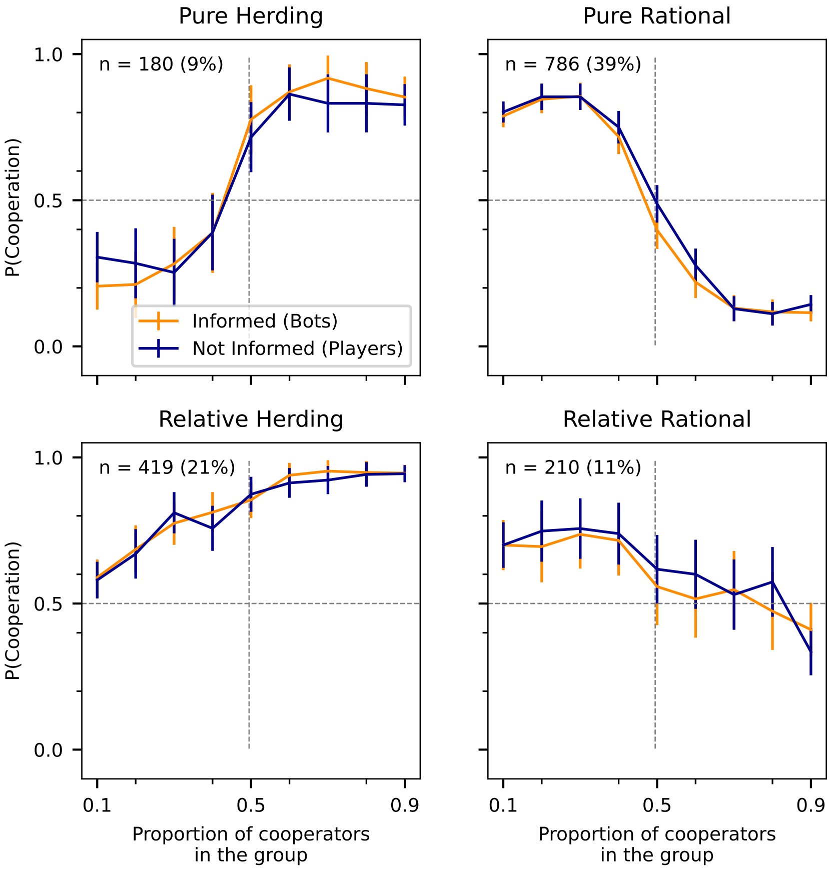

Figure 8: Probability to play C by condition the participants were assigned to: they were informed that their opponents are bots, and they were not informed during the game (they receive this information only in the debriefing stage).

	No change	Herding	Rational	Rel. herding	Rel. rational	All
Mostly D	5.22	5.00	12.98	2.39	15.24	8.71
Other	5.22	8.33	19.47	1.19	4.76	10.22
Follow Majority	5.22	38.33	11.32	13.60	4.76	12.32
Random	14.68	22.78	12.47	7.40	17.62	13.32
Follow Minority	3.23	2.22	30.53	0.48	9.52	13.97
Mostly C	66.42	23.33	13.23	74.94	48.10	41.46
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 3: Proportion of answers about own choices.

	No change	Herding	Rational	Rel. herding	Rel. rational	All
Follow Minority	2.49	2.78	4.83	2.63	2.86	3.51
Mostly D	7.71	5.00	6.23	5.25	12.38	6.86
Other	15.67	8.33	16.54	8.35	11.90	13.42
Random	23.63	15.56	11.58	13.37	20.00	15.62
Mostly C	25.62	27.22	18.07	40.81	27.14	26.14
Follow Majority	24.88	41.11	42.75	29.59	25.71	34.45
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 4: Proportion of answers about other choices.

A.3 Impact of Information on Participants' Perceptions of Behavior

Figure 9 provides a comparative view of participant beliefs about personal and others' behaviors under two distinct conditions: informed of playing against bots (gray bars) and uninformed (dark orange bars). Various strategic categories are identified on the y-axis, and the x-axis shows respective percentages.

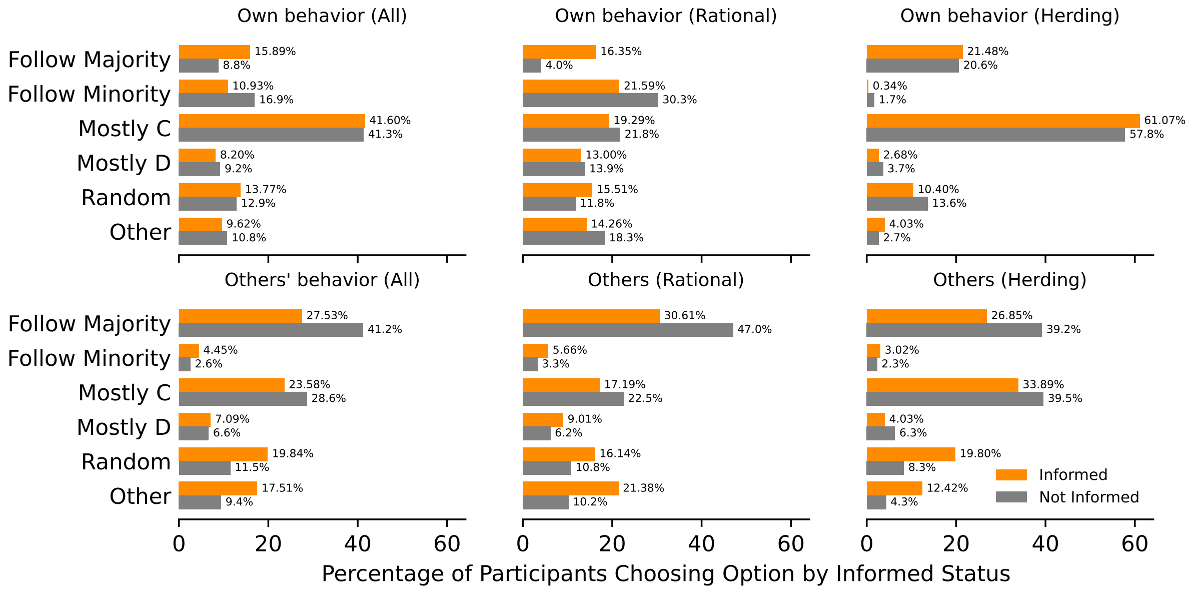

Figure 9: Comparison of participants' beliefs about personal and others' behaviors under informed and uninformed conditions.

The top row, about own behavior, reflects consistent personal behavior beliefs across conditions. Participants seem to maintain a stable understanding of their own strategies, regardless of whether they know they are playing against bots.

The bottom row reveals significant perceptual variations. When uninformed, participants more often thought their opponents were following the majority. Conversely, when informed that their opponents were bots, they attributed more "random" or "other" strategies to them.

These findings highlight the nuanced ways in which information about the nature of opponents can shape player perceptions and beliefs. While personal strategies remain consistent, the understanding of others' strategies shifts significantly, demonstrating the power of context and information in shaping strategic decision-making.

P-values displayed in Fig. 6 in the main text shown above the bars are derived from Fisher's exact test. They indicate the statistical significance between conditions for each strategy. To account for the multiple comparisons, we apply the Bonferroni correction to control the Type I error rate.

A.4 Time to Answer

The analysis of the time to answer provides insights into participants' decision-making processes during the game. It captures the time taken by each participant to make a decision, starting from the moment they see the final decisions and payoff from the last round. By examining the decision-making time across various rounds, we can gauge how participants' learning and effort evolved throughout the experiment.

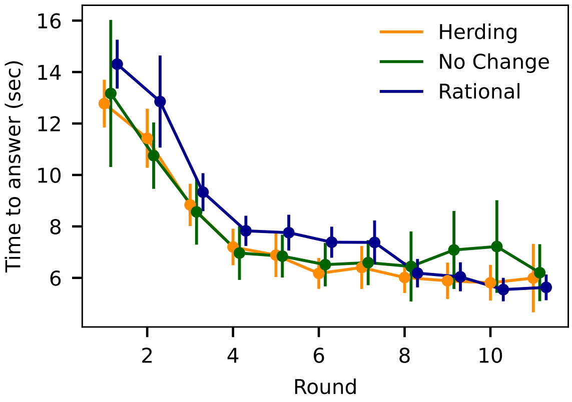

Figure 10: Average decision-making time for each participant per round, accompanied by a 99% standard error around the mean. This figure illustrates how the time to answer varies across different rounds of the game, reflecting the complexity and participants' engagement with the task.

The Fig. 10 represents the average decision-making time per participant for each round, providing a snapshot of the time needed to understand, evaluate, and act upon the information received. This data helps identify any potential learning effects or fatigue that might influence decision-making as the game progresses. However, given the overlapping 99% standard error, we can conclude that there is no credible evidence for differences across types.

The comparative analysis between the informed and uninformed conditions could help to identify systematic difference in effort or learning. Despite differences in awareness regarding the engagement with bots, we found little evidence of differences in the time taken to answer between the two groups, see Fig. 11

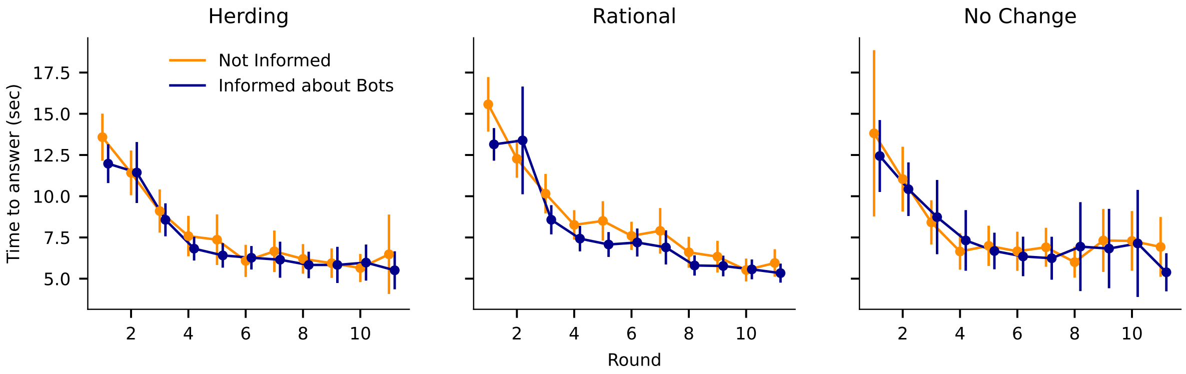

Figure 11: Average decision-making time for each participant per round, accompanied by a 99% standard error around the mean. The orange line denotes response times for participants unaware they were engaging with automated players (bots), while the blue line denotes those who were informed of this fact.

The lack of significant difference in response times between the two conditions suggests that knowledge of playing against bots may not substantially influence the immediacy or deliberation in decision-making. This aligns with the broader findings of the study, where awareness of the bot players did not dramatically alter the actual decisions made by the participants.

A.5 Probability to Choose Cooperation and Risk Attitude

We employed a linear probability regression model to investigate the factors influencing the probability of choosing Cooperation (C) among participants. Two separate models (Model 1 and Model 2) were constructed to examine different sets of explanatory variables.

Model Variables:

- **CRRA:** The Constant Relative Risk Aversion parameter obtained from the BRET task, included in Model 1 but not Model 2. It quantifies the participant's level of risk aversion.
- **Boxes Collected:** This variable, used in Model 2, represents the number of boxes collected by a participant in the BRET.
- **Male:** A binary variable indicating the gender of the participant.
- **Log(Age):** The natural logarithm of the participant's age.
- **Education:** Categorical variables representing different education levels of the participants.
- **Negotiation Experience:** A binary variable representing whether the participant has negotiation experience or not.
- **Charitable Giving:** Categorical variables capturing different ranges of charitable giving in British Pounds.

Model Summary: Model 1 includes CRRA as an explanatory variable, while Model 2 replaces it with Boxes Collected. Both models consider demographic and personal attributes like gender, age, education, negotiation experience, and charitable giving.

The R-squared values indicate the proportion of the variance in the dependent variable that is predictable from the independent variables, while the adjusted R-squared accounts for the number of predictors in the model. The number of observations for both models is 1997.

These models provide insights into the relationships between individual characteristics and the propensity to choose Cooperation (C) in the strategic game, offering a better understanding of how different factors may contribute to this choice.

	Probability to chose C	
	Model 1	Model 2
(Intercept)	0.365***	0.398***
	(0.069)	(0.067)
CRRA	0.013	
	(0.008)	
Coxes Collected		0.000
		(0.000)
Male	−0.027*	−0.026*
	(0.010)	(0.010)
Log(Age)	0.048**	0.050**
	(0.016)	(0.016)
Education		
Don't know / not applicable	−0.153	−0.151
Graduate degree (MA/MSc/MPhil/other)	0.018	0.017
High school diploma/A-levels	0.018	0.018
No formal qualifications	0.102	0.104
Secondary education (e.g. GED/GCSE)	0.025	0.025
Technical/community college	0.042	0.042
Undergraduate degree (BA/BSc/other)	0.039	0.038
Negotiation Experience		
No	0.009	0.009
Yes	0.006	0.005
Charitable Giving		
£1-£50	0.023	0.023
£50-£75	0.006	0.006
£75-£100	0.077***	0.077***
	(0.022)	(0.022)
£101-£200	0.021	0.022
£201-£500	0.008	0.009
£501+	0.033	0.034
R^2	0.024	0.023
Adj. R^2	0.015	0.015
Num. obs.	1997	1997

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5: Linear Probability Regression.

	Pr(C)	(SE)
(Intercept)	3.841	(0.037)***
Herding	0.102	(0.057)
Pure rational	−0.027	(0.039)
Relative herding	−0.015	(0.044)
Relative rational	0.018	(0.054)
No Change	−0.005	(0.028)
Male	0.046	(0.028)
R^2	0.004	
Adj. R^2	0.001	
Num. obs.	1997	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6: Linear Probability regression on the proportion/propbability of participant to choose C.

Effect across behavioural types When comparing the average risk propensities of participants across different behavioural types (e.g., herding), we found no credible evidence of differences, as shown in Table 6.

A.6 Chi-Square and Bayes Factor Analysis for Comparing Cooperation Across Conditions

As pre-registered, we employed χ^2 tests to assess whether the proportions of participants choosing to cooperate ($P(C)$) differed between the 'Informed' and 'Not Informed' conditions.

To complement our frequentist approach, we ran a Bayesian test of association using the R package `BayesFactor`.

From this test, we obtain a Bayes Factor (BF), which serves as a Bayesian counterpart to the p-value, allowing us to quantify the strength of evidence for the null hypothesis—that there is no difference in $P(C)$ across conditions.

Specifically, we obtain the contingency table for each cooperation level (f) for the two decisions (A and B) across the two conditions. For each contingency table (one per cooperation level), we compute first the p-value of the χ^2 using R and then the BF using the `contingencyTableBF` function from the *BayesFactor* package. Specifically, we assume default priors using a joint multinomial sampling plan (`indepMulti"`).

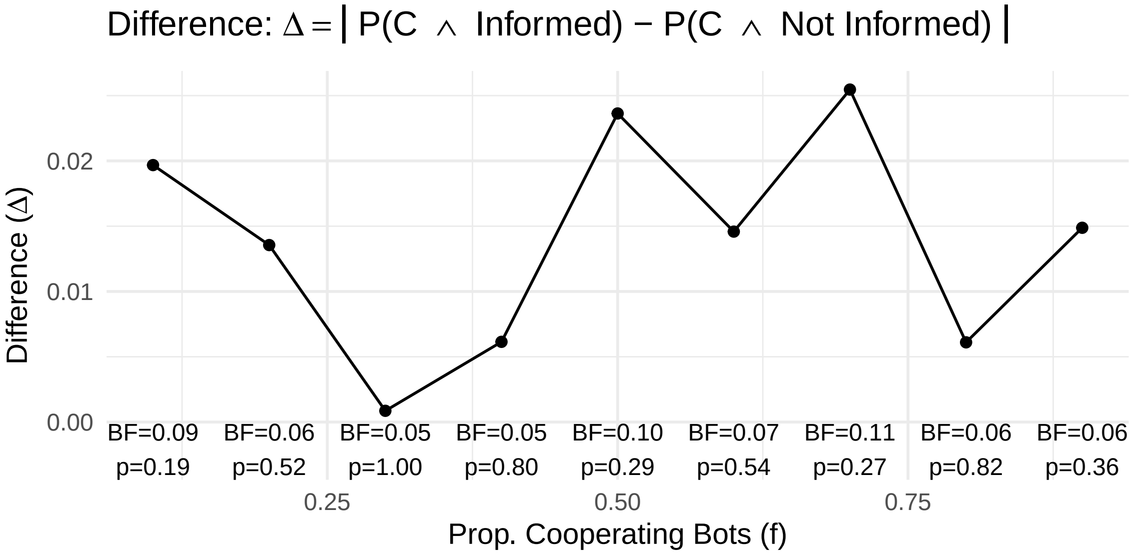

Figure 12: Difference in the proportion of participants choosing to cooperate, denoted as $P(C)$, between the 'Informed' and 'Not Informed' conditions across varying levels of bot cooperation. Annotations below each point indicate the p-value from the chi-square test and the Bayes Factor (BF), which provide measures of statistical significance and evidence strength, respectively.

We show the results in Figure 12. This figure plots the difference in $P(C)$ at different levels of bot cooperation. Below each point, we annotate both the corresponding p-value (χ^2 test) and BF. Importantly, our analyses did not find evidence for an effect on $P(C)$ across conditions. All p-values exceeded 0.19, and all Bayes Factors are substantially below 1.

Aggregate Analysis We extended our analysis to explore participants' behavior both before and after the 'switching point' (f = 0.5), which is the moment when 'Cooperate' ceases to be the optimal strategy. Consistent with our earlier findings, the data shows no significant effect on cooperation rates. Specifically, we observed only a negligible 1% difference in the rate of cooperation after the switching point (p-value of 0.16 and Bayes Factor of 0.06). Similarly, participants' behavior prior to the switching point also showed an inconsequential difference in cooperation rates—less than 1%—with a p-value of 0.36 and a Bayes Factor of 0.04.

A.7 Power Analysis Figure

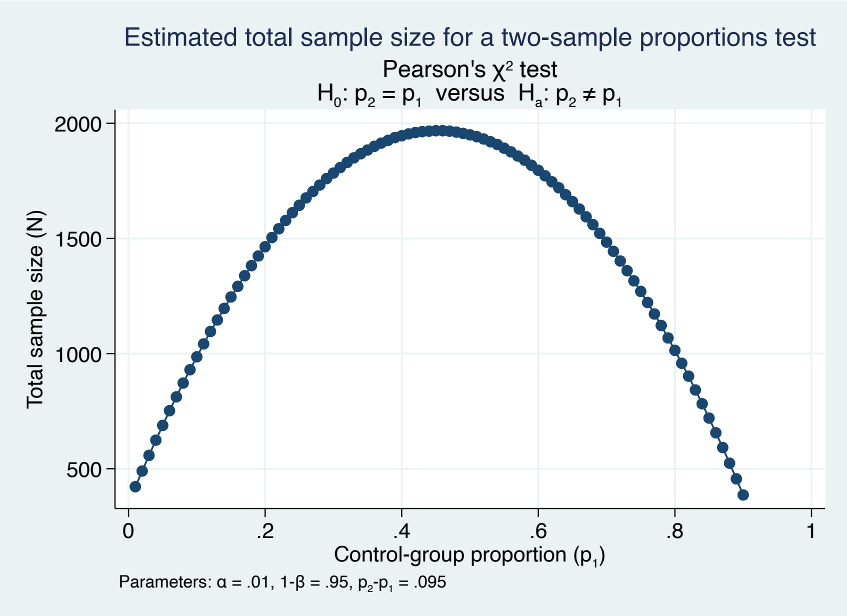

Figure 13: Required Sample Size for Detecting a 9.5% Difference in Probability to Cooperate ($P(C)$) using a χ^2 test with `twoproportions` in Stata. The chart illustrates that a minimum sample size of 2,000 is needed when the control group's $P(C)$ is 0.5. Assumptions include a statistical power of 0.95 (1-β) and a significance level of 0.01 (α). Note that the x-axis labeled "Control-group proportion" corresponds to our $P(C)$ for the informed group, i.e., the proportion of participants choosing C when informed.