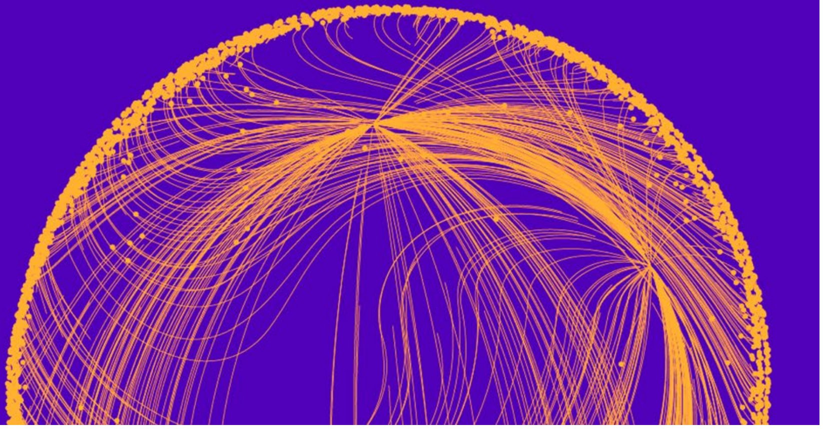
Helsinki GSE

HELSINKI GSE DISCUSSION PAPERS 52 · 2025

Hot Hands or Cold Bots? Belief Reactions to Sequential Predictions from AI, Humans, and Random Sources

Theo Herold
Marco Lambrecht
Erik Wengström



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

HANKEN

Aalto University

Helsinki GSE Discussion Papers

Helsinki GSE Discussion Papers 52 · 2025

Theo Herold, Marco Lambrecht, and Erik Wengström

Hot Hands or Cold Bots? Belief Reactions to Sequential Predictions from AI, Humans, and Random Sources

ISBN 978-952-7543-51-1 (PDF)
ISSN 2954-1492

Helsinki GSE Discussion Papers:
https://www.helsinkigse.fi/discussion-papers

Helsinki Graduate School of Economics
PO BOX 21210
FI-00076 AALTO
FINLAND

Helsinki, December 2025

Hot Hands or Cold Bots? Belief Reactions to Sequential Predictions from AI, Humans, and Random Sources*

Theo Herold[†], Marco Lambrecht[‡], Erik Wengström[§]

December 16, 2025

Abstract

People often draw inferences from sequences of past performance, sometimes perceiving patterns even in random outcomes. This has fueled debates regarding phenomena such as the hot hand and gambler's fallacies. With the growing use of artificial intelligence (AI) systems for forecasting and decision support, it becomes important to understand how people form beliefs from sequences of outcomes attributed to such systems. We report results from a preregistered online experiment (N = 900) in which identical outcome sequences were attributed to an AI model, a human forecaster, or a random device. Belief updating in response to higher prior success rates was strongest for human forecasters, weakest for random devices, and intermediate for AI. Reactions to streaks were similar for AI and human sources, in contrast to the strong reversal expectations observed for random sequences. Performance feedback did not alter the relative reliance on AI versus human sources. Overall, AI is perceived as quasi-human—imbued with some intentionality, yet not fully agentic. (169 words)

Keywords: Artificial intelligence; beliefs; hot hand; decision analysis; experimental economics

JEL Codes: C91, D81, D83, D91

[*] We would like to thank seminar audiences at Helsinki GSE, CNEE Workshop, ESA Helsinki, University of Stavanger, Samfunnsøkonomenes forskermøte Trondheim, LEE Workshop Benicasim, ESA Beijing, the Finnish Economic Association and NCBEE Stavanger for comments. We are also thankful to Mathieu Chevrier, Gergely Hajdu, and Hans-Theo Normann for very useful discussions. Financial support by Hanken School of Economics is gratefully acknowledged.

[†] Hanken School of Economics & Helsinki GSE

[‡] University of Stavanger

[§] Lund University

1 Introduction

People often draw inferences from the past performance of others. When faced with a sequence of previous successes and failures, individuals sometimes perceive patterns, even when outcomes are random. These patterns shape expectations about the future and influence behavior in domains ranging from sports to finance. A large body of work discusses such systematic belief reactions, most notably the *hot hand* and *gambler's fallacy* biases (Tversky & Kahneman, 1971; Gilovich, Vallone, & Tversky, 1985). The former refers to the belief that success breeds success, while the latter denotes the expectation of reversals after streaks.

In this paper, we ask whether people react differently to sequences of outcomes when these sequences are generated by *artificial intelligence* (AI) rather than by humans or inanimate random devices. As AI tools are increasingly used in prediction and decision-making tasks, people's interpretations of their successes and failures are likely to shape behavior across many domains. Understanding how individuals form and adjust beliefs in response to AI performance is therefore important both for economic decision making and for the design of human–AI interaction.

Our study builds on two literatures: the study of belief formation in sequential environments, and research on attitudes toward algorithmic prediction. Work in the first area has provided evidence on systematic misconceptions of randomness and autocorrelation, as well as offering theoretical rationales for such perceptions (Tversky & Kahneman, 1971; Gilovich et al., 1985; Barberis, Shleifer, & Vishny, 1998; Rabin, 2002; Rabin & Vayanos, 2010; Miller & Sanjurjo, 2018). This literature further documents that belief distortions depend on the specifics of the sequences and how the underlying process is perceived. In particular, when outcomes are thought to reflect human ability or intentional action, individuals tend to expect continuation (hot hand); when outcomes are seen as random or mechanical, they expect reversals (gambler's fallacy). The evidence indicates that such perceptions of agency shape belief updating across domains (Ayton & Fischer, 2004; Burns & Corpus, 2004; Caruso, Waytz, & Epley, 2010; Offerman & Sonnemans, 2004). This suggests that AI, which is technically inanimate yet often perceived as intentional or agentic, may occupy an intermediate position between human and mechanical sources.

The second strand of research we relate to focuses on how people evaluate and rely on algorithmic predictions. A central early insight from this literature is the existence of *algorithm aversion*, the tendency to avoid algorithms after observing their errors, even when they outperform human forecasters (Dietvorst, Simmons, & Massey, 2015). Under some conditions, this pattern reverses, giving rise to *algorithm appreciation* when algorithms are perceived as competent or objective (Logg, Minson, & Moore, 2019). Subsequent work explores when and why people accept or reject algorithmic advice (Dietvorst, Simmons, & Massey, 2018; Jung & Seiter, 2021; Holzmeister, Holmén, Kirchler, Stefan, & Wengström, 2022; Dargnies, Hakimov, & Kübler, 2024; Fu & Hanaki, 2024), and how such tendencies influence behavior (Chevrier, Corgnet, Guerci, & Rosaz, 2024; Leib, Köbis, Rilke, Hagens, & Irlenbusch, 2024; Klingbeil, Grützner, & Schreck, 2024).

We combine these literatures by examining how people form beliefs when faced with sequential predictions generated by an AI. To this end, we conduct a preregistered online experiment with 900 participants recruited via Prolific. Participants evaluate sequences of outcomes derived either from stock-market predictions made by large language models (ChatGPT or Microsoft Copilot) or human forecasters, or from randomly generated sequences produced via simulated dice rolls. In each condition, participants observe 24 *identical* sequences of eight binary outcomes. Our approach allows us to isolate *belief reactions*—how expectations differ across sequences—within subject, and compare these reactions cleanly across sources of outcomes.

The experiment proceeds in two stages. In the first stage, participants decide whether to "count on" the next outcome to be successful after observing sequences of eight consecutive previous outcomes. We systematically vary whether those sequences contain more prior successes, more failures, or long streaks of successes, and whether these streaks occurr early or late in the sequence. This enables us to study whether individuals react to sequences differently depending on the source of outcomes. In the second stage, participants receive randomized feedback about their prior performance and then chose whether to continue with AI or human predictions. By design, some are told they performed well, others poorly, allowing us to causally study how positive and negative feedback affects source choice.[1]

Our preregistered hypotheses test three questions. We ask whether people react differently to (i) sequences with more successes and (ii) sequences containing streaks depending on whether outcomes are attributed to AI, a human, or random source. We also investigate whether (iii) positive versus negative feedback affects reliance on AI versus human sources differently. Our findings document both differences and similarities across sources. First, participants weight past successes most heavily when predictions are attributed to humans, least heavily when they are attributed to random devices, and intermediately when they are attributed to AI. Second, responses to streaks occur more strongly in the Random condition than in the Human and AI conditions, with no detectable difference between the latter two. Third, reactions to positive versus negative feedback do not differ significantly between AI- and Human-generated predictions. Overall, the evidence suggests that people interpret AI as more intentional than inanimate objects but not fully agentic—that is, quasi-human: in some respects alike, yet in others occupying an intermediate position between human and mechanical sources.

The rest of the paper is structured as follows. In Section 2, we review the related literature. Section 3 describes the experimental framework and Section 4 presents our analysis and results. Finally, Section 5 summarizes the findings and discusses their implications.

2 Related literature

As briefly mentioned in the introduction, our paper connects two strands of research: (i) the literature on belief formation in sequential environments, particularly the *hot hand* and

[1] Note that the feedback signal remains truthful—only the assignment to feedback conditions is random. For details, see Section 3.

gambler's fallacy biases, and (ii) studies on trust and confidence in algorithmic or artificial intelligence (AI) decision-making. In this section, we provide a more detailed discussion of each of these literatures and outline how our paper relates to them.

2.1 Beliefs and Sequential Patterns

Early work in psychology and behavioral economics has documented that individuals systematically misperceive randomness. Tversky and Kahneman (1971) demonstrated that people tend to view small samples as highly representative of the underlying population—a phenomenon later formalized as the "law of small numbers." Extending these insights, (Gilovich et al., 1985) introduced the notion of the *hot hand fallacy*, showing that individuals perceive illusory streaks of success even in purely random sequences. This observation has since inspired a vast literature on subjective beliefs about autocorrelation in sequential outcomes.

Building on these foundations, a series of theoretical models has sought to formalize the mechanisms behind such belief distortions. Responding to regularities in financial markets, (Barberis et al., 1998) propose a model of regime-shifting beliefs in which investors overreact to sequences of good or bad news, inferring spurious changes in the underlying data-generating process. (Rabin, 2002) develop a model in which agents overinfer from small samples, thereby generating expectations of reversals (the gambler's fallacy), while (Rabin & Vayanos, 2010) extend this framework to dynamic environments, predicting both underreaction and overreaction to streaks depending on their length and perceived persistence. (Miller & Sanjurjo, 2018) provide complementary evidence that finite sequences can naturally give rise to patterns consistent with the law of small numbers, helping to rationalize why such beliefs may persist in experience.

Empirical studies confirm that misperceptions of randomness and persistence shape behavior in practice. For example, (Greenwood & Shleifer, 2014) show that investors' expectations extrapolate from recent returns, an applied manifestation of hot hand reasoning. Similarly, (Pelster, 2020) find that both hot hand and gambler's fallacy patterns manifest in investor decisions. Beyond financial markets, (Croson & Sundali, 2005) document evidence of both fallacies among casino patrons, while (Suetens, Galbo-Jørgensen, & Tyran, 2016) and (Polin & Benisaac, 2023) analyze lottery data and find that law-of-small-numbers beliefs and hot hand reasoning coexist in natural environments.

Experimental work has further illuminated the psychological mechanisms underlying these sequential belief distortions. (Bloomfield & Hales, 2002) show that individuals predict continuations and reversals in random walks as if market regimes periodically shift. (Asparouhova, Hertzel, & Lemmon, 2009) find that individuals tend to expect reversals after short streaks but continuations after long ones, consistent with alternating perceptions of randomness and persistence. (Powdthavee & Riyanto, 2015) demonstrate that individuals are willing to pay for predictions even when outcomes are transparently random, provided that a short streak of successful forecasts is observed. (Offerman & Sonnemans, 2004) and (Huber, Kirchler, & Sutter, 2010) show that both hot hand and gambler's fallacy reasoning can emerge across different experimental contexts, depending on how the sequence-generating process is framed or experienced. Collectively, these studies suggest that perceptions of the underlying process critically shape how individuals update beliefs from sequential information.

Several studies have sought to explain what factors determine whether individuals exhibit hot hand or gambler's fallacy reasoning.[2] One line of research emphasizes characteristics of the process itself: (Burns & Corpus, 2004) find that expectations of continuation or reversal depend on whether the sequence-generating mechanism is perceived as random or non-random. Other work highlights the role of intentionality and perceived agency. (Caruso et al., 2010) show that the more intentional and goal-directed a process is perceived to be, the greater the tendency to expect positive recency. Finally, (Ayton & Fischer, 2004) argue that these belief patterns hinge on whether outcomes are viewed as resulting from human skill or from inanimate, chance-based processes.

Our work extends this literature by introducing AI as a new class of signal generator—

[2] See (Oskarsson, Van Boven, McClelland, & Hastie, 2009) for a comprehensive overview of this literature.

neither purely human nor entirely inanimate.[3] By comparing belief reactions to identical sequences across AI, human, and random (dice-roll) treatments, we test whether individuals process sequential information differently when it originates from an artificial agent. This approach leverages the conceptual ambiguity of AI—as simultaneously mechanical and intentional—to examine how source perceptions shape sequential belief formation.

2.2 Algorithm Aversion and Appreciation

A second strand of research examines how people evaluate and rely on algorithmic predictions. (Dietvorst et al., 2015) introduced the concept of *algorithm aversion*, showing that people prefer human forecasters to algorithms even when the latter are objectively more accurate—particularly after observing algorithmic errors. Subsequent work has identified conditions under which this tendency reverses, documenting instances of *algorithm appreciation* when algorithmic competence is made salient or tasks are perceived as objective (Logg et al., 2019). More recent studies explore when and why people excessively accept or reject algorithmic advice (Dietvorst et al., 2018; Jung & Seiter, 2021; Holzmeister et al., 2022; Dargnies et al., 2024; Fu & Hanaki, 2024), as well as the behavioral implications of such patterns of reliance (Chevrier et al., 2024; Leib et al., 2024; Klingbeil et al., 2024).

Our study contributes to this literature by isolating the role of beliefs over sequential outcomes. Whereas prior work infers attitudes toward algorithms from delegation or reliance decisions, we directly measure how individuals form expectations about future success based on identical performance sequences attributed to human or algorithmic sources. By holding outcomes constant, we separate perception from performance and examine relative belief updating within subjects, rather than absolute levels of trust. This design allows us to assess how beliefs respond to variation in sequential evidence, making our findings less tied to a particular task or context. In the second stage of our experiment, participants choose which source to observe next after experiencing positive or

[3] We emphasize perception rather than ontology: AI systems are technically inanimate but often treated as quasi-agentic by users.

negative performance histories. This stage captures how prior experiences with human or AI forecasters influence subsequent information-seeking behavior—a feature that, to our knowledge, is absent from previous studies.

3 Experimental setup

This section outlines our experimental framework. We begin by introducing the main decision tasks and the sequences incorporated into our design, followed by our preregistered research hypotheses.[4] We then describe how the sequences used in the main experiment were generated during a preparatory phase. Finally, we detail the stages of the main experiment and its implementation.

3.1 Main decision task

In our main decision task, we examine individuals' beliefs about the next binary outcome after observing a short history of past outcomes. Importantly, we do not focus on overall confidence in the next outcome. Instead, we examine how beliefs respond to specific sequence patterns. Participants review sequences of eight binary outcomes—presented as either "Success" or "Failure" in the experiment, but denoted as 1 for success and 0 for failure in this section for conciseness—and decide whether to "count on" the next outcome being successful. Crucially, all participants observe the same set of sequences, albeit in random order. But, across treatments, the sequences are either attributed to an AI model, a human forecaster, or a random device (dice). We thus implement a between-subjects design with three treatment conditions: *AI*, *Human*, and *Dice*. The second stage of our experiment introduces an additional treatment: *Positive feedback* or *Negative feedback* regarding performance in the first stage.

We construct our hypotheses by grouping sequences into blocks, see Table 1 . Each sequence is paired with counterparts from other blocks that differ along a single dimension

[4] We preregistered our hypotheses in the AEA RCT Registry; see https://www.socialscienceregistry.org/trials/11057.

(success rate, denoted by inversion pair, or timing, denoted by reversion pair). All other features are held constant. This design allows us to isolate specific belief reactions by comparing responses across blocks. Our hypotheses are the following:

Hypothesis 1: Overall success rate of past outcomes

To measure how beliefs react to changes in the overall success rate we compare sequences with a majority of successes to their inverses, where every success is swapped with a failure and vice versa. For example, `(1 1 1 1 0 1 0 1)` (6 successes) is compared with its inversion `(0 0 0 0 1 0 1 0)` (2 successes). These sequences share the same alternation pattern and timing of outcomes but differ in overall success rate. Particularly, we compare belief responses across the high-success sequences (blocks 4–6) and the low-success sequences (blocks 1–3) shown in Table 1 based on the following hypothesis:

H1 (Success) The effect of overall success rate (comparing a sequence with its inversion) on belief reactions is the same for AI, human, and random sources.

Hypothesis 2: More successful outcomes towards the end

To test for effect of observing more successful outcomes towards the end of a sequence, we compare sequences that are exact time reversals of each other, such that every outcome appears in the opposite order. For example, `(0 0 0 0 0 1 0 1)` is paired with its reversed version `(1 0 1 0 0 0 0 0)`. These pairs allow us to isolate the effect of outcome timing while keeping the total number of successes constant. We want to know whether participants react to temporal clustering of outcomes, so we focus on sequences that involve a short streak of at least four identical outcomes.[5] In the experiment, this corresponds to comparing sequences in blocks 1 and 5 to their reversion pairs in blocks 2 and 4 (see Table 1). We compare reactions across AI, human, and random sources and state the following hypothesis:

[5] See Offerman and Sonnemans (2004) for discussion of hot hand, gambler's fallacy, and related cold hand beliefs.

H2 (Recent success): The effect of observing more successful outcomes towards the end of the sequence on belief reactions is the same across sources.

Hypothesis 2A: Hot streaks at the end

We also provide a more focused test if participants exhibit hot hand beliefs (expecting recent successes to continue) by removing sequences that involve streaks of four or more unsuccessful outcomes as in Hypothesis 2. This leaves us with sequences that contain streaks of at least four consecutive successes concentrated toward the end of the sequence and sequences where successes appear earlier. For example, `(1 0 1 0 1 1 1 1)` (success streak at the end) is compared with `(1 1 1 1 0 1 0 1)` (success streak earlier). Again, these pairs differ in the temporal clustering of successes while maintaining their overall success rate. In the experiment, this corresponds to comparing sequences in block 5 to their reversion pairs in block 4 in Table 1. To compare across sources we state the following hypothesis:

H2A (Recent success streak): The effect of observing a hot streak (a run of successes) at the end of the sequence on belief reactions is the same across sources.

Hypothesis 3: The effect of feedback

In a second stage of our design, we assess how evaluative feedback affects source preference. After completing the sequence judgments discussed above, participants receive feedback telling them they performed well or poorly when deciding on whether to ”count on” the outcomes or not.[6] Participants then choose whether they prefer to receive a sequence from AI or a human for a final repetition of the task. This tests whether positive versus negative

[6] We achieve variation in feedback by randomizing participants across treatments in which the modal response is correct or incorrect. Specifically, for each sequence, we also collected the subsequent (ninth) outcome to assess whether participants accurately decide to count on success. We find most of our sequences, on different occasions, preceding both a successful and an unsuccessful ninth outcome. We exploit this variation for randomized treatment in the second stage of the main experiment.

Table 1: Sequences

Sequence id (j)	Block	Sequence	Reversion pair id (r)	Inversion pair id (e)	High alternation sequence
1	1	**00000101**	1	1	0
2	1	**00000110**	2	2	0
3	1	**00001001**	3	3	0
4	1	**00001010**	4	4	0
5	2	**10100000**	1	5	0
6	2	**01100000**	2	6	0
7	2	**10010000**	3	7	0
8	2	**01010000**	4	8	0
9	3	**00101010**	5	9	1
10	3	**01001010**	6	10	1
11	3	**01010010**	6	11	1
12	3	**01010100**	5	12	1
13	4	**11110101**	7	4	0
14	4	**11110110**	8	3	0
15	4	**11111001**	9	2	0
16	4	**11111010**	10	1	0
17	5	**10101111**	7	8	0
18	5	**01101111**	8	7	0
19	5	**10011111**	9	6	0
20	5	**01011111**	10	5	0
21	6	**11010101**	11	9	1
22	6	**10110101**	12	10	1
23	6	**10101101**	12	11	1
24	6	**10101011**	11	12	1

feedback has different effects on participants' inclination to use AI versus human sources. We formulate the following hypothesis:

H3 (Feedback): The impact of receiving positive versus negative performance feedback on participants' choice of predictor (AI vs. human) is the same across the AI and human conditions.

3.2 Preparatory phase

We obtained the sequences for our main experiment separately for each source. The sequences generated by AI models are based on series of stock price predictions. We acquired historical S&P 500 data and randomly selected a stock and point in time.[7] We informed a large language model (LLM), either ChatGPT or Microsoft Copilot, about the stock price movement over five consecutive trading days and asked for a prediction whether the stock price will rise or fall on the sixth day.[8] We repeated this task several times, generating sequences of successful and unsuccessful predictions.

We conducted the same task with participants on Prolific to generate sequences for the human treatment. We showed the performance of a randomly selected stock over five consecutive trading days and asked them to predict whether the stock will rise or fall on the sixth day, relative to the previous trading day.[9] Participants repeated the task multiple times, each time for a randomly chosen stock at a random point in time.

For the dice treatment, we randomly chose an ex-ante threshold (in the range of one to six) and simulate sequences of dice rolls in Stata. Outcomes that exceed the threshold are labeled successful, and failure otherwise.

Overall, we obtained a comprehensive set of sequences of successful and unsuccessful outcomes during this phase. From this set, we selected the sequences previously shown in

[7] The stock market data covers all trading days for all listed companies on the S&P500 between 2005 and 2020.

[8] Note that we did not reveal the stock and the point in time that were randomly selected. For the exact prompt given to the AI models, see Appendix B.

[9] The task was incentivized. For each correct prediction, participants earned 0.04 GBP (up to a maximum of 1.20 GBP).

Table 1 for our main experiment. Furthermore, we find most of the sequences multiple times in our data. Many of them, on different occasions, precede both a successful and an unsuccessful ninth outcome. We exploit this variation for randomized treatment in the second stage of the main experiment.

3.3 Main experiment

Our main experiment consists of several stages, with Stages 1 and 2 forming the core tasks designed to elicit our primary outcome variables. Table 2 outlines the sequential structure.[10]

Table 2: Stages in the experiment

Stage	Description
0	Welcome screen
1	Beliefs over sequential outcomes
2	Source selection
3	Control variables
4	Payment summary

We implement six treatments using a 3 × 2 between-subjects factorial design, varying treatments along two dimensions. The first dimension is the source of sequential outcomes: AI, Human, or Dice. The second dimension is feedback direction: in the PosFeedback condition, the ninth outcomes are such that participants following modal response patterns tend to be correct (generating positive performance feedback); in the NegFeedback condition, modal responses tend to be incorrect (generating negative feedback).[11]

3.3.1 Stage 1: Beliefs over sequential outcomes

At the beginning of the first stage, participants are randomly assigned to one of the six treatments. They receive (treatment-dependent) instructions regarding their task in this

[10] Appendix C provides further details and screenshots of each stage.

[11] Technically, we elicit modal responses from our pilot ($n = 122$). The same responses turned out to be modal in the main experiment (regardless of the source of sequential outcomes). Depending on the sequence, the modal responses occurred with frequencies ranging from approximately 63% to 87%.

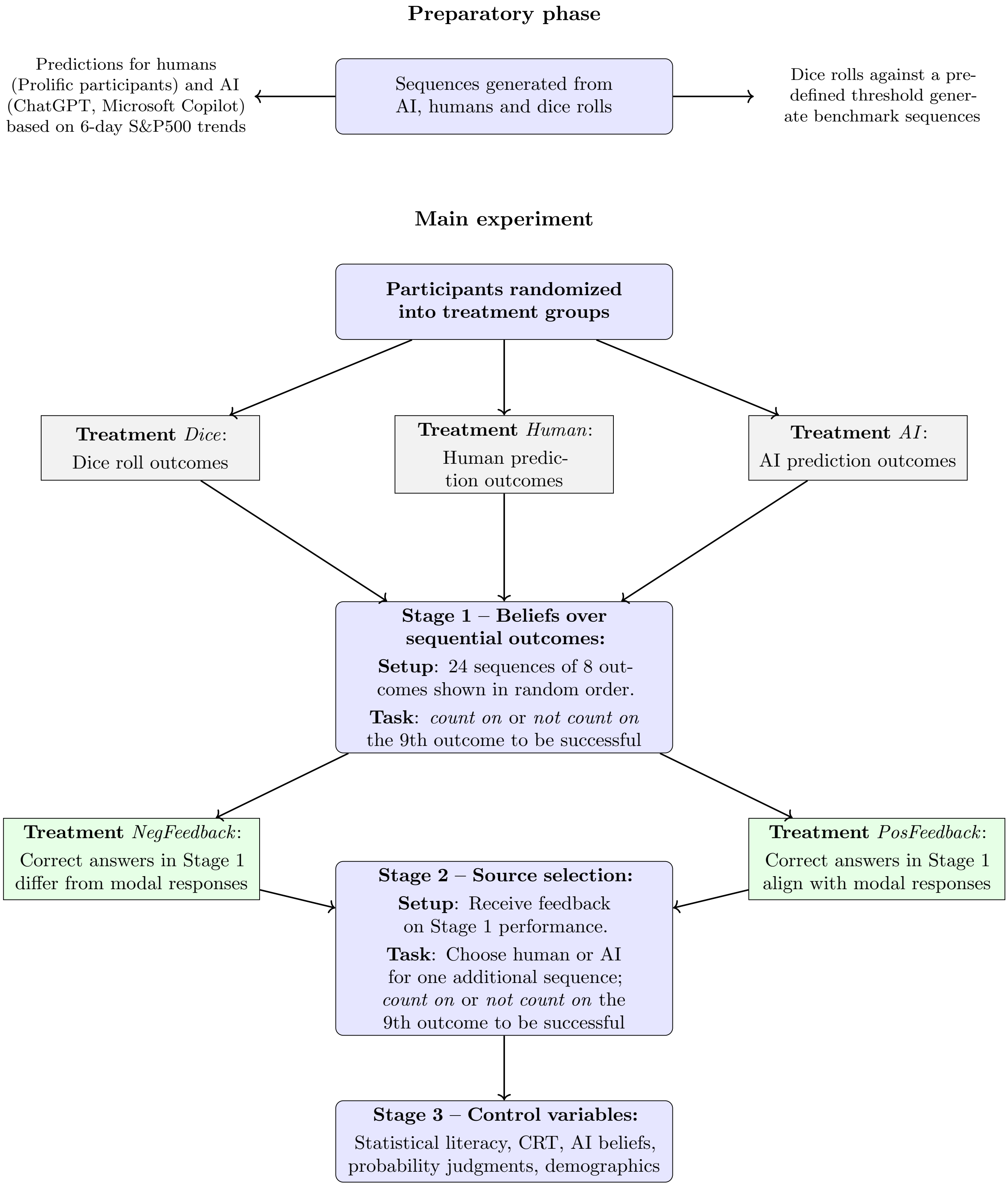

Preparatory phase
Predictions for humans (Prolific participants) and AI (ChatGPT, Microsoft Copilot) based on 6-day S&P500 trends
Sequences generated from AI, humans and dice rolls
Dice rolls against a pre-defined threshold generate benchmark sequences
Main experiment
Participants randomized into treatment groups
Treatment Dice: Dice roll outcomes
Treatment Human: Human prediction outcomes
Treatment AI: AI prediction outcomes
Stage 1 – Beliefs over sequential outcomes:
Setup: 24 sequences of 8 outcomes shown in random order.
Task: count on or not count on the 9th outcome to be successful
Treatment NegFeedback: Correct answers in Stage 1 differ from modal responses
Treatment PosFeedback: Correct answers in Stage 1 align with modal responses
Stage 2 – Source selection:
Setup: Receive feedback on Stage 1 performance.
Task: Choose human or AI for one additional sequence; count on or not count on the 9th outcome to be successful
Stage 3 – Control variables:
Statistical literacy, CRT, AI beliefs, probability judgments, demographics

Figure 1: Overview of Experimental Design

stage, knowing that they will have to pass a comprehension quiz in order to continue with the experiment. In treatments AI and Human, they learn that they will receive information on sequential outcomes that refer to predictions regarding the price of (randomly chosen) financial assets. Importantly, they are informed that the predictions in each sequence were made consecutively by the same AI model (in treatment AI) or the same participant in a previous study (in treatment Human). Conversely, in the dice treatment, we explain that each sequence they will see relates to a series of dice rolls against an unknown but fixed threshold.

In all treatments, they learn that their task is to decide whether they count on the subsequent outcome to be successful or not. They also learn that their task is repeated 24 times, and that the sequences they see are not related to one another. In particular, we inform them that in the dice treatment the random threshold may vary from sequence to sequence. Similarly, in treatments AI and Human, each sequence may show outcomes generated from different AI models or human participants, respectively (and again concern different sets of stocks at different times). We also provide an example of their decision and inform them about the payment of this task. After completing the comprehension quiz, participants continue to the decision screens of this stage.[12]

We show each participant 24 sequences, one at a time, and in random order. They automatically continue to the next screen showing the next sequence once they made their decision whether to *count on* or to *not count on* the subsequent outcome to be successful. Among the 24 decisions, participants encounter each sequence along with its inversion and reversion. This allows us to identify belief reactions within-subject, and compare reactions (rather than decisions) across treatments.

[12] If they fail to answer the comprehension questions, we send them back to the instruction screen. We include the comprehension questions and screenshots of the full instructions of the AI and dice treatment in Appendix C.

3.3.2 Stage 2: Source selection

After providing beliefs over the continuation of 24 sequences participants proceed to the second stage where they receive feedback on their performance. We count each participants number of correct decisions and inform them accordingly. However, dependent on randomly assigned treatment (PosFeedback or NegFeedback), the correct decisions for the sequences in Stage 1 either align with modal responses or not.[13] This design creates variation in correct responses across treatments, which we exploit in our analysis.

Specifically, we examine how participants react to variation in feedback when choosing between a human- or AI-generated sequence for one final round of the Stage 1 task.[14] Participants in the AI and Human treatments may repeat the task with their Stage 1 source, or switch to the other respective source. Participants in the dice treatment can not repeat the task with a sequence based on dice rolls, but their decision between a human- or AI-generated sequence may serve as a benchmark.[15]

After choosing the source of the sequence, participants then move on to their final round of the task. They receive a single sequence (which differs from the 24 sequences in Stage 1) and decide whether they count on the subsequent outcome to be successful or not.

3.3.3 Stage 3: Control variables

Following the main decision tasks, participants complete a series of survey items. In particular, we provide them with incentivized questionnaires assessing statistical literacy and cognitive reflection (CRT).[16] We also ask them to report their AI expertise and beliefs,

[13] Note that all accuracy assessments are based on sequences and ninth outcomes collected during the preparatory phase.

[14] Up to this point, participants are not informed about the existence of other sources.

[15] When facing this decision, participants in the dice treatment also receive detailed information about the process that is underlying the sequences (i.e., consecutive predictions regarding prices of randomly selected stocks—see Appendix C).

[16] We elicit statistical literacy using *Module 7: Probability Numeracy* of the 2020 Health and Retirement Study (HRS 2020) developed by Peter Hudomiet (https://hrs.isr.umich.edu/documentation/modules). The test consists of four items. To test cognitive reflection, we ask three of the four questions originally presented by Thomson and Oppenheimer (2016). For each correct response in these tests, participants are rewarded with a bonus of 0.05 GBP.

probability judgments, and demographic characteristics.[17] After providing their answers, participants receive a payment summary which informs them about the rewards they have earned during the previous stages, and the experiment concludes.

3.4 Implementation details

We conducted the experiment on Prolific (https://www.prolific.com/). Coding (including randomization of treatment assignment) was implemented using oTree (Chen, Schonger, & Wickens, 2016). Preregistration of the experiment and methods was done on The American Economic Association's Registry for Randomized Controlled Trials (AEA RCT Registry).[18] We aimed at 150 participants in each of the six treatments, yielding a total sample of 900 participants. We collected data across two sessions in April and May 2025.[19]

Each participant received a fixed participation fee of 1.50 GBP and could earn a maximum up to 3.00 GBP based on performance.[20] The median participant spent about 14 minutes on the experiment, and the average payoff was 2.29 GBP. These values are very similar across treatments except for payoff, which is significantly lower in the NegFeedback treatment compared to the PosFeedback treatment.

4 Empirical analysis

This section presents the results of our empirical analysis. We first describe our preregistered approach to analyzing the experimental data. Next, we provide descriptive statistics and corresponding statistical tests for participants' decisions during the experiment, followed by

[17] Participants across treatments were similar in their background characteristics, suggesting successful randomization. Summary statistics are provided in Appendix D.

[18] For additional details on the design and implementation of the main experiment and the preparatory phase, see Appendix A and Appendix C. For details on the preregistration, see Appendix E and `https://www.socialscienceregistry.org/trials/11057`.

[19] Specifically, we collected data from 400 participants on April 25th, and the remaining 500 observations on May 1st.

[20] One first stage decision was randomly selected for payment, awarding an additional 0.75 GBP if the participant's response was correct. A bonus of 0.40 GBP was awarded for the decision task in the second stage if answered correctly. Finally, participants could earn up to 0.35 GBP by answering seven quiz questions in Stage 3 correctly.

the results for each hypothesis. Finally, we report additional exploratory, non-preregistered analysis.

4.1 Empirical approach

In line with our preregistration, we exclude the fastest and slowest five percent of respondents, leaving a sample size of 810 participants for the analysis. Table 3 depicts the number of participants across all 6 treatment axes.

Table 3: Number of Participants by treatment.

	NegFeedback	**PosFeedback**	**Total**
Dice	141	134	275
AI	128	146	274
Human	136	125	261
Total	405	405	810

At the center of our analysis are participants' beliefs about the continuation of sequences. For each sequence, we record a binary variable which indicates whether the participant believes the ninth outcome in the sequence to be successful or not. To test our hypotheses, we construct within-subject measures that capture how each participant's beliefs react to the sequence manipulations described in Section 3.1. For H1, we measure each participant's average reaction to success, i.e. the difference in responses between sequences with a majority of successes (blocks 4–6 in Table 1) and their low-success counterparts (blocks 1–3). For H2, we measure the average reaction to recent success, i.e. the difference between sequences with successes concentrated late (blocks 1 and 5) versus early (blocks 2 and 4). For H2A, we focus on the subset involving recent success streaks, that is, sequences with runs of consecutive successes in the second half (block 5) versus the first half (block 4). We compare these measures across source conditions using Mann-Whitney U tests to test our hypotheses. We complement these tests with mixed logistic regression models that account for the data's structure without aggregation.

Specifically, for H1, we use data from all blocks and estimate

$$\begin{aligned} decision_j^i = &\alpha \; + \; \beta_1 \, Dice + \beta_2 \, Human + \beta_3 \, success_j \\ &+ \; \beta_4 \, (Dice \times success_j) + \beta_5 \, (Human \times success_j) + FE_e + u_i + \epsilon_j^i \end{aligned} \tag{1}$$

where the binary variable $decision_j^i$ indicates whether subject i believes the ninth outcome in sequence j to be successful or not. *Dice* and *Human* correspond to the respective Stage 1 treatments, with AI serving as the reference category. The binary variable $success_j$ is equal to 1 if sequence j contains more successes than failures (blocks 4-6), and 0 otherwise (blocks 1-3). FE_e represents fixed effects for inversion sequence pairs e described in Table 1, capturing sequence-specific heterogeneity. The random intercept $u_i \sim \mathcal{N}(0, \sigma_u^2)$ accounts for unobserved heterogeneity at the individual level, while the residual error $\epsilon_j^i \sim \text{Logistic}(0, 1)$ reflects the logistic link function in the binary outcome model. We also estimate an extension of the above baseline specification that includes all individual–level control variables elicited in Stage 3.

For H2, we estimate a similar mixed-effects logit model as in (1), but focusing on the effects of successes late in the sequence. Using data from blocks 1, 2, 4 and 5, we estimate the model

$$\begin{aligned} decision_j^i = &\alpha \; + \; \beta_1 \, Dice + \beta_2 \, Human + \beta_3 \, recent\_success_j \\ &+ \; \beta_4 \, (Dice \times recent\_success_j) + \beta_5 \, (Human \times recent\_success_j) \\ &+ FE_r + u_i + \epsilon_j^i \end{aligned} \tag{2}$$

where $recent\_success_j$ is equal to 1 if sequence j contains more successes in the later half compared to the first half (blocks 1 and 5), and 0 otherwise (blocks 2 and 4). For H2A, we estimate an (otherwise identical) variation that replaces $recent\_success_j$ with $recent\_success\_streak_j$, which is equal to 1 if sequence j contains a streak of successes in the second half (block 5), and 0 if the streak of successes occurs in the first half (block 4). In these specifications, we consider fixed effects for reversion sequence pairs FE_r. We complement these specifications with versions that include our individual–level control variables elicited in Stage 3.

In Stage 2, we observe whether participants choose to receive a sequence generated by an AI model or by a human participant from a previous study. For subjects in four of our treatments (*(AI, PosFeedback)*, *(AI, NegFeedback)*, *(Human, PosFeedback)*, and *(Human, NegFeedback)*), we can define a variable that indicates whether they choose the same source that they have encountered during Stage 1 or not. We perform a permutation test on this variable, $choose\_same\_source^i$, across the four treatments. This test is designated to detect whether effects of receiving varying feedback on performance (*PosFeedback* versus *NegFeedback*) differ across treatment dimensions *Human* and *AI*.

Furthermore, we estimate the logit model

$$\begin{aligned} choose\_same\_source^i = \alpha + \beta_1\, human^i +\ \beta_2\, PosFeedback^i + \\ \beta_3\, (human^i \times PosFeedback^i)\ +\ stage1\_correct^i +\ \epsilon^i \quad (3) \end{aligned}$$

where $human^i$ is equal to 1 if subject i are in the human treatment, and 0 if they are in the AI treatment. $PosFeedback^i$ is equal to 1 if subject i received positive feedback, and 0 if they received negative feedback. We control for correct responses during the first stage by including $stage1\_correct^i$. $\epsilon^i \sim \text{Logistic}(0,1)$ is a residual error term. We also estimate an extension of the above specification that includes individual–level control variables elicited in Stage 3.

4.2 Descriptives

Table 4 presents basic summary statistics by treatment. Across treatments, participants expressed a similar number of positive beliefs about the ninth outcome in each sequence, with medians at 12 out of 24. Statistical tests reveal no significant differences (Kruskal-

Wallis and Mann-Whitney U tests, $p > 0.149$).

Table 4: Summary statistics across all treatment axes of main variables of interest.

	NegFeedback		**PosFeedback**	
	Mean	Median	Mean	Median
Dice				
Times participant counted on prediction	12.71	12.00	12.70	12.00
Number of correct decisions	10.56	10.00	14.06	15.50
Participant chose AI-algorithm in Stage 2	0.84	1.00	0.79	1.00
AI				
Times participant counted on prediction	12.30	12.00	12.53	12.00
Number of correct decisions	8.00	7.00	16.45	18.00
Participant chose AI-algorithm in Stage 2	0.47	0.00	0.65	1.00
Human				
Times participant counted on prediction	12.65	12.00	11.80	12.00
Number of correct decisions	7.94	7.00	16.22	18.00
Participant chose AI-algorithm in Stage 2	0.88	1.00	0.75	1.00

Note: See Appendix D for summary statistics of control variables across treatments.

As one might expect, participants in the *PosFeedback* treatments consistently gave more correct responses than participants in the *NegFeedback* treatments (Mann-Whitney U tests, $p < 0.001$). The difference between *PosFeedback* and *NegFeedback* in the AI treatments is larger compared to the Dice treatment (permutation test, two-sided $p < 0.001$), but similar as in the Human treatment (permutation test, two-sided $p > 0.725$).

Stage 2 choices revealed substantial variation: participants in the Human treatment were much more likely to select the AI sequence for Stage 2 compared to participants in the AI treatment (Mann-Whitney U test, $p < 0.001$). Having neither seen sequences generated by an AI-algorithm or Humans in the first stage, participants in the Dice treatment favored the AI sequence for stage 2, similar to the human treatment (Mann-Whitney U test, $p > 0.872$) and significantly more so than in the AI treatment (Mann-Whitney U test, $p < 0.001$).

4.3 Hypotheses testing

We now address the hypotheses laid out in Section 3. We adhere closely to our preregistration and report the results for each test it specifies.

4.3.1 H1 – Belief reactions to previous success

For this hypothesis, we focus on the effect of sequence inversion across treatments. Figure 2 shows the average reaction of our participants, generally indicating a positive reaction, i.e., that sequences with more successes induce more optimistic beliefs in subsequent success. There is no statistically significant difference between the AI and Human treatments, with average inversion reactions of 0.560 and 0.608, respectively (Mann-Whitney U test, $p > 0.186$). This effect is considerably smaller in the Dice treatment, at 0.296, and statistically different from the AI and Human treatments (Kruskal-Wallis and Mann-Whitney U tests, $p < 0.001$).

The results of our mixed logit model are presented in Table 5, with the first column estimating the baseline model in (1) and the second column including individual-level controls. We primarily report the point estimates from the first column in the text, unless stated otherwise. Our main focus is how $success_j$ (indicating sequences with a majority of successes) interacts with the treatments.

We find the coefficient on Dice to be positive and significant (0.718; $p < 0.001$), indicating that—when sequences contain a majority of failures ($success_j = 0$)—participants in the Dice treatment were substantially more likely than those in the AI treatment to expect the ninth outcome to be successful. In contrast, the coefficient on Human is negative and significant (–0.190; $p < 0.01$), meaning that participants in the Human treatment were less likely than those in the AI treatment to count on the ninth outcome under a majority of failures ($success_j = 0$). Sequences with a majority of successes ($success_j = 1$) substantially increase the likelihood that participants in the AI baseline treatment expect the ninth outcome to be successful (2.659; $p < 0.001$).

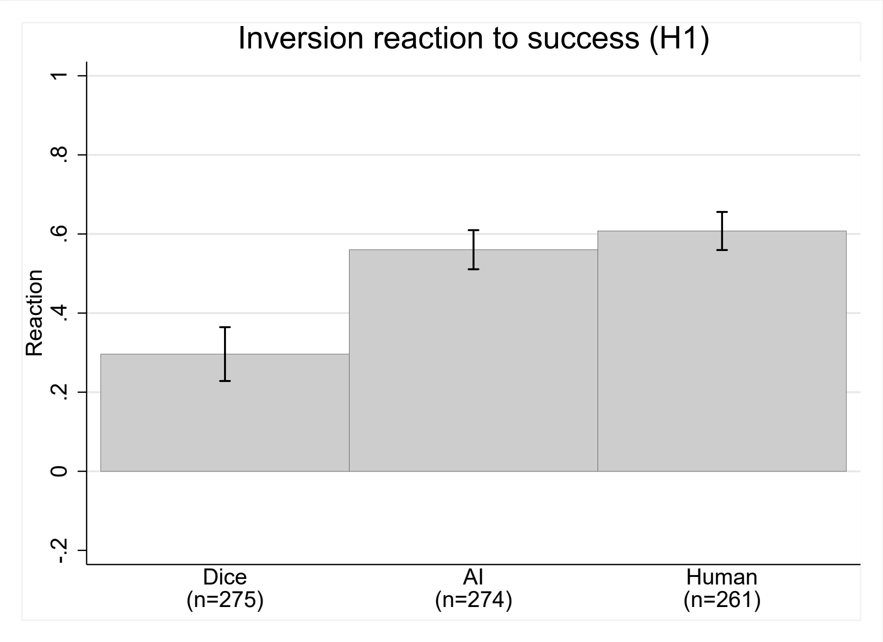


Figure 2: Propensity to count on the ninth outcome between inversion pairs, across treatment. Specifically, the y-axis measures individuals' average difference to believe in a successful outcome when observing a sequence with a majority of successes versus a sequence with a majority of failures.

Table 5: Mixed-effects logit regression in model (1) assessing belief reactions over majority successful outcomes to majority unsuccessful outcomes (H1).

	(1)	(2)
Dep. Var: $decision_j$	No controls	With controls
Dice	0.718^{***} (0.0684)	0.718^{***} (0.0673)
Human	-0.190^{**} (0.0735)	-0.189^{**} (0.0722)
$success_j$	2.659^{***} (0.0623)	2.659^{***} (0.0623)
Dice × $success_j$	-1.380^{***} (0.0813)	-1.381^{***} (0.0813)
Human × $success_j$	0.297^{***} (0.0899)	0.298^{***} (0.0899)
Statistical literacy		-0.147^{***} (0.0315)
Cognitive reflection test (CRT)		0.0322 (0.0275)
Demographics	No	Yes
Survey controls	No	Yes
Constant	-1.243^{***} (0.0757)	-1.427^{**} (0.471)
Observations	19 440	19 440

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. Controls include performance on 4 statistical literacy questions, 3 cognitive reflection questions, 3 belief-based fallacy questions, self–reported demographic variables age^i, $gender^i$, $profession^i$, $income^i$, $education^i$ and self-reported expertise and belief about artificial intelligence. See Appendix E for more details on variables.

However, the effect of success is moderated by treatment. The interaction term $Dice \times success_j$ is negative and statistically significant (–1.380; $p < 0.001$), implying that the belief-updating response to majority-success sequences is considerably weaker in the Dice treatment relative to the AI treatment. Conversely, the interaction $Human \times success_j$ is positive and statistically significant (0.297; $p < 0.001$), indicating stronger belief reactions in the Human treatment relative to AI. The inclusion of controls in column (2) does not meaningfully alter the treatment or interaction coefficients.

Thus, we reject hypothesis H_1. Participants in the AI treatment reacted more strongly to inversion—i.e., were more likely to count on the ninth outcome following sequences with a majority of successes—compared to those in the Dice treatment, but less so than in the Human treatment.

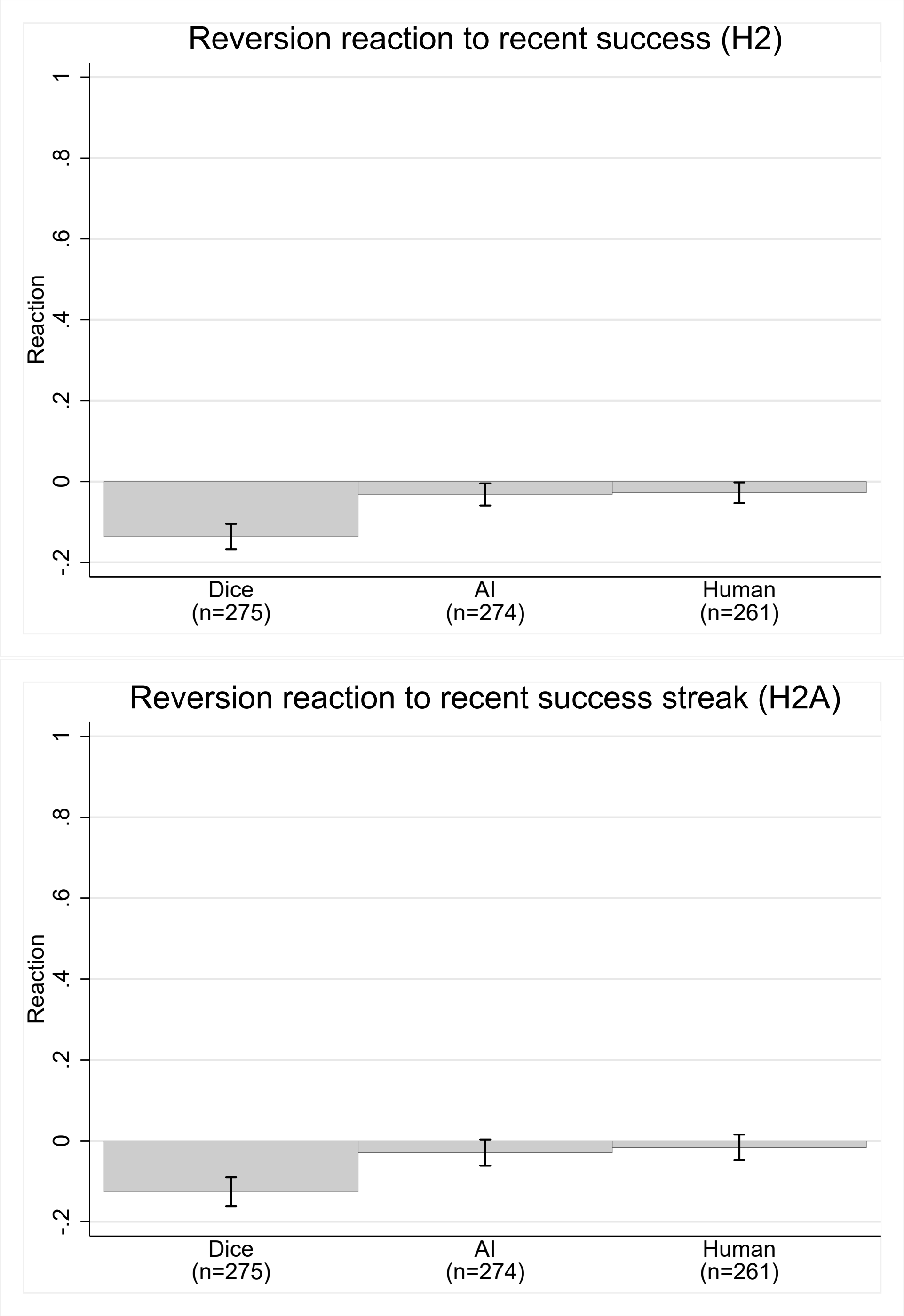
4.3.2 H2 – Belief reactions to streaks

We next examine within-subject reversion reactions. The results are presented in the top panel of Figure 3.

Reversion reactions in the Dice treatment show a negative response of –0.136, whereas reactions in the AI and Human treatments are close to zero. Statistical tests confirm that these reactions differ significantly across treatments (Kruskal-Wallis test, $p < 0.001$), driven by the Dice treatment differing from both AI and Human (Mann-Whitney U tests, $p < 0.001$). In contrast, reactions in the AI and Human treatment do not differ significantly from one another (Mann-Whitney U test, p = 0.767).

Table 6 presents the regression results of specification (2). The variable $recent\_success_j$ captures whether more successes occurred early or late in the sequence. Relative to the AI treatment, participants in the Dice treatment anticipate success more strongly (0.308, $p < 0.001$) when the latter half of the sequence contains more failures ($recent\_success_j = 0$). Generally, late successes reduce expectations of success, a pattern consistent with anticipations of outcome reversion.[21]

[21] A negative reaction is consistent with beliefs that recent success increases the likelihood of a subsequent



Reversion reaction to recent success (H2)

Reversion reaction to recent success streak (H2A)

Figure 3: Propensity to count on the ninth outcome between reversion pairs, across treatment. Specifically, the y-axis measures individuals' average difference to believe in a successful outcome when observing a sequence with late successes versus a sequence with early successes.

Table 6: Mixed-effects logit regression in model (2) assessing belief reactions over streaks in the second half to streaks in the first half of sequences (H2).

	(1)	(2)
Dep. Var: $decision_j$	No controls	With controls
Dice	0.308^{***} (0.0825)	0.304^{***} (0.0816)
Human	-0.0786 (0.0834)	-0.0773 (0.0822)
$recent\_success_j$	-0.197^{**} (0.0750)	-0.197^{**} (0.0750)
Dice × $recent\_success_j$	-0.644^{***} (0.106)	-0.644^{***} (0.106)
Human × $recent\_success_j$	0.0256 (0.107)	0.0257 (0.107)
Statistical literacy		-0.148^{***} (0.0348)
Cognitive reflection test (CRT)		0.0245 (0.0304)
Demographics	No	Yes
Survey controls	No	Yes
Constant	-1.134^{***} (0.0807)	-1.238^{*} (0.548)
Observations	12 960	12 960

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. Controls include performance on 4 statistical literacy questions, 3 cognitive reflection questions, 3 belief-based fallacy questions, self–reported demographic variables age^i, $gender^i$, $profession^i$, $income^i$, $education^i$ and self-reported expertise and belief about artificial intelligence.

In the AI treatment, exposure to sequences with more successes in the second half significantly reduces the likelihood of counting on the ninth outcome (-0.197, $p < 0.01$). In the Dice treatment, the interaction term is negative and significant (-0.644, $p < 0.001$), indicating that participants reacted more strongly to reversion cues. By contrast, reactions in the Human treatment do not differ significantly from those in AI.

Overall, we partly reject hypothesis H_2. As hypothesized, participants in the AI and Human treatment display similar reactions to sequence reversions that concentrated successes towards the end of the sequences. However, contrary to Hypothesis H_2, participants in the Dice treatment demonstrated a significantly stronger reaction compared to the two other treatments.

4.3.2A H2A – Belief reactions to hot streaks

The bottom panel of Figure 3 shows how participants react to reversion of sequences with hot streaks. The picture is essentially the same as for hypothesis H_2.[22] Table 7 shows the corresponding regression results when using specification (2) with $recent\_success\_streak_j$ instead of $recent\_success_j$. We find that participants in the AI treatment anticipate an increased likelihood for an unsuccessful outcome when the later half of the sequence is a streak of successful outcomes ($recent\_success\_streak_j = 1$; -0.402, $p < 0.05$). Again, the Dice treatment stands out (interaction term of -0.784, $p < 0.001$), while the Human condition does not differ significantly from AI.

This leads us to partly reject hypothesis H_{2A}. In the AI treatment, reactions to hot streak reversion are less pronounced compared to the Dice treatment, but are not significantly different from reactions in the Human treatment.

failure which—given random outcomes—aligns with the *gambler's fallacy*).

[22] The Dice treatment shows a reversion reaction of –0.126, which is statistically different from both the AI and Human treatments (Kruskal-Wallis and Mann-Whitney U tests, p < 0.001). Effects in the AI and Human treatments are not statistically different from one another (Mann-Whitney U test, $p = 0.790$).

Table 7: Mixed-effects logit regression in model (2) assessing belief reactions over streaks of success in the second half to a streaks of success in the first half of sequences (H2A).

	(1)	(2)
Dep. Var: $decision_j$	No controls	With controls
Dice	-1.345*** (0.281)	-1.405*** (0.284)
Human	-0.0994 (0.291)	-0.0815 (0.295)
$recent\_success\_streak_j$	-0.402* (0.160)	-0.400* (0.159)
Dice × $recent\_success\_streak_j$	-0.784*** (0.210)	-0.788*** (0.210)
Human × $recent\_success\_streak_j$	0.171 (0.230)	0.167 (0.230)
Statistical literacy		0.364** (0.140)
Cognitive reflection test (CRT)		0.312* (0.124)
Demographics	No	Yes
Survey controls	No	Yes
Constant	3.558*** (0.239)	2.675 (2.113)
Observations	6 480	6 480

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. Controls include performance on 4 statistical literacy questions, 3 cognitive reflection questions, self–reported demographic variables age^i, $gender^i$, $profession^i$, $income^i$, $education^i$ and self-reported expertise and belief about artificial intelligence.

4.3.3 H3 – Reactions to variation in belief performance

In the second stage, subjects receive feedback on their performance in the first stage before choosing between an AI or Human sequence for another repetition of the Stage 1 task. Depending on treatment (*PosFeedback* or *NegFeedback*), their performance is likely to be either relatively good or relatively bad. We investigate how this variation affects choices by analyzing whether participants in the AI and Human treatments choose the same source that they had encountered during Stage 1. Statistical testing shows no significant difference in reactions to performance variation (permutation test, two-sided $p > 0.580$).

Table 8 shows the results of our corresponding logistic regressions. In the Human treatment we observe a strong tendency to switch to the AI-generated sequence (-1.891, $p < 0.001$). However, the interaction term between treatment and feedback is not statistically significant, indicating that reactions to performance variation do not differ between AI-treated and Human-treated participants. Including control variables in Column (2) does not materially change these results.[23]

Thus, we cannot reject hypothesis H_3. Participants' responses to variation in feedback about their Stage 1 performance do not significantly differ between the AI and Human treatment conditions.

4.4 Exploratory analysis

This section provides exploratory analysis that was not included in the preregistration. We report predicted probabilities from the main models used in Section 4, as well as additional analysis concerning control variables.

4.4.1. Predicted probabilities over main models

Figure 4 illustrates marginal probabilities of the interaction terms in the regression

[23] Note that Columns (3) and (4) present two specifications that were not preregistered. These regressions show that the positive feedback treatment significantly influenced participants to choose the same source they experienced in Stage 1—but only when the highly correlated variable representing correct answers is removed. The interaction term, however, remains insignificant.

Table 8: Logit regression in model (3) to assess Stage 2 choice of Human or AI based on Stage 1 success (H3).

	(1)	(2)	(3)	(4)
Dep. Var: $choose\_same\_source_j$	No controls	With controls	No controls	With controls
Human	-1.891***	-1.948***	-1.890***	-1.942***
	(0.319)	(0.318)	(0.320)	(0.319)
PosFeedback	0.491	0.505	0.747**	0.761**
	(0.358)	(0.368)	(0.248)	(0.262)
Human × PosFeedback	0.162	0.110	0.158	0.100
	(0.419)	(0.427)	(0.419)	(0.427)
Number of correct	0.0305	0.0304		
	(0.0307)	(0.0306)		
Statistical literacy		-0.189		-0.192
		(0.149)		(0.148)
Cognitive reflection test (CRT)		0.0890		0.0923
		(0.123)		(0.123)
Demographics	No	Yes	No	Yes
Survey controls	No	Yes	No	Yes
Constant	-0.370	-0.0371	-0.125	0.139
	(0.302)	(2.021)	(0.177)	(2.060)
Observations	535	534	535	534

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors in parentheses. Controls include performance on 4 statistical literacy questions, 3 cognitive reflection questions, self–reported demographic variables age^i, $gender^i$, $profession^i$, $income^i$, $education^i$ and self-reported expertise and belief about artificial intelligence. Columns (3) and (4) present two specifications that were not preregistered.

models (1), (2) and (3).

From the first panel of Figure 4, we can infer that the probability of counting on the ninth outcome rises substantially when the majority of outcomes in a sequence are successes. This shift is largest in the Human and AI treatments, from slightly more than 20% to about 80%. By contrast, the Dice treatment shifts from approximately 40% to around 65%.

The middle panel depicts probabilities for sequences with more successful outcomes in the second half of the sequence (relative to the first half). In the Dice treatment, predicted probabilities fall from almost 60% to about 45%. In contrast, beliefs of participants in the Human and AI treatments appear relatively stable.

Finally, the lower panel refers to the predicted probability of choosing either the AI or Human treatment in the second stage after feedback. The AI and Human treatments vary in levels but follow the same trend as in the original logit regressions, i.e., the predicted probability of choosing AI is higher than for choosing Human, and slightly higher when receiving positive compared to negative feedback.

4.4.2. Predicted probabilities over control variables

We explore whether additional patterns emerge when focusing on the control variables in our analysis. These patterns should be interpreted with caution, as they were not preregistered and serve primarily to generate hypotheses for future research.

We elicited several measures of reasoning ability and susceptibility to cognitive biases and combined them into participants' rationality scores. This index is defined as the standardized sum of correct responses on the statistical literacy and CRT tasks, minus the standardized number of fallacy-indicating responses on the fallacy questionnaire:

$$\text{Rationality score}_i \quad = \quad z(\text{Statistical literacy}_i) \;+\; z(\text{CRT}_i) \;-\; z(\text{Fallacy score}_i)$$

Higher values indicate greater rationality in terms of stronger numeracy and reflective

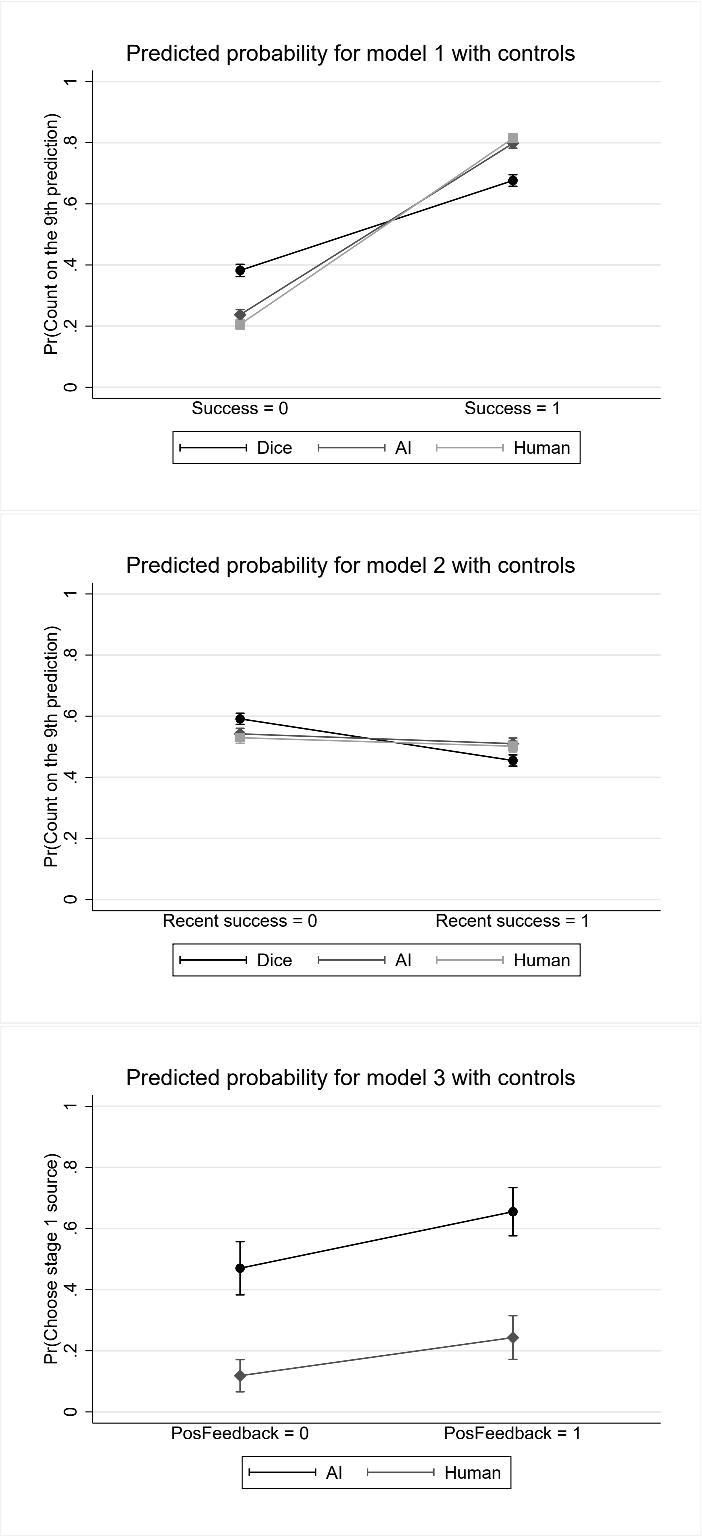


Figure 4: Predicted probability for interaction terms in models (1) (top panel), (2) (middle panel) and (3) (bottom panel).

thinking combined with fewer fallacy-prone judgments. We dichotomize this variable at the sample median, coding participants with scores above 0.2 as ”above-median rational” and those below as ”below-median rational” to define a binary variable *rationality*.

Figure 5 illustrates marginal probabilities when including interaction terms between *rationality* and *success*, *recent_success*, and *PosFeedback* in the models (1), (2) and (3).[24] From its first panel, we infer that the predicted choice probabilities rise when the sequence's majority outcomes are successes. The increase is large and similar in the AI and Human treatments across rationality types. The slope of the Dice treatment differs noticeably across rationality types, with those who score above the median showing a similar trend as participants in the AI and Human treatments.

The middle panel reveals distinct patterns in the AI and Human treatments across rationality types when the sequence contains more successes in the second half. For participants with lower rationality scores, the predicted probability decreases when success is more recent, whereas for those who score above the median, the effect is minimal. In the Dice treatment, reactions are largely similar across rationality types, with predicted probabilities decreasing in both groups when the latter half of the sequence includes more successful outcomes.

The bottom panel illustrates how the predicted probability of choosing the same source as in Stage 1 varies with feedback. The pattern is broadly similar across rationality groups. If anything stands out, it is that participants in the Human treatment appear less responsive to positive feedback—i.e., they continue switching to AI despite an encouraging prior experience. However, the confidence intervals are substantial.

[24] Appendix F shows similar plots for gender, self-reported AI-scepticism and self-reported AI-expertise. Although most patterns appear broadly similar, the tentative exploration indicates some variation that may merit further investigation, particularly concerning feedback effects across subgroups.

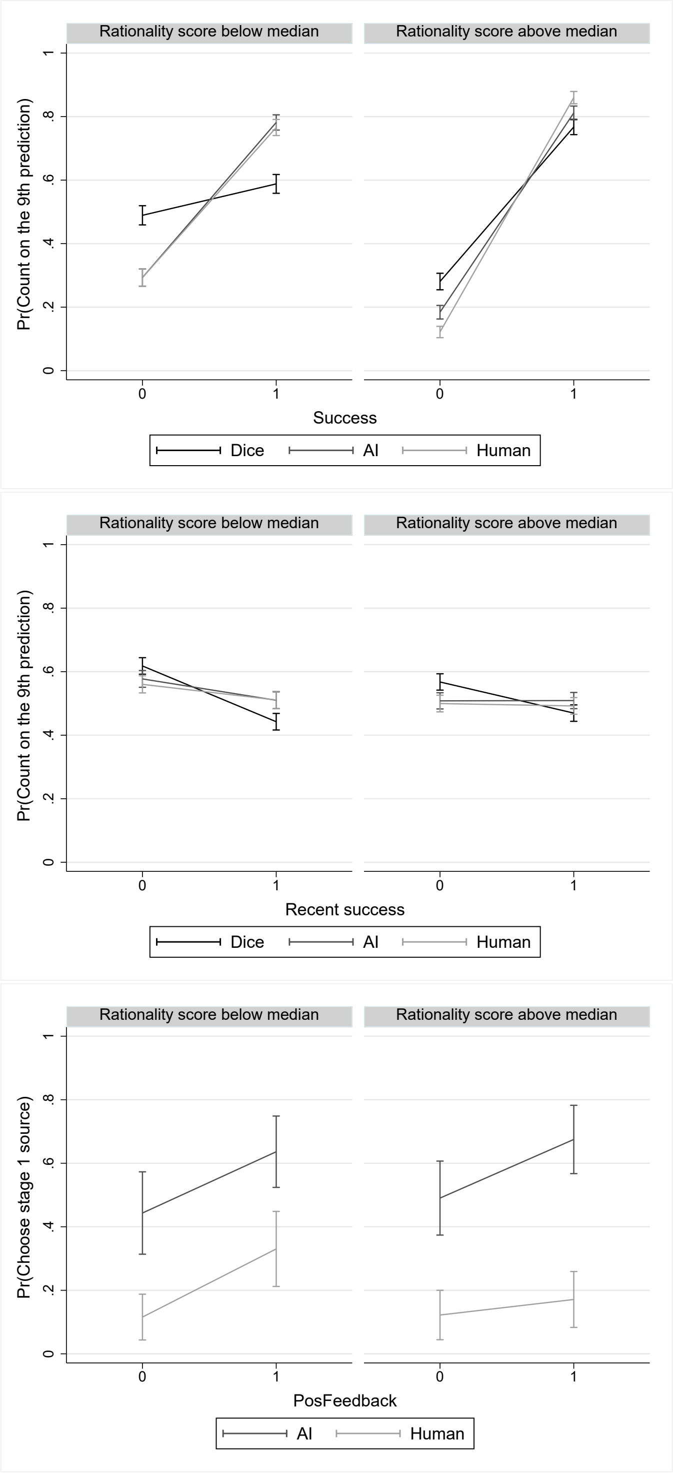


Figure 5: Predicted probabilities for interaction terms in models (1) (top panel), (2) (middle panel) and (3) (bottom panel) when including additional rationality score interaction. Individual-level control variables are included in the models.

5 Conclusion

This paper examines how people update beliefs from sequences of outcomes when those sequences are attributed to AI models, human forecasters, or a random device. Using a preregistered online experiment in which all participants evaluated identical sequences of outcomes, we documented three central findings. First, belief updating in response to higher prior success rates is strongest when sequences are attributed to humans, weakest for random (dice) sources, and lies between these two for AI. Second, reactions to streaks and recent successes differ sharply for random sources—where reversal expectations are pronounced—but are similar and notably weaker for AI and human sources. Third, randomized feedback on participants' own performance does not meaningfully alter beliefs or choices between AI and human forecasters. These results are robust across specifications and individual-level controls.

Our findings contribute to two strands of literature. They extend research on belief formation and sequential inference by showing that perceived source agency affects how individuals extrapolate from short outcome sequences. Consistent with prior work linking intentionality to continuation beliefs (Caruso et al., 2010; Ayton & Fischer, 2004), our results suggest that AI is treated as "quasi-human": more agentic than mechanical processes but not fully equivalent to human forecasters. At the same time, the results speak to the literature on algorithm aversion and appreciation (Dietvorst et al., 2015, 2018; Logg et al., 2019; Jung & Seiter, 2021; Holzmeister et al., 2022; Dargnies et al., 2024; Fu & Hanaki, 2024). In our environment, attributing sequences to AI did not result in lower expected success relative to human-attributed sequences. Instead, belief updating in response to prior successes and failures was stronger when sequences were attributed to human decision makers. By holding outcome sequences fixed across sources, our design allows us to separate source perceptions from differences in realized performance.

Our findings are informative about how individuals interpret sequences of forecasts and how this differs across sources. Because participants reacted to AI sequences in a manner broadly similar to human forecasts—but distinctly unlike their responses to random sources—individuals appear to perceive AI predictions as originating from an agentive, skill-based process. In applied settings such as financial forecasting, sports analytics, or macroeconomic outlooks, people may interpret short sequences of AI predictions as informative in ways that differ systematically from how they interpret mechanically generated data. For research on belief formation, the results show that the perceived nature of the source matters: identical sequences lead to different inferences depending on whether individuals believe the predictions come from an AI system, a human forecaster, or a random mechanism. Models of belief updating may therefore benefit from explicitly incorporating source-type perceptions when seeking to explain or predict how people evaluate sequential information.

Building on these findings, several promising directions for future research emerge. Our exploratory analyses suggest that gender, AI expertise, attitudes toward automation, and measures of rationality may shape how participants react to AI forecasts and personal experiences. These insights open up opportunities for future work that zooms in on these dimensions through dedicated designs, enabling a more detailed understanding of belief formation in human–AI interactions. Furthermore, in our design, AI forecasts were generated by a large language model in a preparatory phase; examining whether similar patterns extend to other forms of algorithms—such as transparent statistical models or specialized forecasting systems—would help illuminate how different types of AI elicit quasi-human perceptions. Future work could also investigate the welfare implications of belief formation by assessing the correctness of participants' inferences relative to the underlying data-generating process, an aspect we deliberately abstracted from in the present study.

Finally, our results may inform discussions around the design of decision-support systems. Such systems may need to recognize that users often treat AI as partially agentic. Clear communication of uncertainty, calibrated performance summaries, and interfaces that contextualize streaks may help align user beliefs with actual model performance. Regulators and platform designers might consider encouraging standardized reporting of accuracy and calibration, especially in domains where misinterpreting sequences can lead to costly errors.

In sum, AI predictions are neither treated as fully mechanical nor fully human. Instead, AI occupies an intermediate, quasi-human position in users' mental models, and this distinction meaningfully shapes belief updating from short sequences. Understanding when and why AI is perceived this way is a key task for future research as algorithmic forecasts become increasingly integrated into everyday decision-making.

Declarations

Declaration of competing interests

None.

Declaration of generative AI and AI-assistance in the writing process

The authors used ChatGPT to assist with language and formatting in Latex. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

The authors gratefully acknowledge the financial support of Hanken School of Economics.

References

Asparouhova, E., Hertzel, M., & Lemmon, M. (2009). Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers. *Management Science*, *55*(11), 1766–1782.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369–1378. doi: 10.3758/BF03206327

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*(3), 307-343. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304405X98000270 doi: https://doi.org/10.1016/S0304-405X(98)00027-0

Bloomfield, R., & Hales, J. (2002). Predicting the next step of a random walk: Experimental evidence of regime-shifting beliefs. *Journal of Financial Economics*, *65*(3), 397-414.

Burns, B. D., & Corpus, B. (2004). Randomness and inductions from streaks: “Gambler’s fallacy” versus ”hot hand“. *Psychonomic bulletin & review*, *11*(1), 179–184.

Caruso, E. M., Waytz, A., & Epley, N. (2010). The intentional mind and the hot hand: Perceiving intentions makes streaks seem likely to continue. *Cognition*, *116*(1), 149–153.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88-97. Retrieved from https://www.sciencedirect.com/science/article/pii/S2214635016000101 doi: https://doi.org/10.1016/j.jbef.2015.12.001

Chevrier, M., Corgnet, B., Guerci, E., & Rosaz, J. (2024). *Algorithm credulity: Human and algorithmic advice in prediction experiments.* (Working paper. Université Côte d’Azur, Emlyon, and Burgundy School of Business)

Croson, R., & Sundali, J. (2005). The gambler’s fallacy and the hot hand: Empirical data from casinos. *Journal of risk and uncertainty*, *30*(3), 195–209.

Dargnies, M.-P., Hakimov, R., & Kübler, D. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*. (Published online in Articles in Advance, June 19) doi: 10.1287/mnsc.2022.02774

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. doi: 10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, *64*(3), 1155–1170.

Fu, Q., & Hanaki, N. (2024). People rely on chatgpt more than peers in decision-making. *Nature Human Behaviour*. (Forthcoming) doi: 10.1038/s41562-024-01848-0

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295-314.

Greenwood, R., & Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, *27*(3), 714–746.

Holzmeister, F., Holmén, M., Kirchler, M., Stefan, M., & Wengström, E. (2022). Delegation decisions in finance. *Management Science*, *69*(8), 4828–4844. Retrieved from https://doi.org/10.1287/mnsc.2022.4555 doi: 10.1287/mnsc.2022.4555

Huber, J., Kirchler, M., & Sutter, M. (2010). The hot hand belief and the gambler’s fallacy in investment decisions under risk. *Theory and Decision*, *68*, 445–462. doi: 10.1007/s11238-009-9160-3

Jung, M., & Seiter, M. (2021). Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study. *Journal of Management Control*, *32*, 495–516. doi: 10.1007/s00187-021-00326-3

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on ai—an experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. *The Economic Journal*, *134*(658), 766–784.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Miller, J. B., & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, *86*(6), 2019–2047.

Offerman, T., & Sonnemans, J. (2004). What's causing overreaction? An experimental investigation of recency and the hot hand effect. *The Scandinavian Journal of Economics*, *106*(3), 533–553. doi: 10.1111/j.0347-0520.2004.00372.x

Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological bulletin*, *135*(2), 262.

Pelster, M. (2020). The gambler's and hot-hand fallacies: Empirical evidence from trading data. *Economics Letters*, *187*, 108887.

Polin, B. A., & Benisaac, E. (2023, January). A longitudinal analysis of the hot hand and gambler's fallacy biases. *Judgment and Decision Making*, *18*, 1-1.

Powdthavee, N., & Riyanto, Y. E. (2015). Would you pay for transparently useless advice? A test of boundaries of beliefs in the folly of predictions. *Review of Economics and Statistics*, *97*(2), 257–272. doi: 10.1162/REST_a_00453

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, *117*(3), 775–816. Retrieved 2025-09-18, from http://www.jstor.org/stable/4132489

Rabin, M., & Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, *77*(2), 730–778. doi: 10.1111/j.1467-937X.2009.00585.x

Suetens, S., Galbo-Jørgensen, C. B., & Tyran, J.-R. (2016). Predicting lotto numbers: A natural experiment on the gambler's fallacy and the hot-hand fallacy. *Journal of the European Economic Association*, *14*(3), 584–607. doi: 10.1111/jeea.12147

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. doi: 10.1017/S1930297500007622

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Pediatrics*.

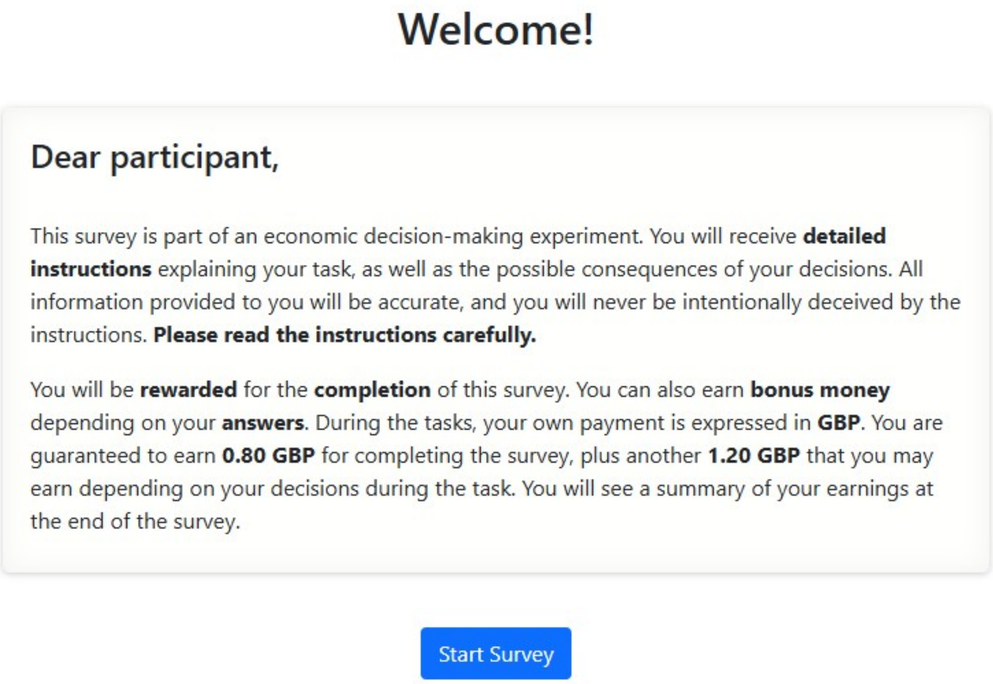
A Experimental instructions and details for preparatory experiment

Table A.1 outlines the sequential structure of the preparatory experiment. Subsection A.2 presents the instructions and interface for the prediction task. Subsection A.3 shows the concluding summary screen.

Table A.1: Stages in preparatory experiment

Step	Screen Description
1	Welcome screen
2	Prediction task
3	Summary page

A.1 Welcome screen



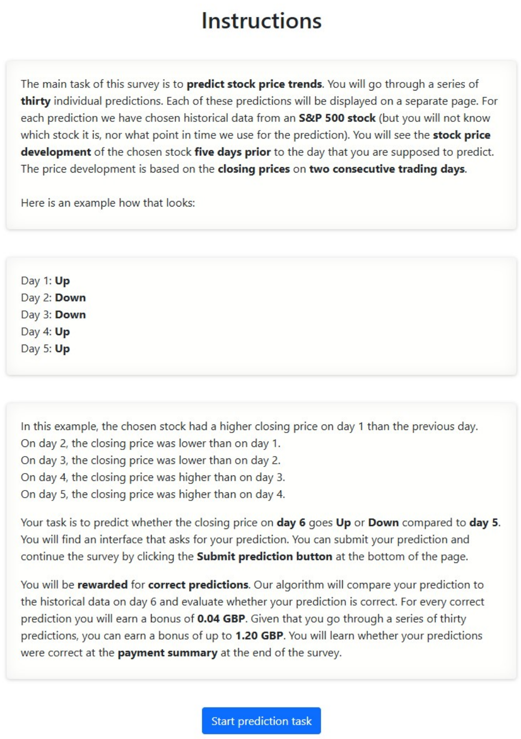
Welcome!

Dear participant,

This survey is part of an economic decision-making experiment. You will receive **detailed instructions** explaining your task, as well as the possible consequences of your decisions. All information provided to you will be accurate, and you will never be intentionally deceived by the instructions. **Please read the instructions carefully.**

You will be **rewarded** for the **completion** of this survey. You can also earn **bonus money** depending on your **answers**. During the tasks, your own payment is expressed in **GBP**. You are guaranteed to earn **0.80 GBP** for completing the survey, plus another **1.20 GBP** that you may earn depending on your decisions during the task. You will see a summary of your earnings at the end of the survey.

A.2 Instructions and main task



Instructions

The main task of this survey is to **predict stock price trends**. You will go through a series of **thirty** individual predictions. Each of these predictions will be displayed on a separate page. For each prediction we have chosen historical data from an **S&P 500 stock** (but you will not know which stock it is, nor what point in time we use for the prediction). You will see the **stock price development** of the chosen stock **five days prior** to the day that you are supposed to predict. The price development is based on the **closing prices** on **two consecutive trading days**.

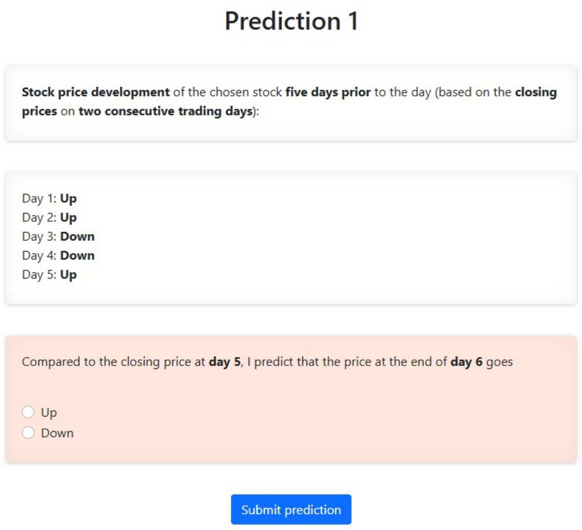
Here is an example how that looks:

Day 1: **Up**
Day 2: **Down**
Day 3: **Down**
Day 4: **Up**
Day 5: **Up**

In this example, the chosen stock had a higher closing price on day 1 than the previous day.
On day 2, the closing price was lower than on day 1.
On day 3, the closing price was lower than on day 2.
On day 4, the closing price was higher than on day 3.
On day 5, the closing price was higher than on day 4.

Your task is to predict whether the closing price on **day 6** goes **Up** or **Down** compared to **day 5**. You will find an interface that asks for your prediction. You can submit your prediction and continue the survey by clicking the **Submit prediction button** at the bottom of the page.

You will be **rewarded** for **correct predictions**. Our algorithm will compare your prediction to the historical data on day 6 and evaluate whether your prediction is correct. For every correct prediction you will earn a bonus of **0.04 GBP**. Given that you go through a series of thirty predictions, you can earn a bonus of up to **1.20 GBP**. You will learn whether your predictions were correct at the **payment summary** at the end of the survey.



Prediction 1

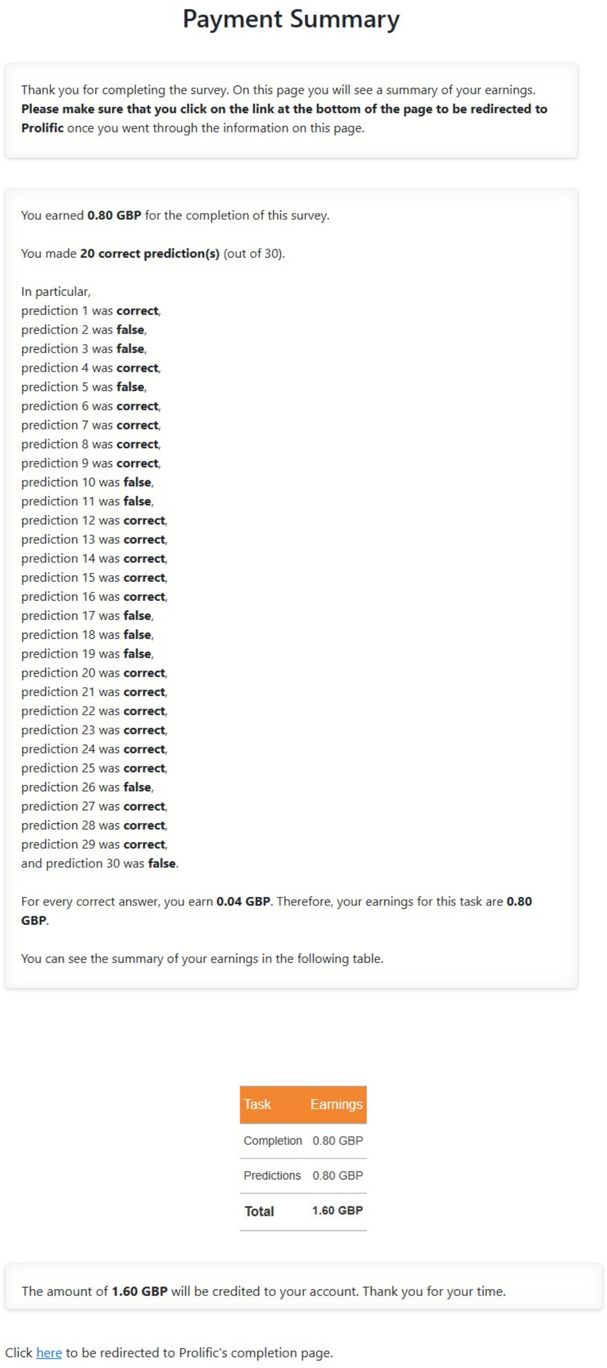
Stock price development of the chosen stock **five days prior** to the day (based on the **closing prices** on **two consecutive trading days**):

Day 1: **Up**
Day 2: **Up**
Day 3: **Down**
Day 4: **Down**
Day 5: **Up**

Compared to the closing price at **day 5**, I predict that the price at the end of **day 6** goes

○ Up
○ Down

A.3 Summary page



Payment Summary

Thank you for completing the survey. On this page you will see a summary of your earnings. **Please make sure that you click on the link at the bottom of the page to be redirected to Prolific** once you went through the information on this page.

You earned **0.80 GBP** for the completion of this survey.

You made **20 correct prediction(s)** (out of 30).

In particular,
prediction 1 was **correct**,
prediction 2 was **false**,
prediction 3 was **false**,
prediction 4 was **correct**,
prediction 5 was **false**,
prediction 6 was **correct**,
prediction 7 was **correct**,
prediction 8 was **correct**,
prediction 9 was **correct**,
prediction 10 was **false**,
prediction 11 was **false**,
prediction 12 was **correct**,
prediction 13 was **correct**,
prediction 14 was **correct**,
prediction 15 was **correct**,
prediction 16 was **correct**,
prediction 17 was **false**,
prediction 18 was **false**,
prediction 19 was **false**,
prediction 20 was **correct**,
prediction 21 was **correct**,
prediction 22 was **correct**,
prediction 23 was **correct**,
prediction 24 was **correct**,
prediction 25 was **correct**,
prediction 26 was **false**,
prediction 27 was **correct**,
prediction 28 was **correct**,
prediction 29 was **correct**,
and prediction 30 was **false**.

For every correct answer, you earn **0.04 GBP**. Therefore, your earnings for this task are **0.80 GBP**.

You can see the summary of your earnings in the following table.

Task	Earnings
Completion	0.80 GBP
Predictions	0.80 GBP
Total	**1.60 GBP**

The amount of **1.60 GBP** will be credited to your account. Thank you for your time.

Click here to be redirected to Prolific's completion page.

B Prompt given to ChatGPT and Microsoft Copilot to elicit sequences

I will give you sequences of the performance of stocks from S&P500 between the year 2005 and 2020. Each sequence represents a different stock at a different point in time, with the stock and the time period being randomly selected. Each sequence presents the outcome of the subsequent stock over five trading days, where 1 = the stock's value has increased compared to the previous trading day, and 0 = the stock's value has decreased compared to the previous trading day.

I want you to predict whether the stocks' value increase or decrease on the sixth trading day, compared to the previous trading day for all of these different stocks. ONLY include your predictions of 1:s and 0:s in your output, nothing else. Please present your predictions in a row and in .csv format.

The sequences are as follows:

C Experimental instructions and details for main experiment

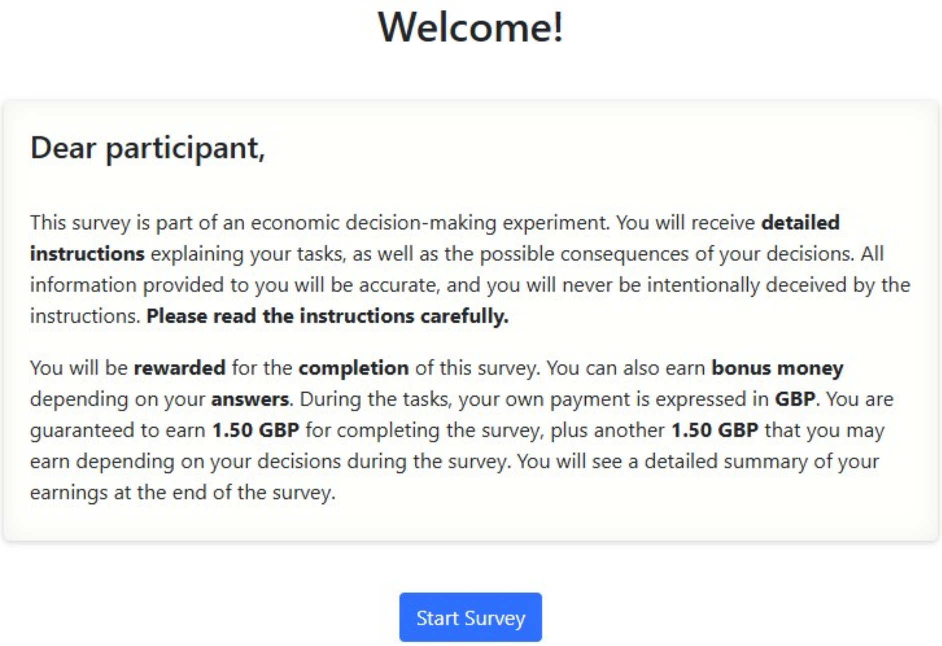
Table C.1 outlines the sequential structure of the experimental stages. Subsection C.1 displays the welcome screen shown to all participants. Subsection C.2 presents the instructions page, comprehension quiz, and interface for the stage 1 belief elicitation task. Depending on the assigned treatment, participants received tailored instructions describing the prediction sequences. As the human and AI treatments are almost structurally identical[25], we include example instructions for the AI and dice treatments. Subsection C.3 displays the stage 2 performance feedback and source-selection task. Participants first received feedback on their prior performance before deciding whether to receive a human or AI sequence. Subsection C.4 contains the debriefing and control elicitation instruments. Subsection C.5 shows the concluding summary screen.

Table C.1: Stages in main experiment

Step	Screen Description
0	Welcome screen
1	Belief reaction task
2	Performance reaction task
3	Debrief and control questions
4	Summary page

[25] For the AI treatment, the instruction text reads: ”You will receive information about prior outcomes of stock market predictions generated by an Artificial Intelligence (AI) model.” For the human treatment, the text reads: ”You will receive information about prior outcomes of stock market predictions generated by a human participant in a previous experiment.” The two treatments are otherwise the same, differing only in the substitution of the term “human” with “AI model”.

C.1 Welcome screen



Welcome!

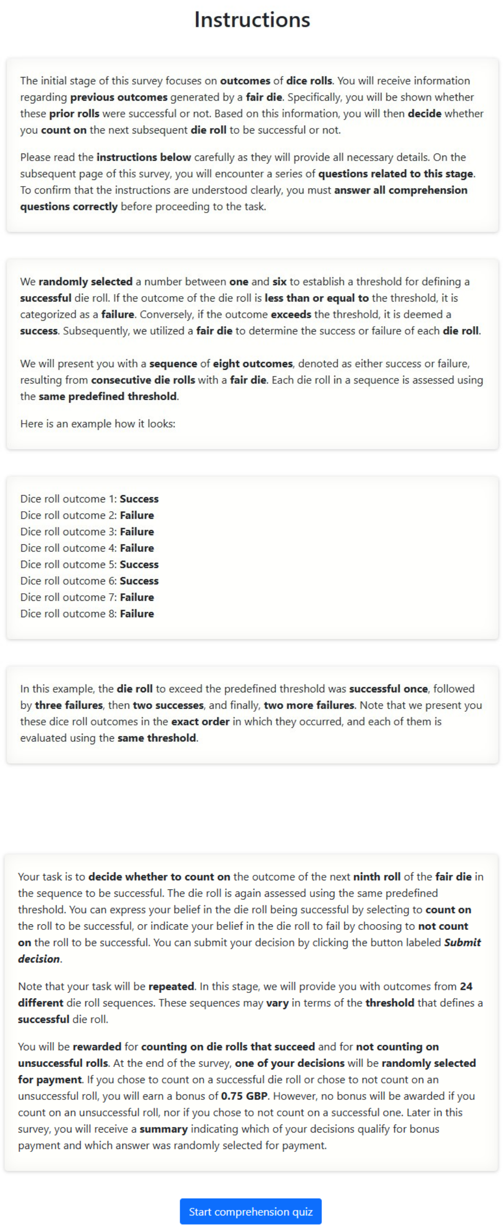
Dear participant,

This survey is part of an economic decision-making experiment. You will receive **detailed instructions** explaining your tasks, as well as the possible consequences of your decisions. All information provided to you will be accurate, and you will never be intentionally deceived by the instructions. **Please read the instructions carefully.**

You will be **rewarded** for the **completion** of this survey. You can also earn **bonus money** depending on your **answers**. During the tasks, your own payment is expressed in **GBP**. You are guaranteed to earn **1.50 GBP** for completing the survey, plus another **1.50 GBP** that you may earn depending on your decisions during the survey. You will see a detailed summary of your earnings at the end of the survey.

C.2 Stage 1 belief reaction task

Instructions, dice treatment



Instructions

The initial stage of this survey focuses on **outcomes** of **dice rolls**. You will receive information regarding **previous outcomes** generated by a **fair die**. Specifically, you will be shown whether these **prior rolls** were successful or not. Based on this information, you will then **decide** whether you **count on** the next subsequent **die roll** to be successful or not.

Please read the **instructions below** carefully as they will provide all necessary details. On the subsequent page of this survey, you will encounter a series of **questions related to this stage**. To confirm that the instructions are understood clearly, you must **answer all comprehension questions correctly** before proceeding to the task.

We **randomly selected** a number between **one** and **six** to establish a threshold for defining a **successful** die roll. If the outcome of the die roll is **less than or equal to** the threshold, it is categorized as a **failure**. Conversely, if the outcome **exceeds** the threshold, it is deemed a **success**. Subsequently, we utilized a **fair die** to determine the success or failure of each **die roll**.

We will present you with a **sequence** of **eight outcomes**, denoted as either success or failure, resulting from **consecutive die rolls** with a **fair die**. Each die roll in a sequence is assessed using the **same predefined threshold**.

Here is an example how it looks:

Dice roll outcome 1: **Success**
Dice roll outcome 2: **Failure**
Dice roll outcome 3: **Failure**
Dice roll outcome 4: **Failure**
Dice roll outcome 5: **Success**
Dice roll outcome 6: **Success**
Dice roll outcome 7: **Failure**
Dice roll outcome 8: **Failure**

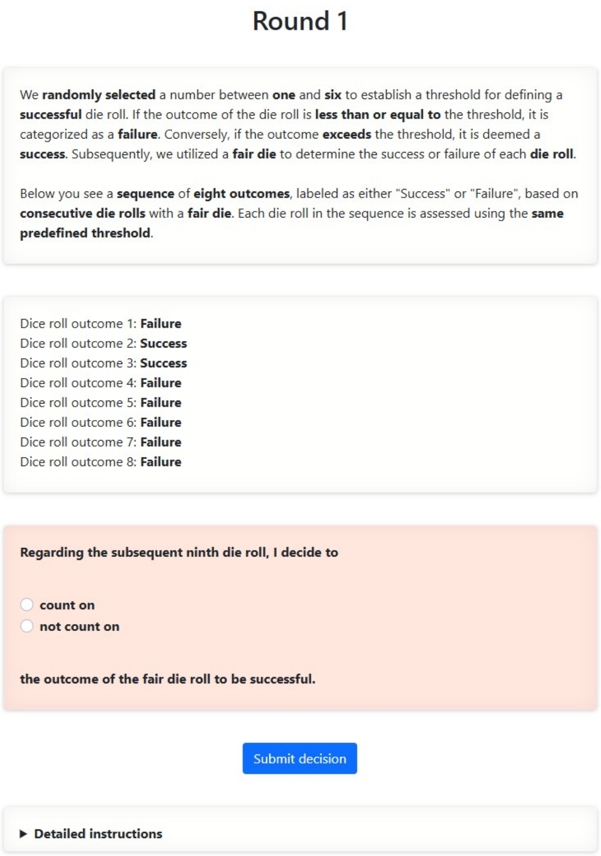
In this example, the **die roll** to exceed the predefined threshold was **successful once**, followed by **three failures**, then **two successes**, and finally, **two more failures**. Note that we present you these dice roll outcomes in the **exact order** in which they occurred, and each of them is evaluated using the **same threshold**.

Your task is to **decide whether to count on** the outcome of the next **ninth roll** of the **fair die** in the sequence to be successful. The die roll is again assessed using the same predefined threshold. You can express your belief in the die roll being successful by selecting to **count on** the roll to be successful, or indicate your belief in the die roll to fail by choosing to **not count on** the roll to be successful. You can submit your decision by clicking the button labeled ***Submit decision***.

Note that your task will be **repeated**. In this stage, we will provide you with outcomes from **24 different** die roll sequences. These sequences may **vary** in terms of the **threshold** that defines a **successful** die roll.

You will be **rewarded** for **counting on die rolls that succeed** and for **not counting on unsuccessful rolls**. At the end of the survey, **one of your decisions** will be **randomly selected for payment**. If you chose to count on a successful die roll or chose to not count on an unsuccessful roll, you will earn a bonus of **0.75 GBP**. However, no bonus will be awarded if you count on an unsuccessful roll, nor if you chose to not count on a successful one. Later in this survey, you will receive a **summary** indicating which of your decisions qualify for bonus payment and which answer was randomly selected for payment.

Example of stage 1 task, dice treatment



Round 1

We **randomly selected** a number between **one** and **six** to establish a threshold for defining a **successful** die roll. If the outcome of the die roll is **less than or equal to** the threshold, it is categorized as a **failure**. Conversely, if the outcome **exceeds** the threshold, it is deemed a **success**. Subsequently, we utilized a **fair die** to determine the success or failure of each **die roll**.

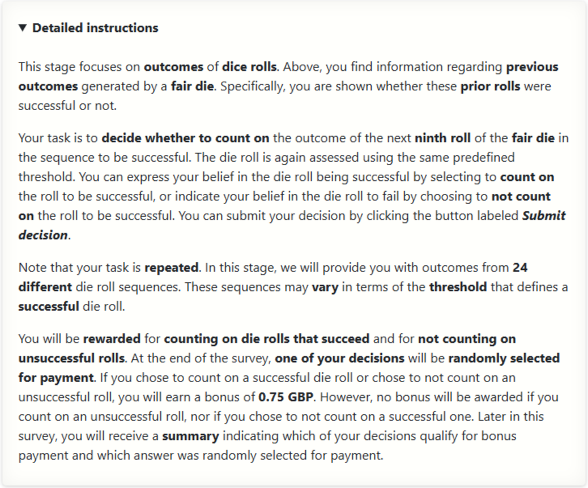
Below you see a **sequence** of **eight outcomes**, labeled as either "Success" or "Failure", based on **consecutive die rolls** with a **fair die**. Each die roll in the sequence is assessed using the **same predefined threshold**.

Dice roll outcome 1: **Failure**
Dice roll outcome 2: **Success**
Dice roll outcome 3: **Success**
Dice roll outcome 4: **Failure**
Dice roll outcome 5: **Failure**
Dice roll outcome 6: **Failure**
Dice roll outcome 7: **Failure**
Dice roll outcome 8: **Failure**

Regarding the subsequent ninth die roll, I decide to

○ **count on**
○ **not count on**

the outcome of the fair die roll to be successful.



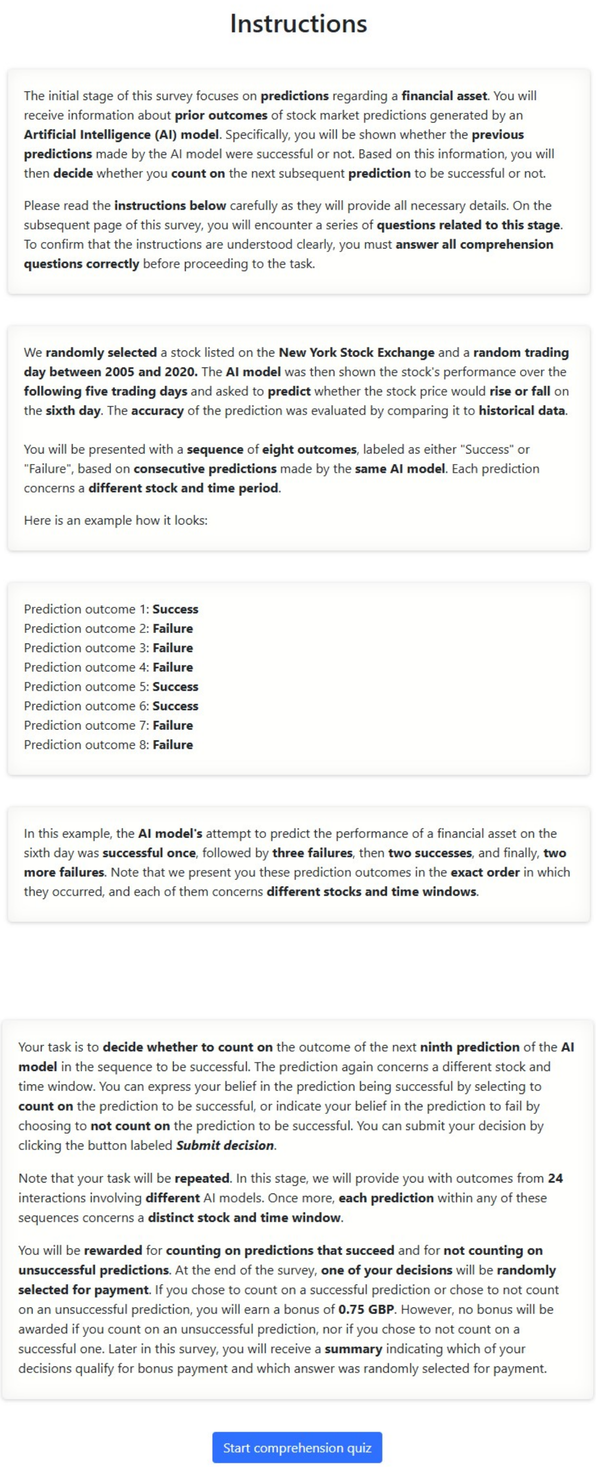
This stage focuses on **outcomes** of **dice rolls**. Above, you find information regarding **previous outcomes** generated by a **fair die**. Specifically, you are shown whether these **prior rolls** were successful or not.

Your task is to **decide whether to count on** the outcome of the next **ninth roll** of the **fair die** in the sequence to be successful. The die roll is again assessed using the same predefined threshold. You can express your belief in the die roll being successful by selecting to **count on** the roll to be successful, or indicate your belief in the die roll to fail by choosing to **not count on** the roll to be successful. You can submit your decision by clicking the button labeled ***Submit decision***.

Note that your task is **repeated**. In this stage, we will provide you with outcomes from **24 different** die roll sequences. These sequences may **vary** in terms of the **threshold** that defines a **successful** die roll.

You will be **rewarded** for **counting on die rolls that succeed** and for **not counting on unsuccessful rolls**. At the end of the survey, **one of your decisions** will be **randomly selected for payment**. If you chose to count on a successful die roll or chose to not count on an unsuccessful roll, you will earn a bonus of **0.75 GBP**. However, no bonus will be awarded if you count on an unsuccessful roll, nor if you chose to not count on a successful one. Later in this survey, you will receive a **summary** indicating which of your decisions qualify for bonus payment and which answer was randomly selected for payment.

Instructions, AI treatment



Instructions

The initial stage of this survey focuses on **predictions** regarding a **financial asset**. You will receive information about **prior outcomes** of stock market predictions generated by an **Artificial Intelligence (AI) model**. Specifically, you will be shown whether the **previous predictions** made by the AI model were successful or not. Based on this information, you will then **decide** whether you **count on** the next subsequent **prediction** to be successful or not.

Please read the **instructions below** carefully as they will provide all necessary details. On the subsequent page of this survey, you will encounter a series of **questions related to this stage**. To confirm that the instructions are understood clearly, you must **answer all comprehension questions correctly** before proceeding to the task.

We **randomly selected** a stock listed on the **New York Stock Exchange** and a **random trading day between 2005 and 2020.** The **AI model** was then shown the stock's performance over the **following five trading days** and asked to **predict** whether the stock price would **rise or fall** on the **sixth day**. The **accuracy** of the prediction was evaluated by comparing it to **historical data**.

You will be presented with a **sequence** of **eight outcomes**, labeled as either "Success" or "Failure", based on **consecutive predictions** made by the **same AI model**. Each prediction concerns a **different stock and time period**.

Here is an example how it looks:

Prediction outcome 1: **Success**
Prediction outcome 2: **Failure**
Prediction outcome 3: **Failure**
Prediction outcome 4: **Failure**
Prediction outcome 5: **Success**
Prediction outcome 6: **Success**
Prediction outcome 7: **Failure**
Prediction outcome 8: **Failure**

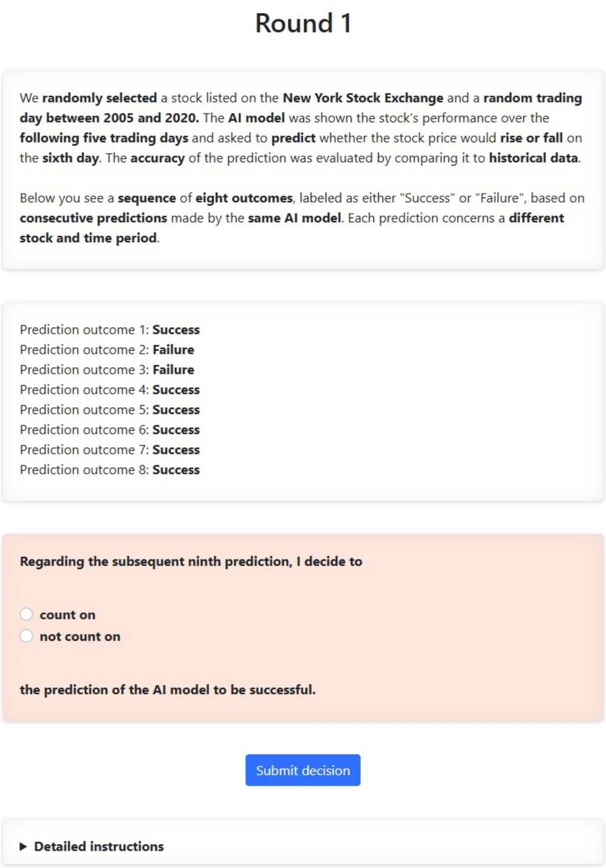
In this example, the **AI model's** attempt to predict the performance of a financial asset on the sixth day was **successful once**, followed by **three failures**, then **two successes**, and finally, **two more failures**. Note that we present you these prediction outcomes in the **exact order** in which they occurred, and each of them concerns **different stocks and time windows**.

Your task is to **decide whether to count on** the outcome of the next **ninth prediction** of the **AI model** in the sequence to be successful. The prediction again concerns a different stock and time window. You can express your belief in the prediction being successful by selecting to **count on** the prediction to be successful, or indicate your belief in the prediction to fail by choosing to **not count on** the prediction to be successful. You can submit your decision by clicking the button labeled ***Submit decision***.

Note that your task will be **repeated**. In this stage, we will provide you with outcomes from **24** interactions involving **different** AI models. Once more, **each prediction** within any of these sequences concerns a **distinct stock and time window**.

You will be **rewarded** for **counting on predictions that succeed** and for **not counting on unsuccessful predictions**. At the end of the survey, **one of your decisions** will be **randomly selected for payment**. If you chose to count on a successful prediction or chose to not count on an unsuccessful prediction, you will earn a bonus of **0.75 GBP**. However, no bonus will be awarded if you count on an unsuccessful prediction, nor if you chose to not count on a successful one. Later in this survey, you will receive a **summary** indicating which of your decisions qualify for bonus payment and which answer was randomly selected for payment.

Example of stage 1 task, AI treatment



Round 1

We **randomly selected** a stock listed on the **New York Stock Exchange** and a **random trading day between 2005 and 2020.** The **AI model** was shown the stock's performance over the **following five trading days** and asked to **predict** whether the stock price would **rise or fall** on the **sixth day**. The **accuracy** of the prediction was evaluated by comparing it to **historical data**.

Below you see a **sequence** of **eight outcomes**, labeled as either "Success" or "Failure", based on **consecutive predictions** made by the **same AI model**. Each prediction concerns a **different stock and time period**.

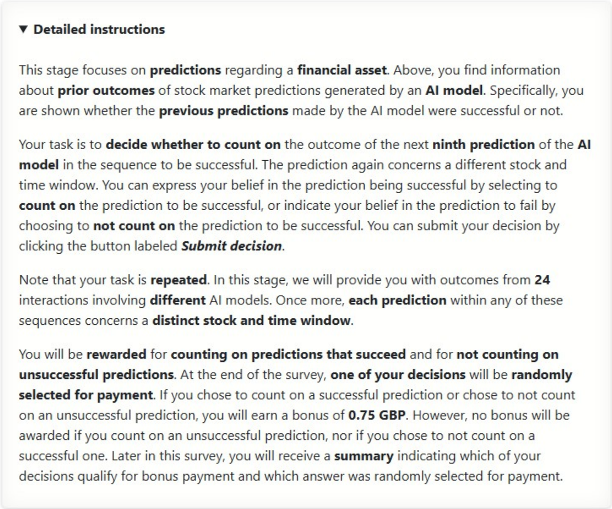
Prediction outcome 1: **Success**
Prediction outcome 2: **Failure**
Prediction outcome 3: **Failure**
Prediction outcome 4: **Success**
Prediction outcome 5: **Success**
Prediction outcome 6: **Success**
Prediction outcome 7: **Success**
Prediction outcome 8: **Success**

Regarding the subsequent ninth prediction, I decide to

○ **count on**
○ **not count on**

the prediction of the AI model to be successful.

▶ **Detailed instructions**



▼ **Detailed instructions**

This stage focuses on **predictions** regarding a **financial asset**. Above, you find information about **prior outcomes** of stock market predictions generated by an **AI model**. Specifically, you are shown whether the **previous predictions** made by the AI model were successful or not.

Your task is to **decide whether to count on** the outcome of the next **ninth prediction** of the **AI model** in the sequence to be successful. The prediction again concerns a different stock and time window. You can express your belief in the prediction being successful by selecting to **count on** the prediction to be successful, or indicate your belief in the prediction to fail by choosing to **not count on** the prediction to be successful. You can submit your decision by clicking the button labeled ***Submit decision***.

Note that your task is **repeated**. In this stage, we will provide you with outcomes from **24** interactions involving **different** AI models. Once more, **each prediction** within any of these sequences concerns a **distinct stock and time window**.

You will be **rewarded** for **counting on predictions that succeed** and for **not counting on unsuccessful predictions**. At the end of the survey, **one of your decisions** will be **randomly selected for payment**. If you chose to count on a successful prediction or chose to not count on an unsuccessful prediction, you will earn a bonus of **0.75 GBP**. However, no bonus will be awarded if you count on an unsuccessful prediction, nor if you chose to not count on a successful one. Later in this survey, you will receive a **summary** indicating which of your decisions qualify for bonus payment and which answer was randomly selected for payment.

C.2.1 Comprehension quiz

Following the instruction page and before the stage 1 belief reaction task, each participant was immediately tasked with answering a comprehension quiz consisting of three multiple-choice questions. Participants could retake the comprehension quiz as many times as required and they could only proceed to viewing the first prediction sequence after answering all three questions correctly. Incorrect answers sent participants back to the instruction page. The comprehension quiz questions are provided below, together with the available alternatives and the highlighted correct answers.

Comprehension quiz, dice roll treatment

1. *What is your task in this stage of the experiment?*
   - (1) Make die roll predictions
   - (2) Remember sequences of die rolls
   - (3) Decide whether to count on the outcomes of die rolls — **correct answer**
2. *What will each sequence you are going to see be based on?*
   - (1) Consecutive outcomes of die rolls, each assessed using different thresholds
   - (2) Consecutive outcomes of die rolls, each assessed using the same threshold — **correct answer**
   - (3) Randomly ordered outcomes of die rolls, each assessed using different thresholds
3. *How many times will the task be repeated in this stage?*
   - (1) 6 times
   - (2) 9 times
   - (3) 24 times — **correct answer**

Comprehension quiz, AI treatment

1. *What is your task in this stage of the experiment?*
   - (1) Make stock market predictions
   - (2) Remember sequences of stock market predictions
   - (3) Decide whether to count on stock market predictions — **correct answer**
2. *What will each sequence you are going to see be based on?*

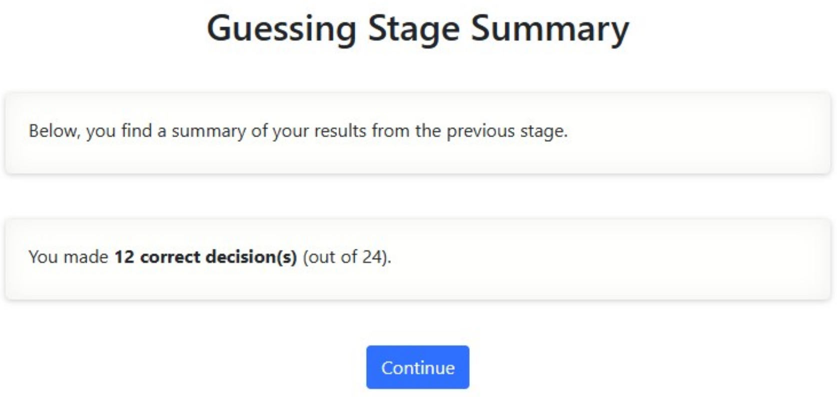
- (1) Consecutive stock predictions by {a participant/an AI model}, same stock and consecutive days
- (2) Consecutive stock predictions by {a participant/an AI model}, different stocks and randomly drawn starting days — **correct answer**
- (3) Randomly ordered stock predictions by {a participant/an AI model}, different stocks and randomly drawn starting days

3. *How many times will the task be repeated in this stage?*

   - (1) 6 times
   - (2) 9 times
   - (3) 24 times — **correct answer**

C.3 Stage 2 performance reaction task

Stage 1 summary

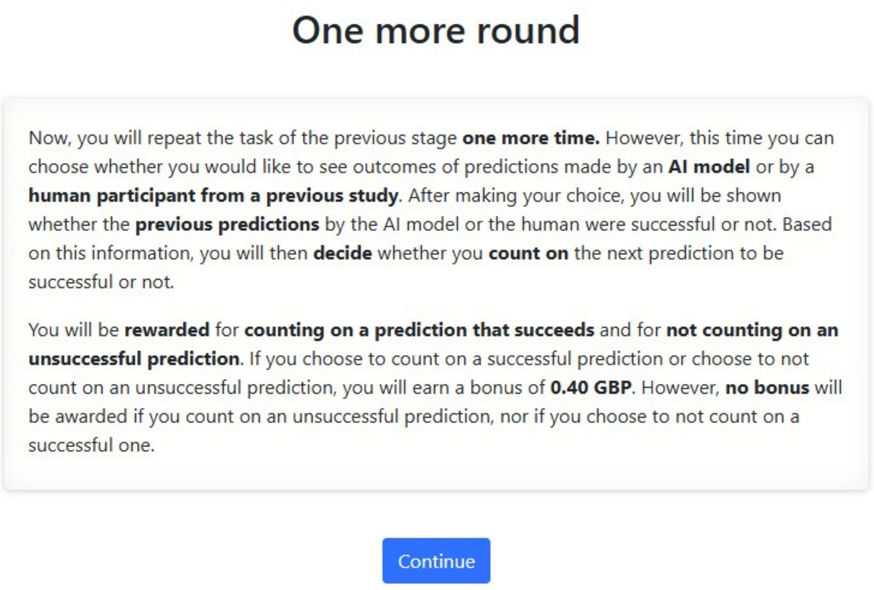


Guessing Stage Summary

Below, you find a summary of your results from the previous stage.

You made **12 correct decision(s)** (out of 24).

Stage 2 instructions, AI and human treatment

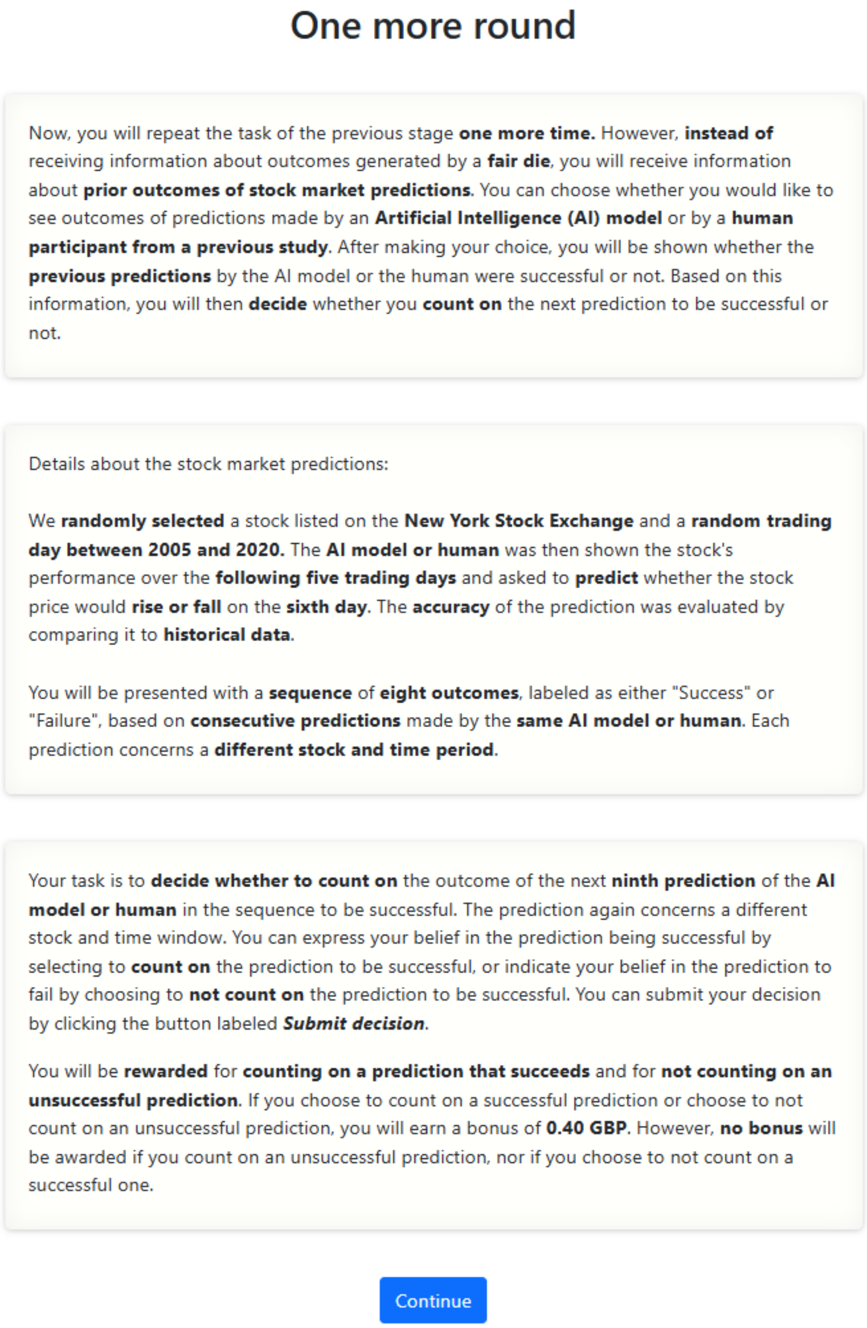


One more round

Now, you will repeat the task of the previous stage **one more time.** However, this time you can choose whether you would like to see outcomes of predictions made by an **AI model** or by a **human participant from a previous study**. After making your choice, you will be shown whether the **previous predictions** by the AI model or the human were successful or not. Based on this information, you will then **decide** whether you **count on** the next prediction to be successful or not.

You will be **rewarded** for **counting on a prediction that succeeds** and for **not counting on an unsuccessful prediction**. If you choose to count on a successful prediction or choose to not count on an unsuccessful prediction, you will earn a bonus of **0.40 GBP**. However, **no bonus** will be awarded if you count on an unsuccessful prediction, nor if you choose to not count on a successful one.

Stage 2 instructions, dice treatment



One more round

Now, you will repeat the task of the previous stage **one more time.** However, **instead of** receiving information about outcomes generated by a **fair die**, you will receive information about **prior outcomes of stock market predictions**. You can choose whether you would like to see outcomes of predictions made by an **Artificial Intelligence (AI) model** or by a **human participant from a previous study**. After making your choice, you will be shown whether the **previous predictions** by the AI model or the human were successful or not. Based on this information, you will then **decide** whether you **count on** the next prediction to be successful or not.

Details about the stock market predictions:

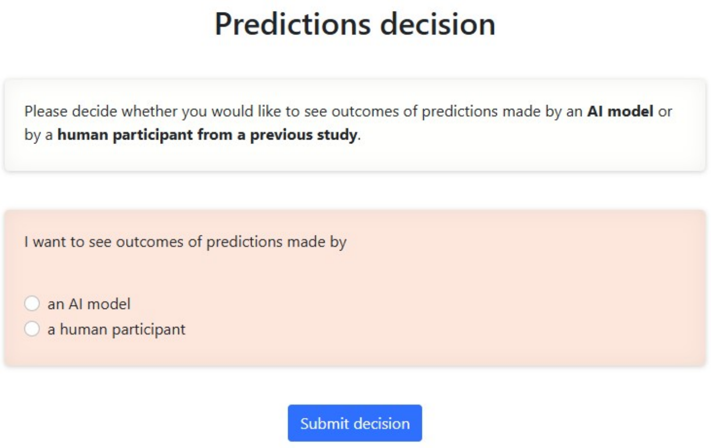
We **randomly selected** a stock listed on the **New York Stock Exchange** and a **random trading day between 2005 and 2020.** The **AI model or human** was then shown the stock's performance over the **following five trading days** and asked to **predict** whether the stock price would **rise or fall** on the **sixth day**. The **accuracy** of the prediction was evaluated by comparing it to **historical data**.

You will be presented with a **sequence** of **eight outcomes**, labeled as either "Success" or "Failure", based on **consecutive predictions** made by the **same AI model or human**. Each prediction concerns a **different stock and time period**.

Your task is to **decide whether to count on** the outcome of the next **ninth prediction** of the **AI model or human** in the sequence to be successful. The prediction again concerns a different stock and time window. You can express your belief in the prediction being successful by selecting to **count on** the prediction to be successful, or indicate your belief in the prediction to fail by choosing to **not count on** the prediction to be successful. You can submit your decision by clicking the button labeled ***Submit decision***.

You will be **rewarded** for **counting on a prediction that succeeds** and for **not counting on an unsuccessful prediction**. If you choose to count on a successful prediction or choose to not count on an unsuccessful prediction, you will earn a bonus of **0.40 GBP**. However, **no bonus** will be awarded if you count on an unsuccessful prediction, nor if you choose to not count on a successful one.

Stage 2 prediction source choice



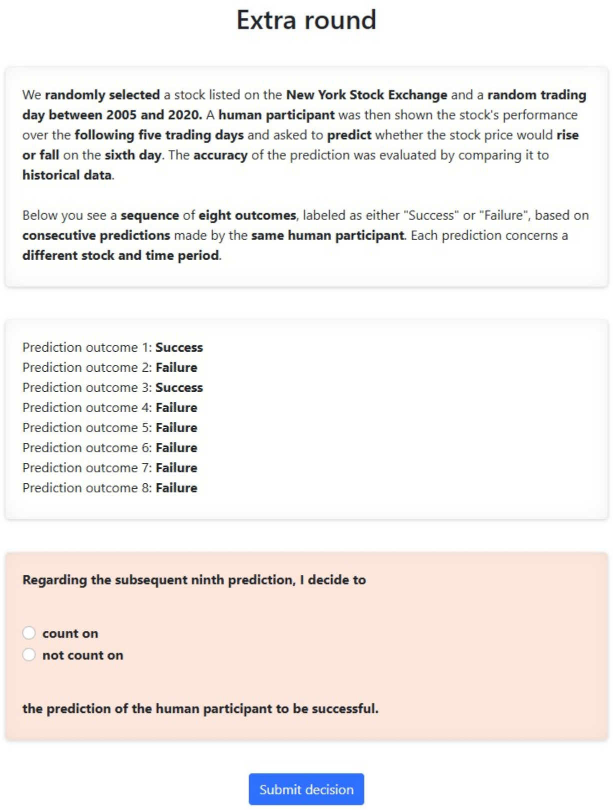
Predictions decision

Please decide whether you would like to see outcomes of predictions made by an **AI model** or by a **human participant from a previous study**.

I want to see outcomes of predictions made by

○ an AI model
○ a human participant

Final round, human treatment



Extra round

We **randomly selected** a stock listed on the **New York Stock Exchange** and a **random trading day between 2005 and 2020.** A **human participant** was then shown the stock's performance over the **following five trading days** and asked to **predict** whether the stock price would **rise or fall** on the **sixth day**. The **accuracy** of the prediction was evaluated by comparing it to **historical data**.

Below you see a **sequence** of **eight outcomes**, labeled as either "Success" or "Failure", based on **consecutive predictions** made by the **same human participant**. Each prediction concerns a **different stock and time period**.

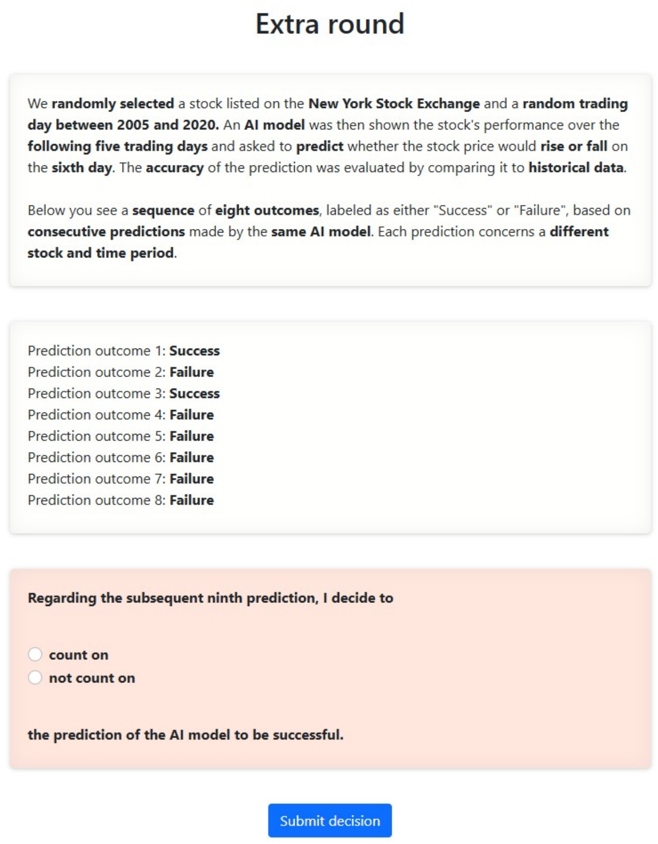
Prediction outcome 1: **Success**
Prediction outcome 2: **Failure**
Prediction outcome 3: **Success**
Prediction outcome 4: **Failure**
Prediction outcome 5: **Failure**
Prediction outcome 6: **Failure**
Prediction outcome 7: **Failure**
Prediction outcome 8: **Failure**

Regarding the subsequent ninth prediction, I decide to

○ **count on**
○ **not count on**

the prediction of the human participant to be successful.

Final round, AI treatment



Extra round

We **randomly selected** a stock listed on the **New York Stock Exchange** and a **random trading day between 2005 and 2020.** An **AI model** was then shown the stock's performance over the **following five trading days** and asked to **predict** whether the stock price would **rise or fall** on the **sixth day**. The **accuracy** of the prediction was evaluated by comparing it to **historical data**.

Below you see a **sequence** of **eight outcomes**, labeled as either "Success" or "Failure", based on **consecutive predictions** made by the **same AI model**. Each prediction concerns a **different stock and time period**.

Prediction outcome 1: **Success**
Prediction outcome 2: **Failure**
Prediction outcome 3: **Success**
Prediction outcome 4: **Failure**
Prediction outcome 5: **Failure**
Prediction outcome 6: **Failure**
Prediction outcome 7: **Failure**
Prediction outcome 8: **Failure**

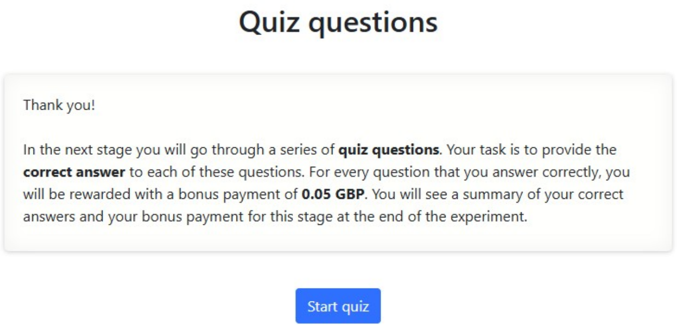
Regarding the subsequent ninth prediction, I decide to

○ **count on**
○ **not count on**

the prediction of the AI model to be successful.

C.4 Debrief and control elicitation

C.4.1 Instructions for CRT and statistical literacy elicitation



Quiz questions

Thank you!

In the next stage you will go through a series of **quiz questions**. Your task is to provide the **correct answer** to each of these questions. For every question that you answer correctly, you will be rewarded with a bonus payment of **0.05 GBP**. You will see a summary of your correct answers and your bonus payment for this stage at the end of the experiment.

Statistical literacy

1. *First, suppose this bowl has 10 white balls and no red balls. You will be asked to draw one ball without looking. On a scale from 0 percent to 100 percent, what is the percent chance that the ball you draw is red?*

   – Numeric input (0–100) – **correct answer 0**

2. *Now suppose that the bowl has 7 white balls and 3 red balls. You will be asked to draw one ball without looking. On a scale from 0 percent to 100 percent, what is the percent chance that the ball you draw is white?*

   – Numeric input (0–100) – **correct answer 70**

3. *Imagine that the weather report tells you that the chance it will rain tomorrow is 70%. Assume that the weather report accurately reports the chance of rain. On a scale from 0 percent to 100 percent, what is the chance it will NOT rain tomorrow?*

   – Numeric input (0–100) – **correct answer 30**

4. *Imagine that whether it rains in your town and whether it rains in New York are unrelated. The chance that it will rain in your town tomorrow is 50%. The chance that it will rain in New York is also 50%. On a scale from 0 percent to 100 percent, what is the chance that it will rain both in your town and in New York tomorrow?*

   – Numeric input (0–100) – **correct answer 25**

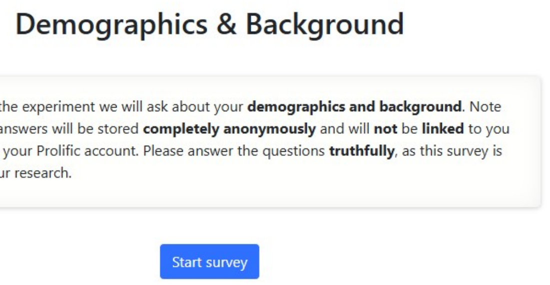
CRT questions

1. *If you are running a race and you pass the person in second place, what place are you in?*

- Open numeric response – **correct answer 2**

2. *A farmer had 15 sheep and all but 8 died. How many are left?*
   - Open numeric response – **correct answer 8**
3. *How many cubic feet of dirt are there in a hole that is 3 feet deep, 3 feet wide, 3 feet long?*
   - Open numeric response – **correct answer 0**

C.4.2 Demographics and background elicitation



Demographics & Background

In this stage of the experiment we will ask about your **demographics and background**. Note that all of your answers will be stored **completely anonymously** and will **not** be **linked** to you personally or to your Prolific account. Please answer the questions **truthfully**, as this survey is important for our research.

AI expertise and beliefs

1. *I am an expert regarding AI models.*
   - Strongly Agree; Agree; Disagree; Strongly Disagree
2. *I am worried that risks associated with AI outweigh the benefits.*
   - Strongly Agree; Agree; Disagree; Strongly Disagree

C.4.3 Fallacy score

1. *When a fair coin toss comes up Heads, it is more likely that the next toss of the same coin also comes up Heads.*
   - Strongly Agree; Agree; Disagree; Strongly Disagree
2. *When a fair coin toss comes up Heads, it is more likely that the next toss of the same coin comes up Tails.*
   - Strongly Agree; Agree; Disagree; Strongly Disagree

3. *It is possible to increase the quality of predicting the toss of a fair coin by considering previous outcomes and applying a systematic strategy.*

    – Strongly Agree; Agree; Disagree; Strongly Disagree

C.4.4 Demographics

1. *What is your highest level of education?*

    – No formal education; Primary education; Secondary education (High school); Bachelor degree; Master degree; PhD or higher

2. *What is your profession?*

    – No Profession; Arts and Entertainment; Business; Industrial and Manufacturing; Law Enforcement and Armed Forces; Science and Technology; Healthcare and Medicine; Other

3. *What is your (approximate) annual income (in GBP)?*

    – below 10,000; between 10,000 and 20,000; between 20,000 and 30,000; between 30,000 and 40,000; between 40,000 and 50,000; higher than 50,000

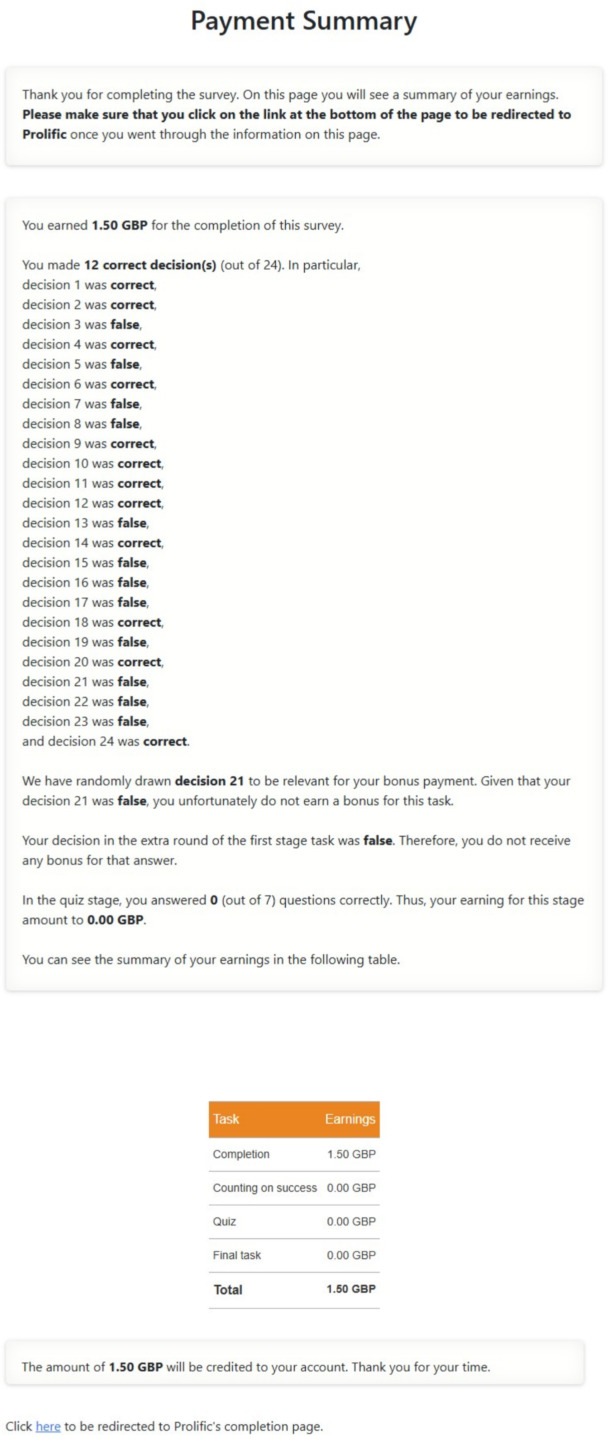
4. *What is your gender?*

    – Female; Male; Non-binary; Prefer not to say

5. *What is your age?*

    – Numeric input

C.5 Summary page



Payment Summary

Thank you for completing the survey. On this page you will see a summary of your earnings. **Please make sure that you click on the link at the bottom of the page to be redirected to Prolific** once you went through the information on this page.

You earned **1.50 GBP** for the completion of this survey.

You made **12 correct decision(s)** (out of 24). In particular,
decision 1 was **correct**,
decision 2 was **correct**,
decision 3 was **false**,
decision 4 was **correct**,
decision 5 was **false**,
decision 6 was **correct**,
decision 7 was **false**,
decision 8 was **false**,
decision 9 was **correct**,
decision 10 was **correct**,
decision 11 was **correct**,
decision 12 was **correct**,
decision 13 was **false**,
decision 14 was **correct**,
decision 15 was **false**,
decision 16 was **false**,
decision 17 was **false**,
decision 18 was **correct**,
decision 19 was **false**,
decision 20 was **correct**,
decision 21 was **false**,
decision 22 was **false**,
decision 23 was **false**,
and decision 24 was **correct**.

We have randomly drawn **decision 21** to be relevant for your bonus payment. Given that your decision 21 was **false**, you unfortunately do not earn a bonus for this task.

Your decision in the extra round of the first stage task was **false**. Therefore, you do not receive any bonus for that answer.

In the quiz stage, you answered **0** (out of 7) questions correctly. Thus, your earning for this stage amount to **0.00 GBP**.

You can see the summary of your earnings in the following table.

Task	Earnings
Completion	1.50 GBP
Counting on success	0.00 GBP
Quiz	0.00 GBP
Final task	0.00 GBP
Total	**1.50 GBP**

The amount of **1.50 GBP** will be credited to your account. Thank you for your time.

Click here to be redirected to Prolific's completion page.

D Full summary statistics across treatment

Table D.1: Summary statistics for Dice and AI

	NegFeedback		**PosFeedback**	
	Mean	Median	Mean	Median
Dice				
Times participant counted on prediction	12.71	12.00	12.70	12.00
Number of correct decisions	10.56	10.00	14.06	15.50
Participant chose AI in stage 2	0.84	1.00	0.79	1.00
Total payoff from participation (in GBP)	0.71	0.60	0.79	0.83
Statistical literacy questions correct out of 4 total	3.22	3.00	3.38	4.00
Cognitive reflection test questions correct out of 3 total	1.93	2.00	1.98	2.00
What is your highest level of education?	4.05	4.00	3.81	4.00
What is your profession?	5.18	6.00	5.47	6.00
What is your (approximate) annual income (in GBP)?	3.45	3.00	3.46	3.00
What is your gender?	1.51	2.00	1.57	2.00
What is your age?	44.73	42.00	40.88	39.00
I am an expert regarding AI models	3.21	3.00	3.04	3.00
I am worried that risks associated with AI outweigh the benefits	2.48	3.00	2.55	3.00
Fallacy score	2.04	3.00	2.19	3.00
AI				
Times participant counted on prediction	12.30	12.00	12.53	12.00
Number of correct decisions	8.00	7.00	16.45	18.00
Participant chose AI in stage 2	0.47	0.00	0.65	1.00
Total payoff from participation (in GBP)	0.69	0.58	0.91	1.10
Statistical literacy questions correct out of 4 total	3.36	4.00	3.32	3.00
Cognitive reflection test questions correct out of 3 total	1.92	2.00	1.87	2.00
What is your highest level of education?	3.91	4.00	3.83	4.00
What is your profession?	5.47	6.00	5.40	6.00
What is your (approximate) annual income (in GBP)?	3.59	4.00	3.34	3.00
What is your gender?	1.48	1.00	1.45	1.00
What is your age?	40.52	39.00	43.39	43.50
I am an expert regarding AI models	3.03	3.00	3.02	3.00
I am worried that risks associated with AI outweigh the benefits	2.50	3.00	2.55	3.00
Fallacy score	2.10	3.00	2.10	3.00

Note: For the statements “I am...”, participants had four options: 1 = ”strongly disagree”, 2 = ”disagree”, 3 = ”agree”, 4 = ”strongly agree”. Fallacy score represents the number of fallacy-indicating answers out of three fallacy questions.

Table D.2: Summary statistics for Human and full sample

	NegFeedback		PosFeedback	
	Mean	Median	Mean	Median
Human				
Times participant counted on prediction	12.65	12.00	11.80	12.00
Number of correct decisions	7.94	7.00	16.22	18.00
Participant chose AI-algorithm in stage 2	0.88	1.00	0.75	1.00
Total payoff from participation (in GBP)	0.65	0.55	0.94	1.15
Statistical literacy questions correct out of 4 total	3.30	3.00	3.34	4.00
Cognitive reflection test questions correct out of 3 total	1.88	2.00	1.86	2.00
What is your highest level of education?	3.91	4.00	4.01	4.00
What is your profession?	5.18	6.00	5.36	6.00
What is your (approximate) annual income (in GBP)?	3.47	3.00	3.42	3.00
What is your gender?	1.54	2.00	1.43	1.00
What is your age?	40.82	39.00	42.20	42.00
I am an expert regarding AI models	2.92	3.00	3.13	3.00
I am worried that risks associated with AI outweigh the benefits	2.58	3.00	2.43	2.00
Fallacy score	2.09	3.00	1.97	3.00
Full sample				
Times participant counted on prediction	12.56	12.00	12.36	12.00
Number of correct decisions	8.87	7.00	15.59	17.00
Participant chose AI-algorithm in stage 2	0.74	1.00	0.73	1.00
Total payoff from participation (in GBP)	0.68	0.55	0.88	1.00
Statistical literacy questions correct out of 4 total	3.29	3.00	3.35	3.00
Cognitive reflection test questions correct out of 3 total	1.91	2.00	1.90	2.00
What is your highest level of education?	3.96	4.00	3.88	4.00
What is your profession?	5.27	6.00	5.41	6.00
What is your (approximate) annual income (in GBP)?	3.50	3.00	3.40	3.00
What is your gender?	1.51	2.00	1.49	1.00
What is your age?	42.09	39.00	42.19	42.00
I am an expert regarding AI models	3.06	3.00	3.06	3.00
I am worried that risks associated with AI outweigh the benefits	2.52	3.00	2.52	3.00
Fallacy score	2.08	3.00	2.09	3.00

Note: For the statements “I am...”, participants had four options: 1 = ”strongly disagree”, 2 = ”disagree”, 3 = ”agree”, 4 = ”strongly agree”. Fallacy score represents the number of fallacy-indicating answers out of three fallacy questions.

E Detailed variable descriptions

Identifying variables

- $i \in I$ – subjects
- $t_1 \in 0, 1, 2$ – stage 1 treatment of Dice, AI, Human
- $t_2 \in 0, 1$ – stage 2 treatment of NegFeedback, PosFeedback
- $j \in \{1, \ldots, 24\}$ – sequences
- $c \in \{1, \ldots, 9\}$ – outcomes within a sequence
- $p_{j,c} \in \{0, 1\}$ – binary indicator: $= 1$ if outcome c in sequence j was successful, otherwise $= 0$
- $r \in \{1, ..., 12\}$ – reversion sequence pair id
- $e \in \{1, ..., 12\}$ – inversion sequence pair id

Outcome variables

- $decision^i_j \in \{0, 1\}$ – for H1, H2, H2A; binary variable: $= 1$ if subject i counts on the subsequent ninth outcome in sequence j to be correct, otherwise $= 0$
- $inversion\_reaction^i \in \{-1, 1\}$ – for H1; average difference in subject i's decisions as reaction to inversion defined as

$$= \frac{1}{12} \left(\sum_{j=13}^{24} decision^i_j - \sum_{j=1}^{12} decision^i_j \right)$$

- $reversion\_reaction^i \in \{-1, 1\}$ – for H2; average difference in subject i's decisions as reaction to reversion defined as

$$= \frac{1}{8} \left(\sum_{j=1}^{4} decision^i_j - \sum_{j=5}^{8} decision^i_j + \sum_{j=17}^{20} decision^i_j - \sum_{j=13}^{16} decision^i_j \right)$$

- $reversion\_reaction\_streak^i \in \{-1, 1\}$ – for H2A; average difference in subject i's decisions as reaction to reversion in sequences with streaks of successes defined as

$$= \frac{1}{4} \left(\sum_{j=17}^{20} decision^i_j - \sum_{j=13}^{16} decision^i_j \right)$$

- $choose\_same\_source^i \in \{0, 1\}$ – for H3; binary variable: $= 1$ if $choose\_ai^i = 0$ and $t_1 = 1$ or $choose\_ai^i = 1$ and $t_1 = 2$, otherwise $= 0$

- $choose\_ai^i \in \{0,1\}$ – for H3; binary variable: $=1$ if subject i chooses an AI generated sequence in stage 2, otherwise $=0$

Main independent variables

- $success_j \in \{0,1\}$ – binary variable

$$= \begin{cases} 1 & \text{if } \sum_{c=1}^{8} \mathbb{1}(p_{j,c}=1) > \sum_{c=1}^{8} \mathbb{1}(p_{j,c}=0) \\ 0 & \text{otherwise} \end{cases}$$

- $recent\_success_j \in \{0,1\}$ – binary variable

$$= \begin{cases} 1 & \text{if } \sum_{c=5}^{8} \mathbb{1}(p_{j,c}=1) > \sum_{c=1}^{4} \mathbb{1}(p_{j,c}=1) \\ 0 & \text{otherwise} \end{cases}$$

- $recent\_success\_streak_j \in \{0,1\}$ – binary variable

$$= \begin{cases} 1 & \text{if } p_{j,c}=1 \ \forall \ c \in \{5,...,8\} \\ 0 & \text{otherwise} \end{cases}$$

Control variables

- $stage1\_correct^i \in \{0,\dots,24\}$ – subject i's number of correct stage 1 decisions[26]
- **Statistical literacy:** $\gamma^i \in [0,4]$ – number of correct answers by subject i (out of four statistical literacy questions)
- **Cognitive reflection test (CRT):** $\vartheta^i \in [0,3]$ – number of correct answers by subject i (out of three CRT questions)
- **AI–expertise:** $\rho^i \in [1,4]$ – self–reported AI expertise level
- **AI–belief:** $\theta^i \in [1,4]$ – self–reported perception of AI risks and benefits
- **Fallacy score:** $\phi^i \in [0,3]$ – number of fallacy–indicating answers by subject i (out of three fallacy questions)
- **Demographics:** χ^i – $m \times 1$ vector constituting self–reported demographic variables age^i, $gender^i$, $profession^i$, $income^i$, $education^i$

[26] This definition differs from our pre-registration, which proposed a shifted version of this variable and—mistakenly—an inverted transformation (12–$correct\_answers$). For clarity and interpretability, we use the straightforward count of correct responses.

F Predicted probabilities for other control variables

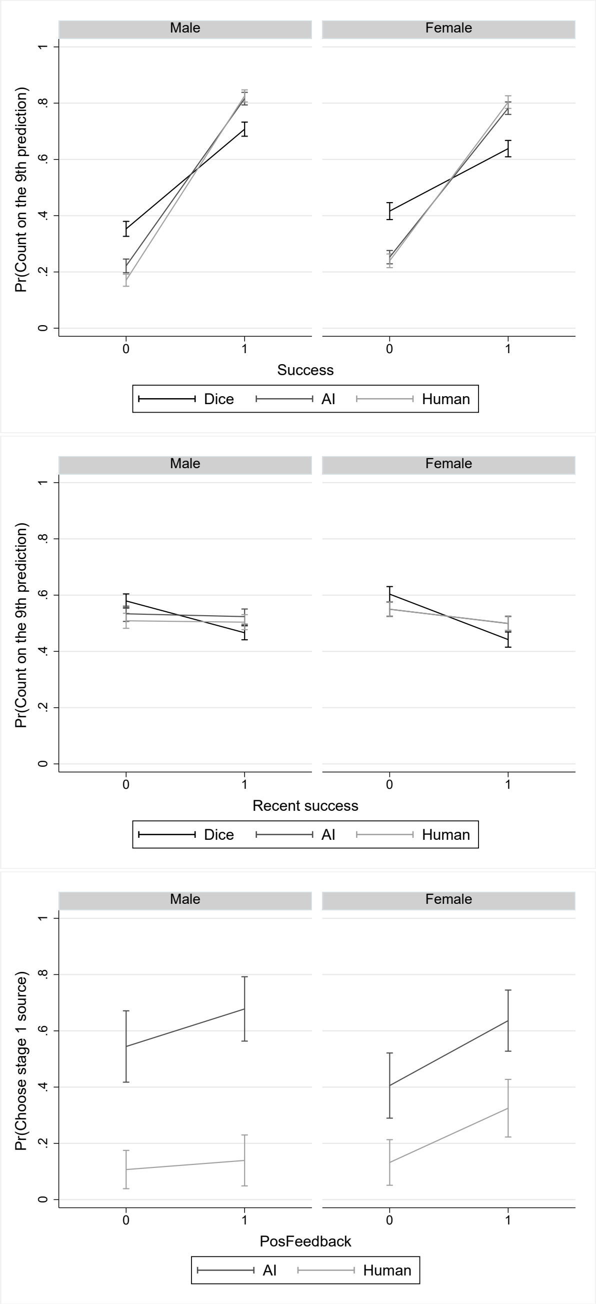


Figure F.1: Predicted probabilities for interaction terms in models (1) (top panel), (2) (middle panel) and (3) (bottom panel) when including additional gender interaction. Individual-level control variables are included in the models.

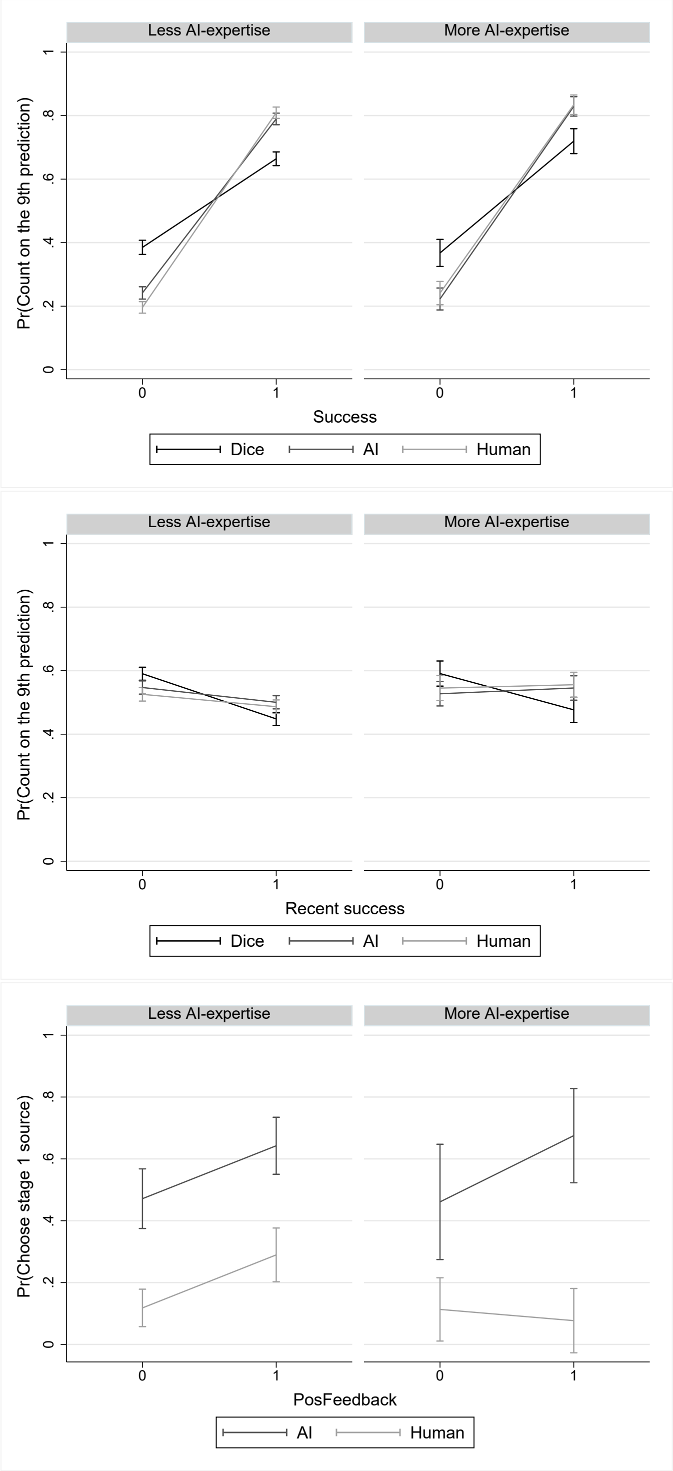


Figure F.2: Predicted probabilities for interaction terms in models (1) (top panel), (2) (middle panel) and (3) (bottom panel) when including additional AI-expertise interaction. Individual-level control variables are included in the models.

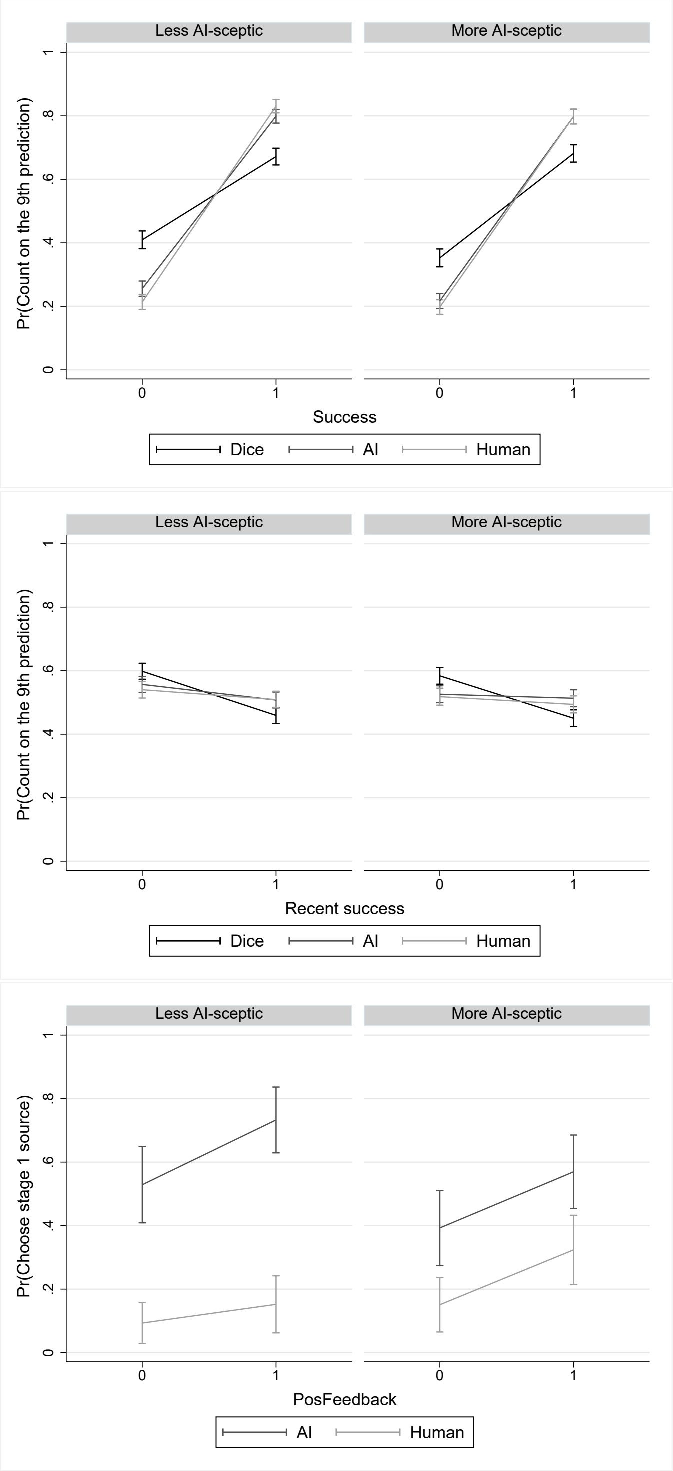


Figure F.3: Predicted probabilities for interaction terms in models (1) (top panel), (2) (middle panel) and (3) (bottom panel) when including additional AI-scepticism interaction. Individual-level control variables are included in the models.