CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

University of Nottingham
UK | CHINA | MALAYSIA

Discussion Paper No. 2026-02

Dominik Suri, Simon Gächter and Sebastian Kube

February 2026

# AI versus humans as authority figures: Evidence from a rule-compliance experiment

CeDEx Discussion Paper Series
ISSN 1749 - 3293


CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Samantha Stapleford-Allen
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 74 86214
Samantha.Stapleford-Allen@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

# AI versus humans as authority figures: Evidence from a rule-compliance experiment

Dominik Suri Simon Gächter Sebastian Kube

February 2026

**Abstract:** AI-driven systems are rapidly moving from decision support to directing human behavior through rules, recommendations, and compliance requests. This shift expands everyday human–AI interaction and raises the possibility that AI may function as an authority figure. However, the behavioral consequences of AI as an authority figure remain poorly understood. We investigate whether individuals differ in their willingness to comply with arbitrary rules depending on whether these rules are attributed to an AI agent (ChatGPT) or to a fellow human. In a between-subject design, 977 US Prolific users completed the coins task: they could earn a monetary payoff by stopping the disappearance of coins at any time, but a rule instructed them to wait for a signal before doing so. There are no conventional reasons to follow this rule: complying is costly and nobody is harmed by non-compliance. Despite this, we find high rule-following rates: 64.3% followed the rule set by ChatGPT and 63.9% complied with the human-set rule. Descriptive and normative beliefs about rule following, as well as compliance conditional on these beliefs, are also largely unaffected by the rule's origin. However, subjective social closeness to the rule setter significantly predicts how participants condition their behavior on social expectations: when participants perceive the rule setter as subjectively closer, conditional compliance is higher and associated beliefs are stronger, irrespective of whether the rule setter is human or AI.

**JEL classification codes:** C91, D91, Z13.

**Keywords:** Artificial intelligence, AI-human interaction, ChatGPT, rule-following, coins task, CRISP framework, social expectations, conditional rule conformity, social closeness, IOS11, online experiments.

---

**Contact:** Suri (corresponding author): Institute for Food and Resource Economics & Center for Economics and Neuroscience, University of Bonn, dsuri@uni-bonn.de. Gächter: Centre for Decision Research and Experimental Economics, University of Nottingham & Institute of Labour Economics, Bonn & CESifo, Munich, Simon.Gaechter@nottingham.ac.uk. Kube: Center for Economics and Neuroscience & Institute for Applied Microeconomics, University of Bonn, kube@uni-bonn.de. **Acknowledgments:** We thank Maren Bermúdez for excellent research assistance. **Funding:** This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy [EXC 2126/1-390838866] and the European Research Council [ERC-AdG 101020453 PRINCIPLES]. **Ethics approval:** Our research complies with all relevant ethical regulations. The study received ethical clearance from the School of Economics Research Ethics Committee at the University of Nottingham (protocol ERCP-2025-056). Informed consent was obtained from all participants. **Research transparency:** The main research hypotheses, survey design and sampling approach were pre-registered on AsPredicted: https://aspredicted.org/am8si2.pdf. A replication package will be made available on OSF.

# 1 Introduction

Artificial intelligence (AI) is rapidly diffusing across a wide range of societal domains.[1] It is transforming medicine and health care (Zhang et al., 2020; Capponi & Daniels, 2023; Biesheuvel et al., 2024), generating efficiency gains and productivity increases in industry (Bharadiya et al., 2023; Rana & Khatri, 2024), supporting sustainability transitions and environmental goals (Satornino et al., 2024; Zejjari & Benhayoun, 2024; Bergougui, 2025), and fostering creativity and idea generation (Lee & Chung, 2024). At the same time, AI raises substantial concerns. It may threaten employment (Krittanawong, 2018; Mirbabaie et al., 2022; Orchard & Tasiemski, 2023), reduce customer acceptance in service contexts (Xu et al., 2024), and undermine research integrity (Eke, 2023; Yusuf et al., 2024), particularly in online and experimental settings (Westwood, 2025). Related to the topic of this paper, AI may also be used as an instrument of social control (e.g., Zuboff, 2019; Tirole, 2021; Beraja et al., 2023; Roberts & Oosterom, 2025). Reflecting these opposing forces, public attitudes toward AI are mixed (Zhang & Dafoe, 2019; Liehner et al., 2023).

Beyond these societal impacts, AI is increasingly capable of engaging in behaviors that resemble human decision-making. Recent research using large language models has shown that these systems can exhibit a wide range of human-like decision patterns in both strategic and social contexts, including cooperation, fairness, trust, and other social preferences (Mei et al., 2024). By systematically varying how models are prompted, researchers can elicit different patterns of behavior and categorize the strategic contexts that produce them (Xie et al., 2025).[2] They can even be a useful tool to simulate pilot data before testing paradigms in the real world (Horton, 2023). These findings suggest not only the sophisticated behavioral flexibility of AI in simulated interactions but also motivate further study into how humans perceive and respond to decisions when AI plays a role. Understanding the fundamental mechanisms that govern human behavior when interacting with AI has therefore become an increasingly important research agenda; not least because AI chatbots respond to human queries in a human-like and authoritative-looking fashion (Chugunova & Sele, 2022; Korinek, 2023; Church, 2024; Naik et al., 2024).

In this paper, we focus on the role of AI versus humans as authority figures and their influence on human rule-following behavior. This is interesting because AI systems are increasingly built as supporting devices. They often become advisors to human decision-making (Köbis et al., 2021), and humans tend to follow their advices (Leib et al., 2024; Walter et al., 2024; Greiner et al., 2025; Yin et al., 2025).[3] But what will happen if AI moves from being taken as an advice giver toward being perceived as a (de facto) rule setter? Do people follow rules set by an AI machine?

Following rules is a fundamental social behavior in all human societies. Rules are a demand on behavior ("Do *x*!", "Don't do *y*!") and a cornerstone of social order (e.g., Brennan & Buchanan, 1985; Kliemt, 2020), structuring and stabilizing social life (Bicchieri, 2006; Gelfand,

[1] The term AI is used to describe technologies that can execute tasks one may expect to require human intelligence (Kaplan, 2016).

[2] Note, however, that large language models are sensitive to framing and order effects (Polachek et al., 2025).

[3] For a broader look at the role of algorithms as advisors, please refer to, for example, Normann et al. (2025) and Hunold and Werner (2025) who investigate this in the field of collusion, or Logg et al. (2019) who discuss algorithmic appreciation (following advice from algorithms) more generally, or Chugunova and Sele (2022) who provide a discussion on algorithmic appreciation versus algorithmic aversion (not following advice from algorithms).

2018; Daston, 2022). They are followed, especially if they come from authority figures (Milgram, 1963; Burger, 2009). While humans typically expect AI systems to follow rules designed by humans (being the authority figures), it remains an open question whether this relationship also operates in the opposite direction. Popular culture—for example, the "Terminator" or "Matrix" franchises—often depicts scenarios in which artificial agents impose rules on humans, often with dystopian overtones. Abstracting from such narratives, a fundamental question emerges: are individuals willing to comply with rules set by AI agents? Can an AI chatbot tell us what to do when we have incentives not to follow its rules? This paper provides an experimental investigation of rule compliance when the origin of the rule is artificial intelligence as compared to human intelligence.

We study this question in an online experiment with 977 US Americans recruited via Prolific Academic, drawn from all major regions of the United States (average age 42.3 years, 49.5% male). Rule compliance is measured using the "coins task" (Suri et al., 2025), which makes it very salient that following the rule means losing money: Participants observe coins disappearing over time and can stop this process at any moment to secure the remaining coins as a payoff. A non-binding rule instructs participants to wait for a specific signal before stopping the process. We implement a between-subject design with two treatment conditions that vary only in the origin of the rule: in one condition, the rule is set by a fellow US Prolific user (`HUMAN`); in the other, the rule originates from ChatGPT (`AI`).

To understand the nature of rule-following, we are guided by the interdisciplinary and experimentally validated CRISP framework (Gächter et al., 2025). CRISP explains compliance with rules ($C$) as a function of unconditional respect for rules ($R$), extrinsic incentives ($I$), conformity with social expectations ($S$), and social preferences ($P$). Our experiment incentivizes people to break the rule because monetary incentives are always negative ($I < 0$), and rule-breaking has no effect on others, rendering social preferences ($P$) irrelevant. Therefore, because $I < 0$ and $P = 0$, there are no conventional reasons for following the rule and rule following can only occur if people follow it out of respect for the rule ($R$) or because they conform with social expectations ($S$) that others think one should follow the rule (normative beliefs) or that others actually follow the rule (descriptive beliefs). We measure $R$ and $S$ by eliciting participants' normative and descriptive beliefs, as well as people's rule compliance conditional on these beliefs.

We find high levels of rule compliance when the rule is set by an AI agent: in the `AI` condition, 64.27% of participants follow the rule. This is strikingly similar to rule following when the rule is issued by another human: 63.88% of participants comply in the `HUMAN` condition. The rule-following rates in the `HUMAN` condition are consistent with a growing literature documenting remarkably stable and high levels of rule-following behavior in abstract tasks (Kimbrough & Vostroknutov, 2016, 2018; Kimbrough et al., 2024; Gächter et al., 2025; Suri et al., 2025; Bicchieri et al., 2025; Suri et al., 2026).

Replicating prior findings, we observe that compliance is strongly conditional on beliefs about others' disapproval of rule violations and others' conformity with the rule. However, these belief-driven patterns do not differ between the `AI` and `HUMAN` conditions. Instead, what systematically predicts belief formation and conditional rule compliance is the participant's *perceived psychological closeness* to the rule setter, measured using the "Inclusion of Other in the Self" (IOS11) scale (Baader et al., 2024). For a given type of rule setter, when she/it is

perceived as subjectively close beliefs are more pronounced and conditional compliance is significantly higher.

Our paper contributes to two strands of literature. First, we contribute to the literature on rule-following behavior, beginning with Kimbrough and Vostroknutov (2016), who introduce the traffic light task. In this task, participants navigate an avatar past traffic lights that initially display red and eventually turn green. Participants earn higher payoffs the faster they reach the destination, yet a rule instructs them to wait until the light turns green. Despite the absence of enforcement, 62.5% of participants comply with the rule. Moreover, the authors show that rule followers subsequently behave more cooperatively in public goods games. Using a simplified version of the same task, Gächter et al. (2025) report average compliance rates of 65%; and 60% in a more abstract variant in which participants move an avatar into a box and are instructed to wait for a signal change before proceeding.

To explain rule-following, Gächter et al. (2025) developed the CRISP framework and showed that both $R$ and $S$ explain rule-following when $I$ suggests breaking the rule and $P$ is irrelevant. Gächter et al. (2025) also provide evidence for the compliance-enhancing influence of $I > 0$ and $P > 0$ when there are externalities for breaking the rule (and $P$ therefore can matter) and when rule breaking is punished severely enough, rendering $I > 0$. In this paper, we study the most basic setting with no externalities (rendering $P$ irrelevant) and no punishment for rule breaking (keeping $I < 0$), because it allows us to focus on the pure effect of potential differences in the perceived authority of AI compared to a human rule setter.

Related experiments employ alternative task designs to study rule compliance. Kimbrough and Vostroknutov (2018) introduce the bucket task, in which participants allocate balls between two buckets, one of which yields lower payoffs per ball despite being prescribed by the rule. They observe rule conformity of 58%. Using the same task in a panel design, Kimbrough et al. (2024) find compliance rates of 60.8% and 62.2% in two waves conducted one month apart. Bicchieri et al. (2025) use the Y-task (named after the Y-shaped maze where people have to move a token to either the left or right branch) and find rule-following rates between 51.2% and 61.6%. Suri et al. (2025) introduce the coins task employed in the present study and report a compliance rate of 58%. Using the same task, Suri et al. (2026) find a closely comparable rule-following rate of 57.8%.

A common feature of these studies is that the rule is set by the experimenter, who is a human being. An important exception is Suri et al. (2026), who ran experiments on Prolific where in some treatments the rule setter is another Prolific worker rather than the experimenter, though still a human being. In the condition most comparable to our `HUMAN` treatment, they report a compliance rate of 55%.

While existing work exclusively focuses on human rule setters, we extend this literature by demonstrating that rule compliance remains equally high when the rule originates from an AI agent. In addition, we provide a within-subject analysis of the full set of social expectations emphasized by the CRISP framework, combining unconditional and conditional rule compliance with detailed elicitation of normative and descriptive beliefs; and considering subjective closeness to the rule setter. We conduct this analysis separately for human and AI rule setters. Finally, our findings reinforce the role of perceived social proximity in shaping rule compliance (see Suri et al., 2026), highlighting promising avenues for future research on the causal mechanisms underlying this relationship.

Second, we contribute to the growing experimental literature on human-AI interactions, which has largely focused on trust in AI systems. In trust games, Livingston et al. (2025) show that participants send, on average, 54% of their endowment to ChatGPT, and still transfer 41% even after learning that ChatGPT is prompted to act selfishly. Comparing human-AI and human-human interactions, participants exhibit similar (Jayasekara et al., 2025) or even higher (de Boer et al., mimeo) levels of trust toward AI agents in such games. de Boer et al. (mimeo) further demonstrate that this pattern persists when decisions are made jointly in groups interacting with ChatGPT. In contrast, in a personnel pre-screening context, Langer et al. (2023) find that participants trust fellow human colleagues more than AI-based systems when evaluating applicants.

In a series of ultimatum, trust, prisoner's dilemma, stag hunt, and coordination games, Dvorak et al. (2025) examine the effects of ChatGPT acting on behalf of human participants. Relative to human–human interactions, they find that AI involvement systematically leads to payoff-decreasing outcomes, characterized by lower levels of fairness, trust, trustworthiness, cooperation, and coordination success. Importantly, these effects arise both when participants explicitly delegate decision-making authority to the AI and when the AI makes decisions autonomously from the outset.

Focusing on cooperation, Akata et al. (2025) examine strategic interaction with large language models in a prisoner's dilemma. They find that participants reciprocate cooperative behavior from AI less frequently than is typically observed in human-human interactions. However, cooperation increases when the chatbot is prompted to first predict the participant's action before choosing its own strategy. Kasberger et al. (2024) also study algorithmic cooperation in prisoner's dilemmas and find that algorithms cooperate less than humans when cooperation is risky or not incentive-compatible.

While this strand of literature has primarily examined whether individuals trust, cooperate with, or rely on AI agents in strategic or advisory contexts, much less is known about whether humans view AI as a legitimate authority and are willing to incur costs to comply with arbitrary AI-issued rules. This distinction matters because trust or instrumental reliance need not translate into normative obligation: AI may be treated as a tool that provides recommendations rather than an entity entitled to prescribe behavior. Our study complements the literature on human–AI interactions by isolating voluntary, unmonitored, and costly compliance, thereby providing a micro-foundation for how AI-generated rules may be received in social and institutional settings.

The remainder of the paper is structured as follows. Section 2 describes our experimental design including the coins task and belief measures in more detail. Section 3 presents our results. Lastly, Section 4 concludes with a brief discussion of our findings alongside future avenues of research.

## 2 Experimental design and procedures

In the following, we first describe the coins task used to measure rule compliance in Section 2.1. We then provide an overview of the two treatment conditions—`AI` and `HUMAN`—in Section 2.2. This is followed by a description of the belief elicitation and conditional rule compliance measures (Section 2.3), as well as the subjective closeness measure (Section 2.4). We

subsequently detail the experimental procedures in Section 2.5 and report information about the participant pool (Section 2.6). Our experiment received approval from the Research Ethics Committee in the School of Economics at the University of Nottingham (protocol ERCP-2025-056). The experimental design, hypotheses, and procedures were pre-registered on AsPredicted.org: https://aspredicted.org/am8si2.pdf. Detailed screenshots of the experimental instructions are provided in the Supplementary Information (SI).

## 2.1 Measuring rule compliance: The coins task

To measure rule compliance, we implement the coins task (introduced by Suri et al., 2025) that we also used in a related paper by Suri et al. (2026). The basic idea of the coins task builds on earlier designs by Kimbrough and Vostroknutov (2016) and Gächter et al. (2025) and constitutes a more abstract alternative to the traffic-light paradigm, with the key advantage that the cost of complying with the rule is both transparent and salient.

In the coins task, participants are initially presented with 20 coins displayed on the screen (see the left panel of Figure 1). One coin disappears every second. At any point—when 20, 19, ..., 1, or 0 coins remain—the participant may press the "Stop"-button[4] to stop the process. Upon doing so, she earns a monetary payoff equal to \$0.10 per remaining coin and must wait until the remaining seconds have elapsed before continuing with the experiment. Participants are explicitly informed that they may press the button at any time, implying that earnings range from a maximum of 20 coins (\$2) to a minimum of zero coins (\$0).

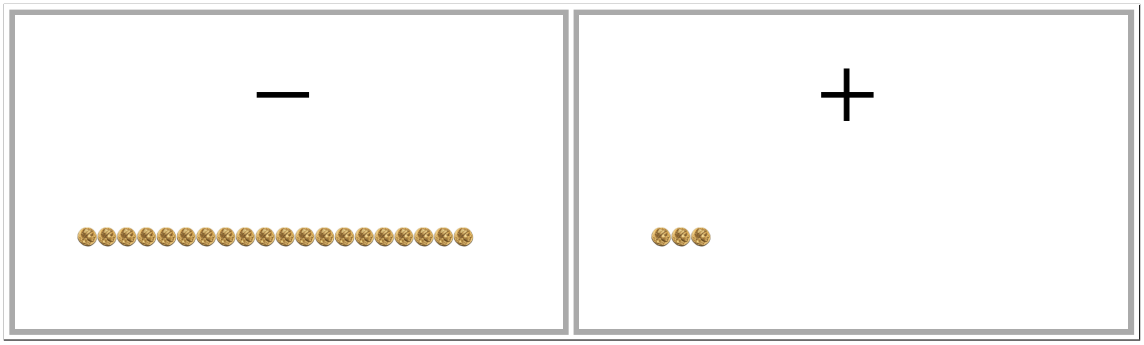


**Figure 1: Visualization of the coins task.** The figure shows a screenshot of the coins task at two different time points: The left image shows the starting position. The sign is "minus" and there are 20 coins available. The right image shows the situation after 17 seconds have passed. The sign is "plus" and there are 3 coins left. Within the coins task, the following non-enforced rule is implemented: "There is a rule for this. The rule is: Press the "Stop"-button after the sign has changed from "minus" to "plus"." The time change occurs after 12 seconds, which is not disclosed to the participants.

In addition to the coins, a sign is displayed on the screen. Initially, the sign shows a minus symbol, which changes to a plus symbol after 12 seconds. At that point, 12 coins have disappeared. Participants are informed that the sign will change after some time but are not told the exact timing. Following Kimbrough and Vostroknutov (2016), Gächter et al. (2025), Suri et al. (2025), and Suri et al. (2026), we introduce an explicit rule linked to this sign change: "There is a rule for this. The rule is: Press the "Stop"-button after the sign has changed from

[4]We changed to press the space bar as in the original setting by Suri et al. (2025) to make the setting less dependent on a functioning keyboard.

'minus' to 'plus'." This implies that following the rule is monetarily costly, but the exact size of these costs (12 coins, $1.20) are ex ante unknown. Importantly, complying with or violating the rule has no monetary consequences beyond those mechanically implied by the number of remaining coins.

Before starting the task, participants are shown a graphical illustration of the coins task (see Figure 1) and are required to answer two comprehension questions correctly in order to proceed. For each of the two situations depicted in the figure, participants are asked to indicate how many coins they would earn if they pressed the "Stop"-button at that point in time. This procedure is designed to ensure understanding of the payoff structure and to highlight—without explicitly stating—that violating the rule does not entail any additional monetary penalty. Each participant completes the coins task exactly once.

## 2.2 Treatment conditions: Varying the rule setter

The experiment adopts a between-subject design in which the origin of the rule is randomly varied across participants. Specifically, participants are assigned to one of two treatment conditions that differ only in who has set the rule in the coins task. In the `AI` condition, the rule is set by the artificial intelligence ChatGPT, whereas in the `HUMAN` condition, the rule is set by a US Prolific user, with the latter mirroring the design of Suri et al. (2026). Participants are explicitly informed of the rule's origin through the following statement in the instructions: "The rule was chosen by the artificial intelligence ChatGPT." in `AI` or "The rule was chosen by a US Prolific user." in `HUMAN`. To ensure that the treatment manipulation is salient and correctly understood, participants must correctly answer an additional comprehension question identifying the origin of the rule before proceeding. Furthermore, we displayed the origin of the rule just before participants started the coins task.

The indication of the rule origin is grounded in actual decision data. For the `HUMAN` condition, we draw on pilot data reported in Suri et al. (2026), in which US Prolific users were asked whether they would implement the rule in the coins task. In that pilot, 28.2% of the 291 participants indicated that they would choose to implement the rule. To construct a comparable basis for the `AI` condition, we presented ChatGPT (model GPT-5 mini, freeware version) with the same task instructions shown to participants and asked whether it would implement the rule. The exact prompt is reported in Section SI-4. This procedure was repeated 100 times, each time in a new chat session, and ChatGPT's responses were recorded as binary choices. Across these iterations, ChatGPT indicated that it would implement the rule in 45 out of 100 cases.

## 2.3 Belief elicitation and conditional rule compliance

In a situation like in our experiment where monetary incentives align with rule-breaking ($I < 0$) and social preferences ($P$) do not matter because rule-breaking does not affect others, the CRISP framework and the experimental evidence supporting it (Gächter et al., 2025) suggest that only an unconditional respect for rules ($R$) and conformity with social expectations ($S$, i.e., descriptive and normative beliefs) can explain rule compliance. Guided by CRISP, we therefore elicit a comprehensive set of belief measures capturing participants' normative and descriptive beliefs regarding rule compliance in the coins task. Our elicitation closely follows

the wording and procedures used in Gächter et al. (2025) and Suri et al. (2026), while extending them along some dimensions.

First, we measure *descriptive beliefs* by asking participants to state their empirical expectations about the share of other participants who followed the rule. Accuracy in this belief elicitation is incentivized using the quadratic scoring rule (Selten, 1998). Second, we elicit *social (second-order) normative beliefs* by asking participants how socially appropriate they believe other US Prolific users would consider rule compliance and, separately, rule violation. This elicitation is incentivized following the coordination-based method proposed by Krupka and Weber (2013).[5] In addition to the CRISP framework, we also elicit *personal (first-order) normative beliefs* regarding the social appropriateness of rule compliance and rule violation from the participant's own perspective because personal beliefs are an important driver of behavior (Bašić & Verrina, 2024). We use the samples' answers to these questions for the incentivization mechanism for the social normative beliefs. Furthermore—and in accordance to Suri et al. (2026)—we also measure measure participants' *general rule-following attitude*—i.e., their views about what one should do in situations like the coins task—using a 4-point Likert scale ranging from "never" (1) to "always" (4) follow the rule.

Moreover, because social expectations can influence rule compliance only if people actually condition their rule compliance on their social expectations, we elicit people's *conditional rule compliance functions*. Specifically, participants are asked whether they would follow the rule conditional on (i) the share of others who disapprove of rule violations and (ii) the share of others who comply with the rule. Following the wording in Gächter et al. (2025), participants report their intended behavior for each of five intervals (0-20%, 21-40%, . . . , 81-100%) of others' disapproval or compliance.[6] Both conditional compliance measures are incentivized using the strategy method (Selten, 1967).

## 2.4 Measuring subjective closeness

In related research (Suri et al., 2026) we have found that people's perceived social closeness to the rule setter is strongly positively correlated with rule compliance. Because AI is a machine and people are aware of that, we expect that the perceived social closeness toward ChatGPT will be lower than to humans. To measure participants' perceived social distance to the rule setter, we employ the "Inclusion of Other in the Self" (IOS11) scale (Baader et al., 2024). The IOS11 is a well-established and validated measure of subjective closeness that has been widely used to capture perceived interpersonal proximity in social contexts (Aron et al., 1992; Gächter et al., 2015; Baader et al., 2024). The scale represents closeness graphically by two circles, one denoting the respondent and the other denoting another agent. Participants indicate perceived closeness by selecting the degree of overlap between the two circles on an 11-point scale, ranging from no overlap (1), corresponding to maximal perceived distance, to near-complete overlap (11), corresponding to very high perceived closeness.

In the experiment, participants adjusted the overlap between the two circles using a slider and were instructed to interpret the resulting configuration as representing their relationship

[5] For a discussion of the reliability of this method to elicit the prevailing norm (with and without incentivization), please refer to Fallucchi and Nosenzo (2022), Aycinena et al. (2024), Charness et al. (2025), and Fallucchi et al. (2026).

[6] Participants were originally asked about the share of others who violate the rule. For comparability with the literature, we transform these measures to reflect the share of others who comply with the rule.

with a specified agent. We varied the identity of the agent such that each participant evaluated their subjective closeness to both (i) ChatGPT and (ii) a fellow US Prolific user, with the order of the two assessments randomized. Illustrations of the task interface are provided in Section SI-3. To construct our measure of subjective closeness toward the rule setter, we use the IOS11 score corresponding to the agent that sets the rule in the coins task—ChatGPT in the `AI` condition and a fellow US Prolific user in the `HUMAN` condition.

## 2.5 Experimental procedures

The experiment was programmed using the oTree software (Chen et al., 2016) and administered via Prolific Academic in December 2025. Participants were recruited from a US-based Prolific pool subject to the following inclusion criteria: at least 18 years of age, native English speaker, US citizen, currently residing in the United States, no prior participation in similar studies by the authors, and provision of informed consent. Using Prolific's built-in filtering tools, we recruited a sample balanced with respect to gender (male/female) and political party affiliation (Democrats/Republicans).

After providing informed consent, participants completed the coins task described in Section 2.1. Subsequently, we elicited belief measures in the following order: general rule-following attitude, normative, and descriptive beliefs, followed by rule compliance conditional on others' disapproval of rule violations and, in a fixed sequence, rule compliance conditional on others' conformity with the rule (see Section 2.3). Participants then completed the IOS11 scale, assessing subjective closeness toward ChatGPT and toward a fellow US Prolific user in random order (see Section 2.4). Next, we administered the ATTARI-12 questionnaire to measure attitudes toward artificial intelligence (Stein et al., 2024).[7] For the ATTARI index, items 2, 4, 7, 8, 10, and 12 were reverse coded and the mean of the 12 substantive items, measured on 5-point Likert scales, was computed.

Following the ATTARI questionnaire, we elicited trust in the rule setter using five items measured on 5-point Likert scales, with the identity of the rule setter varying by treatment condition ("AI" or "other US Prolific users"). The items were: (1) "I generally trust decisions made by [rule setter]." (2) "I usually trust a human decision [*my own decision* in `HUMAN`] more than an [rule setter] decision, even when the [rule setter] is said to be very accurate." (3) "I am willing to rely on [rule setter] when I have to make difficult decisions." (4) "I am skeptical of decisions made by [rule setter]." (5) "I feel at ease when [rule setter] are used to support decisions in everyday life." We constructed a trust index by averaging the items after reverse coding items (2) and (4). Participants then completed a second attention check and a short socio-demographic questionnaire, which included a measure of patience based on the qualitative 11-point Likert scale elicitation method from Falk et al. (2018).

The median completion time of the study was 16 minutes and 23 seconds. Participants received a fixed participation payment of $2.29 and a bonus payment of up to $2.00 based on their decisions in the experiment. To determine the bonus payment, one of the incentivized tasks was randomly selected for each participant: (i) the coins task, (ii) the normative and descriptive belief elicitation, (iii) rule compliance conditional on others' disapproval of rule violations, or (iv) rule compliance conditional on others' conformity with the rule. In each

[7] We added a thirteenth item serving as an attention check, instructing participants to select the response "strongly agree."

case, the maximum attainable bonus was $2.00. The mean total payment per participant was $3.36.

## 2.6 Our participant pool

We pre-registered a sample size of 500 participants for each treatment condition. After applying our pre-registered exclusion criteria[8], the final sample consists of n=487 participants in treatment `AI` and n=490 participants in treatment `HUMAN`. Table 1 reports mean participant characteristics by treatment condition, along with p-values from Pearson's chi-squared tests and Wilcoxon rank sum tests, as well as statistics for the full sample. We find no statistically significant differences in any of the observed characteristics between treatment conditions, suggesting successful randomization.[9]

**Table 1:** Overview of participants' characteristic

| Characteristic | `AI` | `HUMAN` | Full sample | p-value |
|---|---|---|---|---|
| Sample size | 487 | 490 | 977 | |
| Male | 49.90% | 49.18% | 49.54% | 0.874[a] |
| Democrat | 49.49% | 49.18% | 49.33% | 0.976[a] |
| Age | 42.28 | 42.34 | 42.31 | 0.915[b] |
| Northeast | 17.04% | 16.73% | 16.89% | 0.966[a] |
| Midwest | 18.07% | 17.14% | 17.60% | 0.767[a] |
| South | 44.97% | 46.53% | 45.75% | 0.670[a] |
| West | 19.92% | 19.59% | 19.75% | 0.962[a] |
| Patience | 6.73 | 6.76 | 6.74 | 0.804[b] |
| IOS11 to humans | 5.67 | 5.73 | 5.70 | 0.865[b] |
| IOS11 to ChatGPT | 3.98 | 3.75 | 3.86 | 0.274[b] |
| ATTARI | 3.38 | 3.38 | 3.38 | 0.918[b] |
| Trust in rule setter | 2.87 | 2.83 | 2.85 | 0.384[b] |

*Notes:* The table reports means for individual characteristics for the two treatment conditions and the full sample. The p-values are obtained from Pearson's chi-squared tests (a) or Wilcoxon rank sum tests (b).

Participants are, on average, approximately 42 years old and are drawn from all four major US regions—Northeast, Midwest, South, and West. On average, they self-report being slightly patient (6.74 on an 11-point Likert scale with 0 = very impatient and 10 = very patient). Attitudes toward artificial intelligence are mildly positive, as measured by the ATTARI questionnaire (mean = 3.38, SD = 1.10 in both treatment conditions), which is comparable to prior findings in the literature (e.g., Stein et al., 2024). The internal consistency of the 12-item ATTARI scale is excellent (Cronbach's $\alpha$ = .96 in both treatment conditions).

In addition, we collected two measures capturing broader aspects of participants' relationships with the rule setter. First, we elicited *trust* in the rule setter using five survey items. Across both treatment conditions, participants report slightly below-neutral trust, with mean

[8] Among others, we included two attention checks. First, participants were instructed to select "strongly agree" for a specific item in the ATTARI questionnaire. Second, at the end of the study, participants were asked to recall which sign changed in the coins task from a set of four possible changes.

[9] Please note that gender and political party affiliation are balanced by design, as participants were invited accordingly.

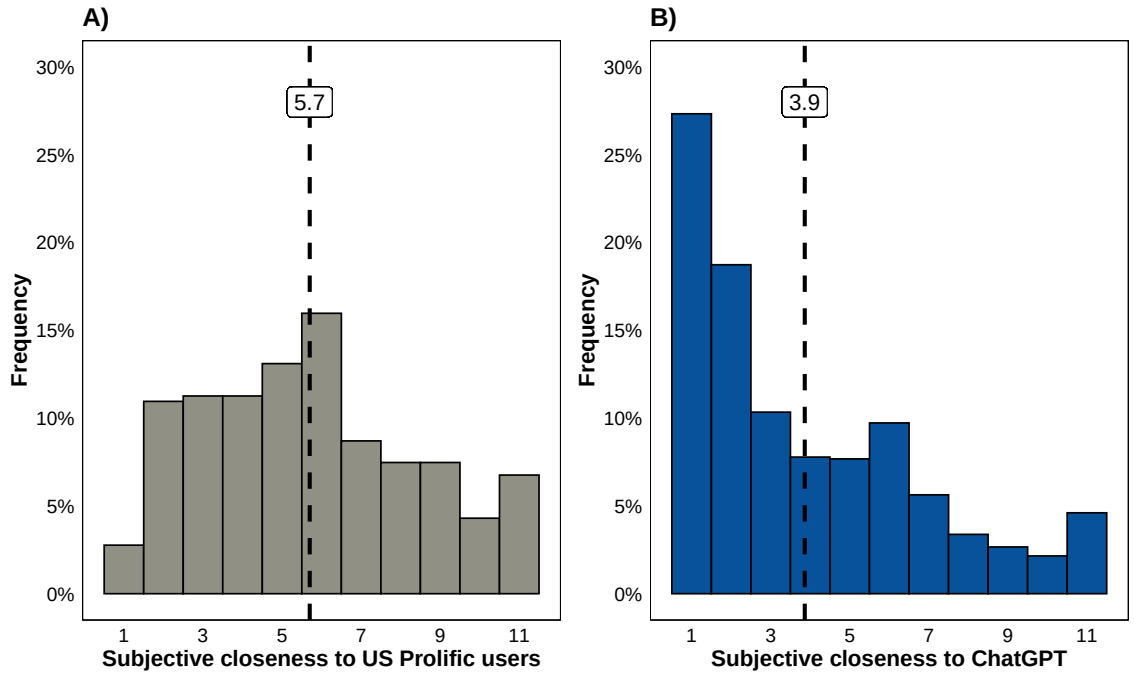



**Figure 2: Subjective closeness to US Prolific users and ChatGPT.** The figure shows the distribution of subjective closeness toward US Prolific users (left, A) and ChatGPT (right, B) using data from the full sample. Subjective closeness is measured by the IOS11 score (Baader et al., 2024), where 1 means subjectively very distant and 11 means subjectively very close. The dashed lines represent the respective means.

values just below 3, the midpoint of the 5-point Likert scale (Cronbach's $\alpha$ = .89 in `AI` and = .78 in `HUMAN`). Second, because previous research (Suri et al., 2026) suggests that subjective closeness is correlated with rule compliance, we measured *subjective closeness* toward (i) a fellow US Prolific user and (ii) ChatGPT. The distributions of these closeness measures are shown in Figure 2. For fellow Prolific users (panel A), subjective closeness is approximately normally distributed, with a mean of 5.7 in the full sample, indicating that relatively few participants feel either very distant (e.g., a score of 1) or very close (e.g., a score of 11). Most participants report intermediate levels of closeness. This pattern closely resembles the distribution reported in Suri et al. (2026), which also examines psychological distance toward US Prolific users. In contrast, subjective closeness toward ChatGPT (panel B) differs markedly. The distribution is significantly different from that toward fellow Prolific users (Kolmogorov–Smirnov and Wilcoxon signed rank tests, both p<.001, for each treatment condition separately and for the full sample) and is strongly right-skewed, indicating that a large share of participants report feeling subjectively distant from the artificial intelligence system.

Lastly, we also asked participants to indicate their agreement with the following statement on a 5-point Likert scale (1="strongly disagree", 5="strongly agree"): "I experienced the rule as coming from a real person rather than just from a computer system." In the `HUMAN` condition—where the rule is actually set by a human—the average response of 3.43 corresponds toward an agreement with the statement (median=4), while if the rule is set by ChatGPT (`AI`) participants on average rather disagree with the statement (mean=2.45, median=2). This significant difference (Wilcoxon rank sum test, p<.001) not only confirms—alongside the control questions—that our manipulation of the rule setter's origin is salient, but, together with the pronounced difference in subjective closeness to a human versus an AI rule setter, also points

to a potential mechanism through which the rule's origin might influence compliance. We therefore proceed to examine the treatment effects of AI versus HUMAN in the next section.

# 3 Results

This section presents the results of our study in two steps. In Section 3.1, we begin with the analysis of the effects of the experimental treatment conditions on rule-following behavior. Guided by CRISP, we examine overall rule compliance in the coins task, the role of normative and descriptive beliefs as well as general rule-following attitudes, and the extent to which participants condition their behavior on social expectations. Subsequently, in Section 3.2 we investigate how subjective closeness to the rule setter is related to rule compliance.

For data analysis we used the R software, version 4.2.2 (R Core Team, 2022). We report additional figures and tables in the SI for explanatory and exploratory purposes. Unless stated otherwise, all p-values reported are based on two-sided hypothesis tests.

## 3.1 Rule compliance with human and AI rule setters

As laid out in the introduction, prior work shows that people often engage with and sometimes trust AI agents, but AI involvement can reduce fairness, cooperation, and coordination success relative to human–human interaction. Therefore, our first pre-registered hypothesis posits that rule compliance in the coins task is higher when the rule is set by a US Prolific user than when it is set by ChatGPT. Recall that, in this task, participants can freely violate the rule to earn a higher payoff without facing any negative consequences, while rule compliance itself is not rewarded. Participants were explicitly informed of these features and demonstrated understanding by correctly answering comprehension questions.

Figure 3A displays rule compliance rates by treatment condition. In the HUMAN condition, 63.88% of participants comply with the rule. This relatively high rate of rule adherence is comparable to rule-following rates reported in the literature when rules are set by the experimenter—and thus implicitly by a human agent—(see, e.g., Kimbrough & Vostroknutov, 2016, 2018; Kimbrough et al., 2024; Gächter et al., 2025; Suri et al., 2025), as well as to the compliance rate reported by Suri et al. (2026) for a setting in which the rule is set by a fellow US Prolific user (55%, n=1,666).

In the AI condition, rule compliance is 0.39 percentage points higher at 64.27%. The difference is not statistically significant relative to the HUMAN condition (Pearson's chi-squared test, p=.951). This null result comes at a surprise given the substantial differences in subjective closeness toward ChatGPT and toward fellow Prolific users documented in the previous section (see Figure 2). Consistent with this descriptive observation, the treatment indicator is insignificant in the corresponding linear probability model that includes additional control variables (see Table 2, Column 1).

Following Suri et al. (2026), we additionally conduct equivalence testing to assess whether the statistically insignificant difference in rule compliance is also economically negligible. We rely on the regression estimates of the AI treatment indicator reported in Table 2, Columns 1 and 2, together with a baseline specification in which rule compliance is regressed solely on the treatment indicator. The corresponding point estimates and 90% confidence intervals are shown in Figure SI-1.

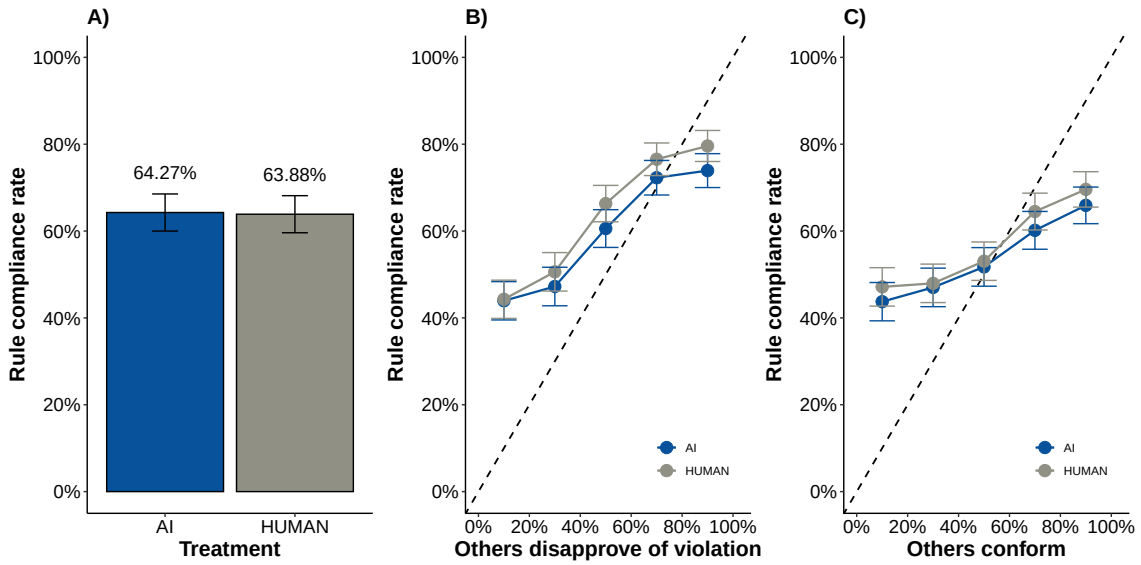



**Figure 3: Rule compliance across treatment conditions.** The figure shows incentivized rule compliance in the one-shot coins task (left, A) and conditional on the percentage (presented as quintiles) of others who disapprove of rule violation (middle, B) as well as conform with the rule (right, C); elicited with the strategy method. The dashed lines are the diagonals. The whiskers represent the 95% confidence intervals.

For all three models, the confidence intervals lie entirely within the region of practical equivalence. This region is defined by the upper bound of the confidence interval of the "experimenter-induced experimenter demand effect" estimated by Suri et al. (2026), corresponding to a 6.7-percentage-point difference in rule compliance. The two one-sided tests therefore indicate statistical equivalence in compliance between rules set by ChatGPT and those set by Prolific users, implying that the statistically insignificant difference across both rule-setter identities is also behaviorally negligible.

> ***Result 1:*** *Compliance with an arbitrary rule in the coins task is high and comparable to levels reported in the literature, and it does not differ by the origin of the rule, whether set by ChatGPT or by a human agent who is a fellow US Prolific user.*

To further examine the determinants of rule compliance, we next analyze the role of beliefs. Table 2, Column 2 shows that several belief and attitude measures significantly predict compliance behavior. In particular, participants' general rule-following attitude about what one should do in situations like the coins task are positively associated with compliance ($\beta$=.12, p<.001). Compliance is also positively related to participants' descriptive beliefs, measured as their incentivized estimate of the share of others who followed the rule ($\beta$=.01, p<.001). By contrast, personal normative beliefs that rule violations are socially appropriate are negatively associated with compliance ($\beta$=-.07, p<.05). When each belief category is added separately to the regression models (see Table SI-1 for `AI` and Table SI-2 for `HUMAN`), social normative beliefs—beliefs about others' views regarding the social appropriateness of rule violations—also become statistically significant. Overall, these findings align closely with our additional hypotheses pre-registered on the basis of the CRISP framework (Gächter et al., 2025) and echo the results reported by Suri et al. (2026).

**Table 2:** Impact of rule setter on rule compliance

| | Dependent variable: Rule compliance | | | |
|---|---|---|---|---|
| | Unconditional | | Conditional | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.64*** | 0.17* | 0.13* | 0.14* |
| | (0.08) | (0.09) | (0.06) | (0.07) |
| AI | 0.01 | −0.01 | −0.03 | −0.03 |
| | (0.03) | (0.03) | (0.02) | (0.02) |
| % others disapprove of violation | | | 0.005*** | |
| | | | (0.0003) | |
| % others conform | | | | 0.003*** |
| | | | | (0.0003) |
| General rule-following attitude | | 0.12*** | | |
| | | (0.02) | | |
| Descriptive belief | | 0.01*** | | |
| | | (0.001) | | |
| Personal normative belief (compliance) | | −0.03 | | |
| | | (0.03) | | |
| Personal normative belief (violation) | | −0.07* | | |
| | | (0.03) | | |
| Social normative belief (compliance) | | −0.001 | | |
| | | (0.04) | | |
| Social normative belief (violation) | | 0.05 | | |
| | | (0.03) | | |
| Controls | Yes | Yes | Yes | Yes |
| Number of participants | 962 | 962 | 962 | 962 |
| Observations | 962 | 962 | 4,810 | 4,810 |
| Adjusted $R^2$ | −0.01 | 0.22 | 0.09 | 0.04 |
| F Statistic | 0.30 | 19.48*** | 49.46*** | 23.44*** |
| Degrees of freedom | (9; 952) | (15; 946) | (10; 4799) | (10; 4799) |

*Notes:* The table reports coefficient estimates from linear probability models. The dependent variable is rule compliance. Columns 1-2 display data from the one-shot coins task. Column 3 (4) displays data from the rule compliance elicitation conditional on the share of others who disapprove of violation (who conform with the rule) using the strategy method. The complete regression table is displayed in Table SI-3. Table SI-4 provides the regression estimates of the corresponding logit models which yield qualitatively similar results for reported coefficients. Standard errors are reported in parentheses. Models 1–2 use heteroskedasticity-robust standard errors. Models 3–4 use standard errors clustered at the subject level. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

***Result 2:*** *Ceteris paribus, rule compliance in the coins task is more likely among participants who (i) believe that others should follow rules in similar decision-making situations (general rule-following attitude) and (ii) expect a higher share of others to comply with the rule (descriptive beliefs), and less likely among participants who (iii) consider rule violations socially appropriate (personal normative beliefs) and (iv) believe that others consider rule violations socially appropriate (social normative beliefs).*

Given that beliefs exhibit similar predictive power across treatment conditions, we next examine whether their levels differ by the origin of the rule. In our pre-registration, we hypothesized that belief measures would be higher in the HUMAN condition than in the AI condition. Figure 4 summarizes all elicited beliefs and attitudes by treatment.

Starting with the *general rule-following attitude* about what one should do in situations like the coins task (Figure 4A), the median response in both conditions is that the rule should always be followed. Neither the medians nor the full distributions differ significantly between treatments (Wilcoxon rank sum test: p=.559, Kolmogorov-Smirnov test: p = .981). Turning to *descriptive beliefs* regarding others' compliance (Figure 4B), participants in both conditions slightly overestimate actual compliance rates. Although the mean descriptive belief is

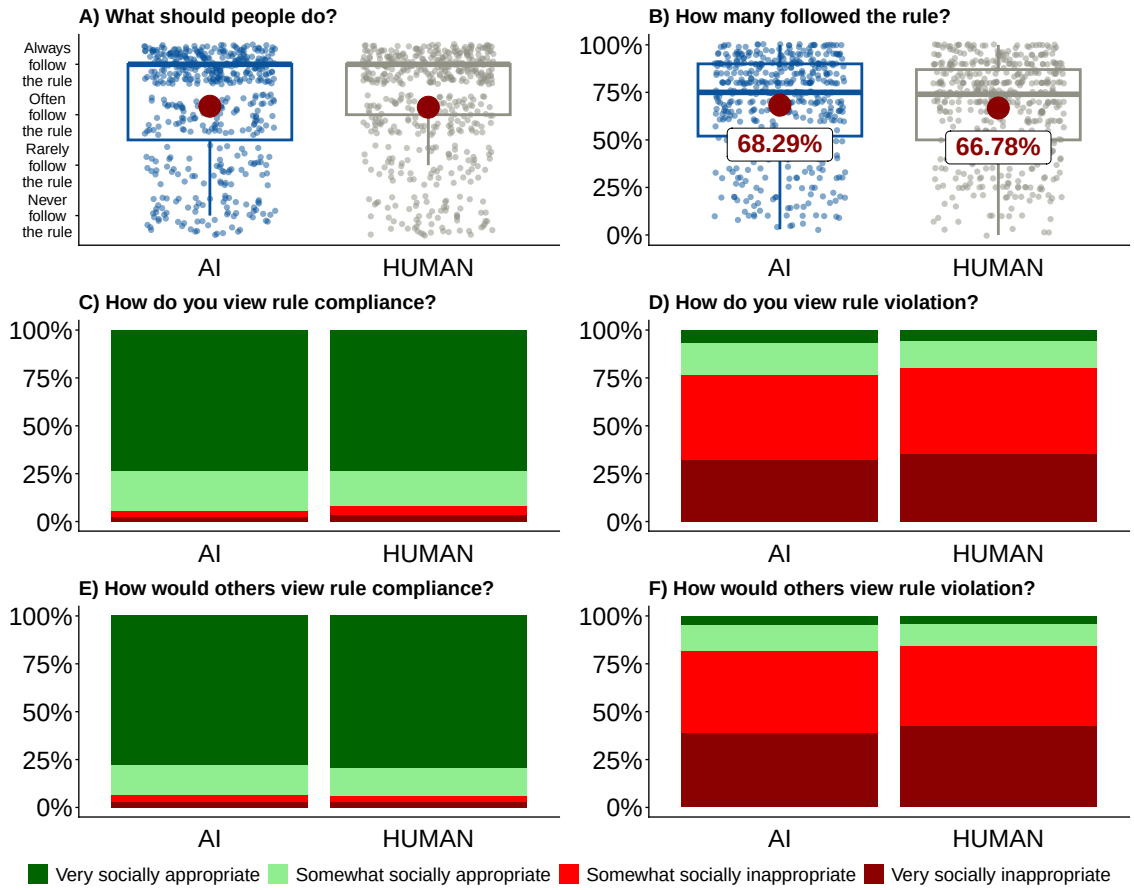



**Figure 4: Beliefs across treatment conditions.** The figure shows the general rule-following attitude (top left, A) and incentivized descriptive (top right, B) and personal as well as social normative beliefs of rule compliance (middle left, C and bottom left, E) and rule violation (middle right, D and bottom right, F); elicited with the quadratic scoring rule (B) or the strategy method (C to F). The solid lines represent the median and the red dots the mean values (A and B).

approximately 1.5 percentage points higher in the AI condition, neither the mean difference nor the distributions differ significantly across treatments (Wilcoxon rank sum test: p=.276, Kolmogorov-Smirnov test: p = .541).

A similar pattern emerges for *normative beliefs* about the social appropriateness of rule compliance and rule violations (Figures 4C–F). Neither personal nor social normative beliefs differ significantly between treatment conditions (Wilcoxon rank sum tests: all p>.191, Kolmogorov-Smirnov tests: all p>.902). Within each treatment condition, however, normative beliefs about rule compliance and rule violation differ markedly (Wilcoxon rank sum tests: all p<.001, Kolmogorov-Smirnov tests: all p<.001).

> ***Result 3:*** *Neither descriptive nor normative beliefs differ by the origin of the rule in the coins task, whether the rule is set by ChatGPT or by a US Prolific user.*

The CRISP framework (Gächter et al., 2025) further predicts that rule compliance depends not only on social expectations *per se* but also on people's *conditional rule compliance function* because social expectations can only influence compliance if people condition their compliance on them. We therefore analyze compliance behavior conditional on (i) the share of others who disapprove of rule violations and (ii) the share of others who comply with the rule. In both cases, participants indicated whether they would follow or violate the rule for each quintile of others' behavior. These elicitations were incentivized using the strategy method.

We begin with conditioning on others' disapproval of rule violations (*conditional conformity with normative beliefs*). Figure SI-2A shows individual compliance profiles. In the AI condition, 34.91% of participants always comply with the rule regardless of others' behavior, compared to 37.35% in the HUMAN condition, classifying them as unconditional rule-followers. Conversely, 17.66% of participants in AI and 15.10% in HUMAN never comply, rendering them unconditional rule-violators. A further 35.73% in AI and 39.19% in HUMAN condition their compliance monotonically on others' disapproval, while the remaining participants exhibit non-monotonic or decreasing compliance patterns.

Average conditional compliance rates are displayed in Figure 3B. When only a small share of others (0-20%) disapprove of rule violations, compliance averages 43.94% in AI and 44.29% in HUMAN. Compliance increases monotonically with the share of others who disapprove, reaching 73.92% in AI and 79.59% in HUMAN when nearly all others (81-100%) disapprove. This pattern is corroborated by a linear probability model, which yields a positive and highly significant slope coefficient ($\beta$=.005, p<.001; see Table 2, Column 3). Pairwise Fisher's exact tests with Holm correction reveal, however, no statistically significant differences between treatment conditions at any quintile (all p>.2).[10]

We next consider compliance conditional on the share of others who conform with the rule (*conditional conformity with descriptive beliefs*). Figure SI-2B displays the corresponding individual profiles. In the AI condition, 30.60% of participants are unconditional rule-followers, compared to 33.67% in the HUMAN condition. Unconditional rule-violators account for 21.77% and 19.80% of participants in AI and HUMAN, respectively. Conditional compliers—who condition their behavior on others' rule conformity—constitute 31.21% in AI and 31.43% in HUMAN, while the remainder again display non-monotonic or decreasing compliance patterns.

[10]The median switching point among participants with monotonically increasing profiles—the third quintile (41-60% disapproval)—also does not differ significantly between conditions (Kolmogorov–Smirnov test, p=.453).

Figure 3C shows average compliance rates by quintile of others' compliance. When only 0-20% of others comply with the rule, compliance averages 43.74% in AI and 47.14% in HUMAN. Compliance rises with the share of others who comply, reaching 65.91% in AI and 69.59% in HUMAN when 81-100% comply. A linear probability model again confirms a positive and highly significant relationship between others' compliance and own compliance ($\beta$=.003, p<.001; see Table 2, Column 4). Pairwise Fisher's exact tests with Holm correction indicate again no statistically significant differences between treatment conditions across quintiles (all p>.8).[11]

> ***Result 4:*** *Participants condition their rule compliance on both normative beliefs about others' disapproval of rule violations and descriptive beliefs about others' compliance; however, conditional compliance does not differ by whether the rule originates from ChatGPT or from a fellow US Prolific user.*

Taken together, rule compliance and the associated belief measures appear largely unaffected by the origin of the rule, echoing the findings of Suri et al. (2026), who document substantial consistency in rule-following across rule setters associated with different political party affiliations. However, as emphasized by Suri et al. (2026), social closeness toward the rule setter may nevertheless play a role in shaping compliance behavior. Although we have already established significant differences in subjective closeness toward ChatGPT versus fellow Prolific users, we explore this potential channel in greater detail in the following section.

## 3.2 Rule compliance and social closeness

Based on our findings in related research (Suri et al., 2026), we further pre-registered the hypothesis that rule compliance depends on the perceived social distance to the originator of the rule, with higher compliance when the rule setter is perceived as subjectively closer to the participant. To test this hypothesis, we first include the IOS11 measure of subjective closeness toward the rule setter in our regression models. Recall that the IOS11 task asks participants to adjust sliders such that the degree of overlap reflects their perceived psychological closeness to (i) ChatGPT and (ii) a fellow US Prolific user. Table 3 reports the corresponding regression results.

Focusing first on rule compliance in the one-shot coins task (unconditional rule compliance), subjective closeness does not exhibit a statistically significant association with compliance behavior, neither in the full sample (Table 3, Columns 1 and 2) nor when estimating the models separately for the AI and HUMAN conditions (Table SI-5, Columns 1–4). In contrast, subjective closeness significantly predicts conditional rule compliance (Table 3, Columns 3 and 4). Specifically, a one-point increase on the 11-point IOS11 scale is, ceteris paribus, associated with an increase of approximately two percentage points in average conditional compliance. This relationship holds for compliance conditional on others' disapproval of rule violations as well as for compliance conditional on others' conformity with the rule. The effect is robust across treatment conditions and is present both in pooled regressions and in treatment-specific subsample estimations.

To complement this regression-based analysis, we additionally implement the pre-registered median split of subjective closeness scores within each treatment condition, as illus-

[11] The median switching point for monotonically increasing profiles—the third quintile (41-60% compliance)—also does not differ across conditions (Kolmogorov-Smirnov test p=.996).

**Table 3:** Impact of subjective closeness toward rule setter on rule compliance

| | Dependent variable: Rule compliance | | | |
|---|---|---|---|---|
| | Unconditional | | Conditional | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.63*** | 0.18* | 0.10 | 0.12 |
| | (0.08) | (0.09) | (0.06) | (0.06) |
| IOS11 toward rule setter | 0.01 | −0.01 | 0.02*** | 0.02*** |
| | (0.01) | (0.005) | (0.004) | (0.005) |
| AI | 0.02 | −0.02 | 0.01 | 0.005 |
| | (0.03) | (0.03) | (0.02) | (0.03) |
| % others disapprove of violation | | | 0.005*** | |
| | | | (0.0003) | |
| % others conform | | | | 0.003*** |
| | | | | (0.0003) |
| General rule-following attitude | | 0.13*** | | |
| | | (0.02) | | |
| Descriptive belief | | 0.01*** | | |
| | | (0.001) | | |
| Personal normative belief (compliance) | | −0.04 | | |
| | | (0.03) | | |
| Personal normative belief (violation) | | −0.07* | | |
| | | (0.03) | | |
| Social normative belief (compliance) | | −0.002 | | |
| | | (0.04) | | |
| Social normative belief (violation) | | 0.05 | | |
| | | (0.03) | | |
| Controls | Yes | Yes | Yes | Yes |
| Number of participants | 962 | 962 | 962 | 962 |
| Observations | 962 | 962 | 4,810 | 4,810 |
| Adjusted $R^2$ | −0.01 | 0.22 | 0.11 | 0.05 |
| F Statistic | 0.52 | 18.35*** | 52.94*** | 25.78*** |
| Degrees of freedom | (10; 951) | (16; 945) | (11; 4798) | (11; 4798) |

*Notes:* The table reports coefficient estimates from linear probability models. The dependent variable is rule compliance. Columns 1-2 display data from the one-shot coins task. Column 3 (4) displays data from the rule compliance elicitation conditional on the share of others who disapprove of violation (who conform with the rule) using the strategy method. The complete regression table is displayed in Tables SI-5 and SI-6. Standard errors are reported in parentheses. Models 1–2 use heteroskedasticity-robust standard errors. Models 3–4 use standard errors clustered at the subject level. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

trated in Figure 2. In the AI condition, ties at the median (n=48) are randomly assigned to either the CLOSE or DISTANT subgroup. No ties exist in HUMAN. Figures SI-3 and SI-4 display unconditional and conditional rule compliance, as well as the corresponding belief measures, separately for each treatment and closeness subgroup. Table 4 summarizes mean values for all treatment × closeness combinations and reports p-values from pairwise Pearson's chi-squared tests and Wilcoxon rank sum tests comparing CLOSE and DISTANT participants within each treatment condition.

Consistent with the regression results, unconditional rule compliance in the one-shot coins task is descriptively higher among participants who perceive the rule setter as subjectively close—by approximately five percentage points in AI and 3.7 percentage points in HUMAN. However, though these differences echo the findings of Suri et al. (2026), they are not statistically significant. By contrast, conditional rule compliance differs sharply by subjective closeness. Across all quintiles of others' behavior, and after correcting for multiple hypothesis testing, participants in the CLOSE subgroups exhibit significantly higher conditional compliance rates than those in the DISTANT subgroups, both when conditioning on others' disapproval of rule violations and on others' conformity with the rule, and both if the rule is set by ChatGPT or a fellow Prolific user.

Belief and attitude measures mirror this pattern. General rule-following attitudes and descriptive beliefs are significantly higher in the CLOSE subgroups within both treatment conditions. Normative beliefs regarding the social appropriateness of rule compliance vary little between CLOSE and DISTANT participants. Normative beliefs concerning the social appropriateness of rule violations, however, differ significantly between the two groups, with participants in CLOSE expressing stronger disapproval of violations.

> ***Result 5:*** *Perceived subjective closeness to the rule setter substantially increases conditional rule compliance. Participants who feel closer to the rule originator condition their behavior more strongly on both normative beliefs about others' disapproval of rule violations and descriptive beliefs about others' rule conformity. This pattern holds irrespective of whether the rule is set by ChatGPT or by a fellow US Prolific user.*

**Table 4:** Mean rule compliance and beliefs given subjective closeness to the rule setter

| | AI Close n=244 | AI Distant n=243 | HUMAN Close n=245 | HUMAN Distant n= 245 | p-values |
|---|---|---|---|---|---|
| Rule compliance | 66.80% | 61.73% | 65.71% | 62.04% | .283 \| .452 |
| Conditional rule compliance (0-20% disapprove) | 50.82% | 37.04% | 50.20% | 38.37% | .003 \| .022 |
| Conditional rule compliance (21-40% disapprove) | 54.92% | 39.51% | 55.92% | 45.31% | . 002 \| .024 |
| Conditional rule compliance (41-60% disapprove) | 69.67% | 51.44% | 73.47% | 59.18% | <.001 \| .003 |
| Conditional rule compliance (61-80% disapprove) | 82.38% | 62.14% | 83.27% | 69.80% | <.001 \| .002 |
| Conditional rule compliance (81-100% disapprove) | 81.56% | 66.26% | 86.12% | 73.06% | <.001 \| .002 |
| Conditional rule compliance (0-20% conform) | 51.64% | 35.80% | 53.47% | 40.82% | .002 \| .026 |
| Conditional rule compliance (21-40% conform) | 56.97% | 37.04% | 54.29% | 41.63% | <.001 \| .026 |
| Conditional rule compliance (41-60% conform) | 59.84% | 43.62% | 58.37% | 47.76% | .002 \| .047 |
| Conditional rule compliance (61-80% conform) | 64.75% | 55.56% | 71.02% | 57.96% | .044 \| .017 |
| Conditional rule compliance (81-100% conform) | 70.90% | 60.91% | 73.88% | 65.31% | .044 \| .049 |
| General rule-following attitude | 3.45 | 2.89 | 3.32 | 2.98 | <.001 \| <.001 |
| Descriptive belief | 73.53% | 63.04% | 70.80% | 62.77% | <.001 \| <.001 |
| Personal normative belief (compliance) | 0.79 | 0.76 | 0.75 | 0.75 | .039 \| .270 |
| Personal normative belief (violation) | -0.45 | -0.24 | -0.46 | -0.33 | <.001 \| .005 |
| Social normative belief (compliance) | 0.78 | 0.81 | 0.82 | 0.79 | .773 \| .186 |
| Social normative belief (violation) | -0.53 | -0.34 | -0.52 | -0.44 | <.001 \| .036 |

*Notes:* The table reports the means of rule compliance and normative beliefs. The corresponding distributions are shown in Figures SI-2 and SI-3. The Close/Distant categorization is based on median splits within each treatment condition. The first p-value represents the comparison between AI Close and AI Distant, while the second p-value represents the comparison between HUMAN Close and HUMAN Distant. P-values are calculated using pairwise Pearson's chi-squared tests for all comparisons, except for general rule-following attitude and normative beliefs, which are assessed using Wilcoxon rank sum tests. P-values for the conditional rule compliance comparisons are adjusted using the Holm correction method. The categories for the general rule-following attitude are: never follow the rule (1), rarely follow the rule (2), often follow the rule (3), and always follow the rule (4). The categories for the normative beliefs are: very socially inappropriate (-1), somewhat socially inappropriate (-1/3), somewhat socially appropriate (1/3), and very socially appropriate (1).

## 4 Discussion and concluding remarks

In this study, we examined rule compliance under two distinct rule-setting regimes: rules originating from an artificial intelligence agent (ChatGPT) or from a fellow US Prolific user. Across all outcome measures, we find that revealing the origin of the rule does not significantly affect behavior. Neither rule compliance nor beliefs differ between the two treatment conditions.

This result aligns closely with the findings of Suri et al. (2026), who document remarkably stable rule-following behavior irrespective of whether rules are attributed to the experimenter, a co-partisan, a political opponent, or an anonymous stranger. Taken together, these findings suggest that—at least in abstract and non-consequential environments—rule compliance appears largely invariant to the identity of the rule setter. This invariance may also explain why compliance rates in studies that do not disclose the rule-setter's identity closely resemble those in settings where the identity is explicitly revealed (see also Kimbrough & Vostroknutov, 2016, 2018; Kimbrough et al., 2024; Gächter et al., 2025; Suri et al., 2025).

We further explored two factors commonly emphasized in the literature on human–AI interaction: attitudes toward AI and trust in AI. Neither factor explains compliance behavior in our data. Attitudes toward AI, measured using the ATTARI questionnaire, do not predict rule compliance when the rule is set by ChatGPT, nor do they correlate significantly with compliance at the individual level. Similarly, trust in the rule setter is virtually identical across the `AI` and `HUMAN` conditions, mirroring experimental evidence showing comparable trust levels in human–AI and human–human interactions (e.g., Jayasekara et al., 2025). As a result, limited variation in trust likely constrains its explanatory power in our setting.

In contrast, we observe meaningful heterogeneity in participants' perceived subjective closeness to the rule setter. While participants report relatively low closeness toward ChatGPT and intermediate levels of closeness toward other Prolific users, this variation proves behaviorally relevant. Greater subjective closeness to the rule setter is associated with higher rule compliance, independent of whether the rule setter is human or artificial. This finding corroborates earlier evidence on the role of social proximity in rule-following behavior (Suri et al., 2026) and suggests that psychological distance, rather than the human or non-human nature of the rule setter, is a key behavioral mechanism worth further investigation.

Consistent with the CRISP framework (Gächter et al., 2025), we further show that rule compliance is closely linked to social expectations $S$ and an unconditional respect for rules $R$. Participants condition their behavior on both normative beliefs (others' disapproval of rule violations) and descriptive beliefs (others' conformity with the rule) and around a third of people follow the rule unconditionally, which is consistent with the importance of the unconditional respect $R$ in rule compliance. Importantly, these patterns hold irrespective of whether the rule is set by a human or by an AI agent, indicating that the expectation-based foundations of rule compliance and the unconditional respect for rules extend to human–AI interaction contexts. This supports the conclusion from Gächter et al. (2025, p. 1342) that "respect for rules and social expectations are basic elements of rule-conformity [...] even without extrinsic incentives and social preferences".

Our findings point to several promising avenues for future research. First, while rule compliance in our study appears insensitive to the identity of the rule setter, this may change in environments with material externalities. Introducing settings in which compliance benefits—

or harms—the rule setter directly could reveal whether distributional concerns or perceived legitimacy differentially affect compliance with human versus AI-generated rules. Second, future work could increase the moral or ethical salience of the rule. In our design, the rule governs an arbitrary task with no broader consequences. Rules embedded in ethical dilemmas or socially consequential domains—such as fairness, discrimination, or safety—may trigger stronger reactions to the nature of the rule setter, particularly when accountability and responsibility become salient. Third, an important extension would be to endogenize the rule-setting process. Allowing participants to vote on, delegate, or reject a rule setter—human or AI—could shed light on the conditions under which artificial agents are granted authority and how such authority shapes subsequent compliance. Finally, future studies could explore dynamic interactions with AI rule setters. Repeated exposure, feedback, or the possibility of contesting AI decisions may gradually alter perceptions of legitimacy, trust, and social proximity, thereby affecting rule-following behavior over time.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Declaration of AI and AI-assisted technologies in the writing process**

During the preparation of this work, we used ChatGPT in order to improve readability and language. After using this service, we reviewed and edited the content as needed and take full responsibility for the content of the publication.

**Authorship contribution statement**

**Dominik Suri:** Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Simon Gächter:** Conceptualization, Writing – review & editing. **Sebastian Kube:** Conceptualization, Writing – review & editing.

## References


Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, *9*, 1380–1390. https://doi.org/10.1038/s41562-025-02172-y.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the Structure of Interpersonal Closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596.

Aycinena, D., Bogliacino, F., & Kimbrough, E. O. (2024). *Measuring Norms: Eliciting normative expectations with Krupka and Weber's method allowing for neutral reports* (SSRN Working Paper No. 5050953). https://doi.org/10.2139/ssrn.5050953.

Baader, M., Starmer, C., Tufano, F., & Gächter, S. (2024). Introducing IOS11 as an extended interactive version of the 'Inclusion of Other in the Self' scale to estimate relationship closeness. *Scientific Reports*, *14*(1), 8901. https://doi.org/10.1038/s41598-024-58042-6.

Bašić, Z., & Verrina, E. (2024). Personal norms — and not only social norms — shape economic behavior. *Journal of Public Economics*. https://doi.org/10.1016/j.jpubeco.2024.105255.

Beraja, M., Kao, A., Yang, D. Y., & Yuchtman, N. (2023). Ai-Tocracy. *The Quarterly Journal of Economics*, *138*(3), 1349–1402. https://doi.org/10.1093/qje/qjad012.

Bergougui, B. (2025). Can Artificial Intelligence Technologies Advance Environmental Sustainability? The Role of Institutional Adaptability and Skill-Biased Technological Transformation. *Sustainable Development*, 1–23. https://doi.org/10.1002/sd.70296.

Bharadiya, J. P., Thomas, R. K., & Ahmed, F. (2023). Rise of Artificial Intelligence in Business and Industry. *Journal of Engineering Research and Reports*, *25*(3), 85–103. https://doi.org/10.9734/JERR/2023/v25i3893.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., Gächter, S., Molleman, L., & Nosenzo, D. (2025). Group identity and peer effects in rule-following. *Journal of Economic Behavior & Organization*, *239*, 107264. https://doi.org/10.1016/j.jebo.2025.107264.

Biesheuvel, L. A., Dongelmans, D. A., & Elbers, P. W. (2024). Artificial intelligence to advance acute and intensive care medicine. *Current Opinion in Critical Care*, *30*(3), 246–250. https://doi.org/10.1097/MCC.0000000000001150.

Brennan, G., & Buchanan, J. M. (1985). *The reason of rules: Constitutional political economy*. Cambridge University Press.

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, *64*(1), 1. https://doi.org/10.1037/a0010932.

Capponi, S., & Daniels, K. G. (2023). Harnessing the power of artificial intelligence to advance cell therapy. *Immunological Reviews*, *320*(1), 147–165. https://doi.org/10.1111/imr.13236.

Charness, G., Dimant, E., Gneezy, U., & Krupka, E. (2025). Experimental methods: Eliciting and measuring social norms. *Journal of Economic Behavior & Organization*, *237*, 107187. https://doi.org/10.1016/j.jebo.2025.107187.

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. https://doi.org/10.1016/j.jbef.2015.12.001.

Chugunova, M., & Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, *99*, 101897. https://doi.org/10.1016/j.socec.2022.101897.

Church, K. (2024). Emerging trends: When can users trust GPT, and when should they intervene? *Natural Language Engineering*, *30*(2), 417–427. https://doi.org/10.1017/S1351324923000578.

Daston, L. (2022). *Rules: A short history of what we live by*. Princeton University Press.

de Boer, H., Luo, H., Musshoff, O., & Hermann, D. (mimeo). *Do we trust humans or AI more? - Experimental evidence on individual and group trust behavior* (Working Paper).

Dvorak, F., Stumpf, R., Fehrler, S., & Fischbacher, U. (2025). Adverse reactions to the use of large language models in social interactions. *PNAS Nexus*, *4*(4), pgaf112. https://doi.org/10.1093/pnasnexus/pgaf112.

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, *13*, 100060. https://doi.org/10.1016/j.jrt.2023.100060.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, *133*(4), 1645–1692. https://doi.org/10.1093/qje/qjy013.

Fallucchi, F., Fromell, H., & Nosenzo, D. (2026). Incentivizing social norm elicitation. *Experimental Economics*, 1–22. https://doi.org/10.1017/eec.2025.10036.

Fallucchi, F., & Nosenzo, D. (2022). The coordinating power of social norms. *Experimental Economics*, *25*(1), 1–25. https://doi.org/10.1007/s10683-021-09717-8.

Gächter, S., Molleman, L., & Nosenzo, D. (2025). Why people follow rules. *Nature Human Behaviour*, *9*, 1342–1354. https://doi.org/10.1038/s41562-025-02196-4.

Gächter, S., Starmer, C., & Tufano, F. (2015). Measuring the Closeness of Relationships: A Comprehensive Evaluation of the 'Inclusion of the Other in the Self' Scale. *PLOS ONE*, *10*(6), e0129478. https://doi.org/10.1371/journal.pone.0129478.

Gelfand, M. (2018). *Rule makers, rule breakers: Tight and loose cultures and the secret signals that direct our lives*. Scribner.

Greiner, B., Grünwald, P., Lindner, T., Lintner, G., & Wiernsperger, M. (2025). Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study. *Management Science*, *72*(1), 302–322. https://doi.org/10.1287/mnsc.2022.02777.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (NBER Working Paper No. 31122). http://www.nber.org/papers/w31122.

Hunold, M., & Werner, T. (2025). Algorithmic price recommendations and collusion: Experimental evidence. *Experimental Economics*, *28*(2), 298–316. https://doi.org/10.1017/eec.2025.9.

Jayasekara, D., Prissé, B., Deng, R., & Ho, J. Q. (2025). *Exploring Trust in Artificial Intelligence (AI) Systems: Insights from a Repeated Trust Game* (SSRN Working Paper No. 5229860). https://doi.org/10.2139/ssrn.5229860.

Kaplan, J. (2016). *Artificial intelligence: What everyone needs to know*. Oxford University Press.

Kasberger, B., Martin, S., Normann, H.-T., & Werner, T. (2024). Algorithmic cooperation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4855849.

Kimbrough, E. O., Krupka, E. L., Kumar, R., Murray, J. M., Ramalingam, A., Sánchez-Franco, S., Sarmiento, O. L., Kee, F., & Hunter, R. F. (2024). On the stability of norms and norm-following propensity: A cross-cultural panel study with adolescents. *Experimental Economics*, *27*(2), 351–378. https://doi.org/10.1007/s10683-024-09821-5.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms Make Preferences Social. *Journal of the European Economic Association*, *14*(3), 608–638. https://doi.org/10.1111/jeea.12152.

Kimbrough, E. O., & Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, *168*, 147–150. https://doi.org/10.1016/j.econlet.2018.04.030.

Kliemt, H. (2020). Economic and Sociological Accounts of Social Norms. *Analyse & Kritik*, *42*(1), 41–96. https://doi.org/10.1515/auk-2020-0003.

Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2.

Korinek, A. (2023). *Language Models and Cognitive Automation for Economic Research* (NBER Working Paper No. 30957). http://www.nber.org/papers/w30957.

Krittanawong, C. (2018). The rise of artificial intelligence and the uncertain future for physicians. *European Journal of Internal Medicine*, *48*, e13–e14. https://doi.org/10.1016/j.ejim.2017.06.017.

Krupka, E. L., & Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, *11*(3), 495–524. https://doi.org/10.1111/jeea.12006.

Langer, M., König, C. J., Back, C., & Hemsing, V. (2023). Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias. *Journal of Business and Psychology*, *38*(3), 493–508. https://doi.org/10.1007/s10869-022-09829-9.

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, *8*(10), 1906–1914. https://doi.org/10.1038/s41562-024-01953-1.

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. https://doi.org/10.1093/ej/uead056.

Liehner, G. L., Biermann, H., Hick, A., Brauner, P., & Ziefle, M. (2023). Perceptions, Attitudes and Trust Towards Artificial Intelligence — An Assessment of the Public Opinion. *Artificial Intelligence and Social Computing*, *72*, 32–41. https://doi.org/10.54941/ahfe1003271.

Livingston, J. A., Rankich, K., & Shen, S. (2025). *Do People Trust Generative AI, and is it Trustworthy? Evidence from Playing Trust Games with ChatGPT* (SSRN Working Paper No. 5260776). https://doi.org/10.2139/ssrn.5260776.

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. https://doi.org/https://doi.org/10.1016/j.obhdp.2018.12.005.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121.

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371. https://doi.org/10.1037/h0040525.

Mirbabaie, M., Brünker, F., Möllmann Frick, N. R., & Stieglitz, S. (2022). The rise of artificial intelligence - understanding the AI identity threat at the workplace. *Electronic Markets*, *32*(1), 73–99. https://doi.org/10.1007/s12525-021-00496-x.

Naik, D., Naik, I., & Naik, N. (2024). Imperfectly Perfect AI Chatbots: Limitations of Generative AI, Large Language Models and Large Multimodal Models. *The International Conference on Computing, Communication, Cybersecurity & AI*, 43–66. https://doi.org/10.1007/978-3-031-74443-3_3.

Normann, H.-T., Rulié, N., Stypa, O., & Werner, T. (2025). *Delegate pricing decisions to an algorithm? experimental evidence* (arXiv preprint arXiv:2510.27636). https://doi.org/10.48550/arXiv.2510.27636.

Orchard, T., & Tasiemski, L. (2023). The rise of generative AI and possible effects on the economy. *Economics and Business Review*, *9*(2), 9–26. https://doi.org/10.18559/ebr.2023.2.732.

Polachek, S. W., Romano, K., & Tonguc, O. (2025). Homo-silicus: Not (yet) a good imitator of homo sapiens or homo economicus. *Journal of the Economic Science Association*, 1–9. https://doi.org/10.1017/esa.2025.10023.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rana, K., & Khatri, N. (2024). Automotive intelligence: Unleashing the potential of AI beyond advance driver assisting system, a comprehensive review. *Computers and Electrical Engineering*, *117*, 109237. https://doi.org/10.1016/j.compeleceng.2024.109237.

Roberts, T., & Oosterom, M. (2025). Digital authoritarianism: A systematic literature review. *Information Technology for Development*, *31*(4), 860–884. https://doi.org/10.1080/02681102.2024.2425352.

Satornino, C. B., Du, S., & Grewal, D. (2024). Using artificial intelligence to advance sustainable development in industrial markets: A complex adaptive systems perspective. *Industrial Marketing Management*, *116*, 145–157. https://doi.org/10.1016/j.indmarman.2023.11.011.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168, Vol. 1). Mohr.

Selten, R. (1998). Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics*, *1*, 43–61. https://doi.org/10.1023/A:1009957816843.

Stein, J.-P., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: Measurement and associations with personality. *Scientific Reports*, *14*, 2909. https://doi.org/10.1038/s41598-024-53335-2.

Suri, D., Gächter, S., Kube, S., & Schultz, J. (2026). *The resilience of rule compliance in a polarized society* (CeDEx Discussion Paper No. 2026-01).

Suri, D., Kube, S., & Schultz, J. (2025). Evaluating Online Data Collection Platforms Using A Simple Rule-Following Task. *Economics Letters*, *255*, 112509. https://doi.org/10.1016/j.econlet.2025.112509.

Tirole, J. (2021). Digital Dystopia. *American Economic Review*, *111*(6), 2007–48. https://doi.org/10.1257/aer.20201214.

Walter, J., Biermann, J., & Horton, J. (2024). *Advised by an algorithm: Learning with different informational resources* (Beiträge zur Jahrestagung des Vereins für Socialpolitik 2024: Upcoming Labor Market Challenges, ZBW - Leibniz Information Centre for Economics). https://hdl.handle.net/10419/302407.

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122. https://doi.org/10.1073/pnas.2518075122.

Xie, Y., Mei, Q., Yuan, W., & Jackson, M. O. (2025). Using large language models to categorize strategic situations and decipher motivations behind human behaviors. *Proceedings of the National Academy of Sciences*, *122*(35), e2512075122. https://doi.org/10.1073/pnas.2512075122.

Xu, Y., Zhou, G., Cai, R., & Gursoy, D. (2024). When disclosing the artificial intelligence (AI) technology integration into service delivery backfires: Roles of fear of AI, identity threat

and existential threat. *International Journal of Hospitality Management*, *122*, 103829. https://doi.org/10.1016/j.ijhm.2024.103829.

Yin, J., Ngiam, K. Y., Tan, S. S.-L., & Teo, H. H. (2025). Designing AI-Based Work Processes: How the Timing of AI Advice Affects Diagnostic Decision Making. *Management Science*, *71*(11), 9361–9383. https://doi.org/10.1287/mnsc.2022.01454.

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, *21*, 21. https://doi.org/10.1186/s41239-024-00453-6.

Zejjari, I., & Benhayoun, I. (2024). The use of artificial intelligence to advance sustainable supply chain: Retrospective and future avenues explored through bibliometric analysis. *Discover Sustainability*, *5*(1), 174. https://doi.org/10.1007/s43621-024-00364-6.

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends* (SSRN Working Paper No. 3312874). https://doi.org/10.2139/ssrn.3312874.

Zhang, Z., Navarese, E. P., Zheng, B., Meng, Q., Liu, N., Ge, H., Pan, Q., Yu, Y., & Ma, X. (2020). Analytics with artificial intelligence to advance the treatment of acute respiratory distress syndrome. *Journal of Evidence-Based Medicine*, *13*(4), 301–312. https://doi.org/10.1111/jebm.12418.

Zuboff, S. (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future and the New Frontier of Power*. PublicAffairs.

# Supplementary Information

for

**AI versus humans as authority figures:**
**Evidence from a rule-compliance experiment**

Dominik Suri*, Simon Gächter, Sebastian Kube
*Corresponding author. E-mail: dsuri@uni-bonn.de.

## Contents

## SI-1 Supplementary Figures

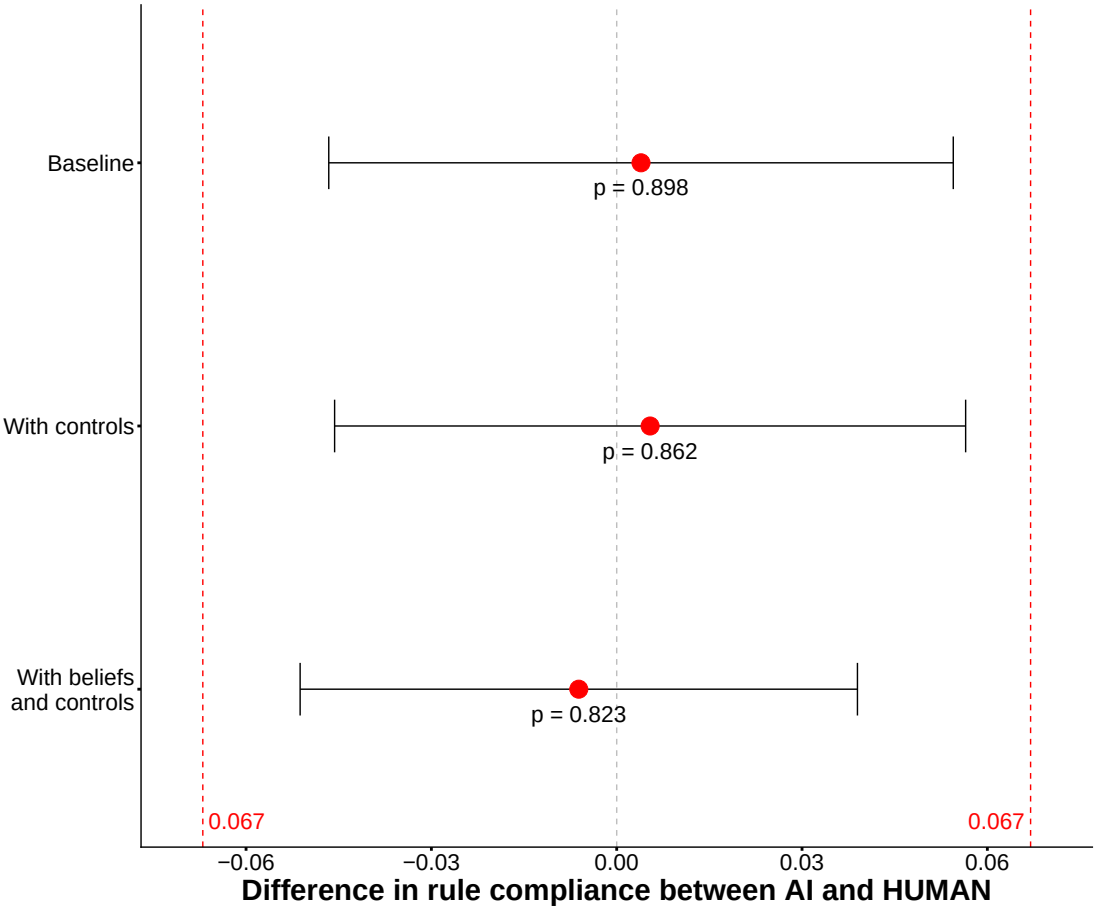



**Figure SI-1: Equivalence testing of differences in rule compliance across treatment conditions.** The figure shows the treatment effect of AI as regression coefficients of linear probability models with 90% confidence intervals. "Baseline" corresponds to a model where the treatment is regressed on rule compliance. "With controls" displays the regression coefficient of AI in Table SI-3, Column 1. "With beliefs and controls" displays the regression estimate of AI in Table SI-3, Column 2. The red dotted line indicates the maximum region of practical equivalence for differences between rule-setter identities outlined in Suri et al. (2026). p-values are obtained using heteroskedasticity-robust standard errors.

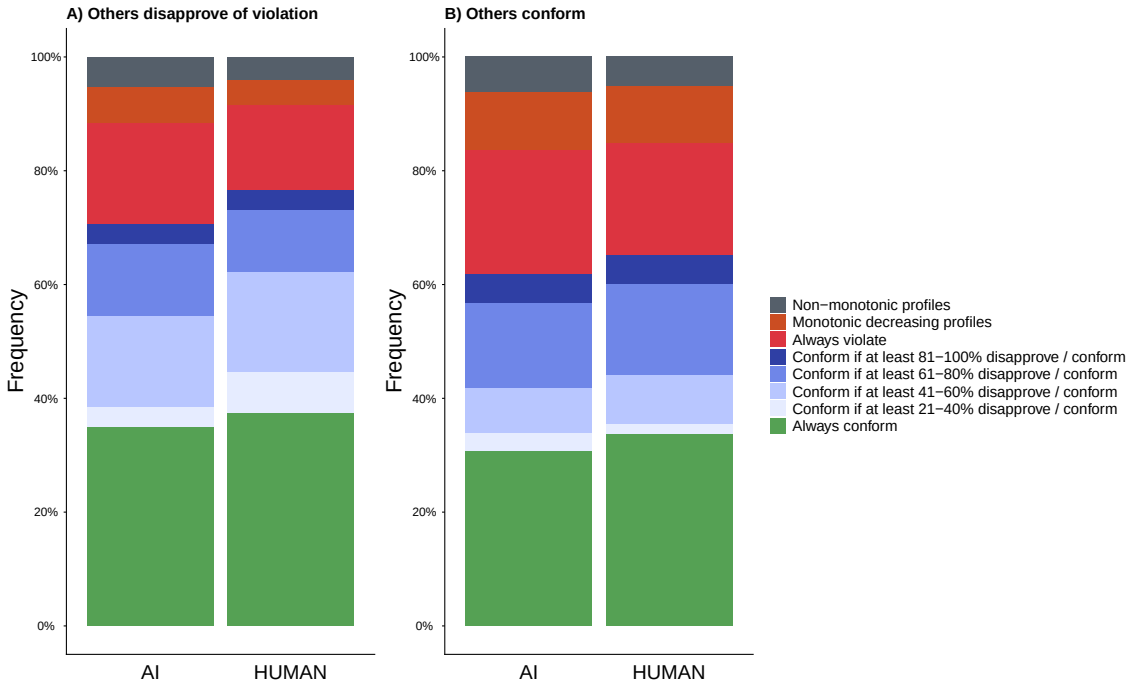



**Figure SI-2: Individual profiles of rule-conformity conditional on social expectations.** The figure shows the individual-level breakdown of Figures 2B (left, A) and 2C (right, B).

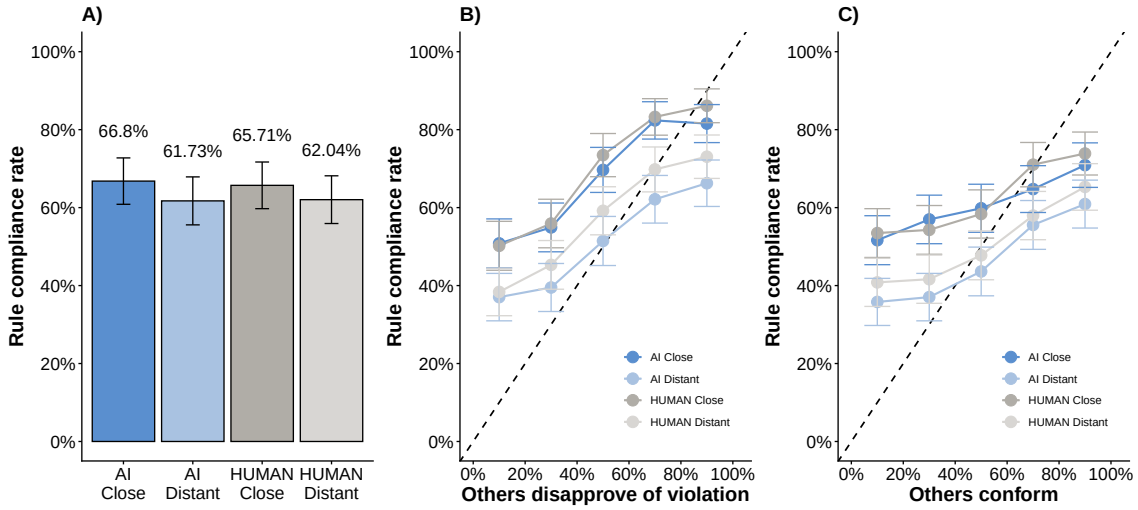



**Figure SI-3: Rule compliance given the subjective closeness to the rule setter.** The figure shows incentivized rule compliance in the one-shot coins task (left, A) and conditional on the percentage (presented as quintiles) of others who disapprove of rule violation (middle, B) as well as conform with the rule (right, C); elicited with the strategy method. The dashed lines are the diagonals. The whiskers represent the 95% confidence intervals.

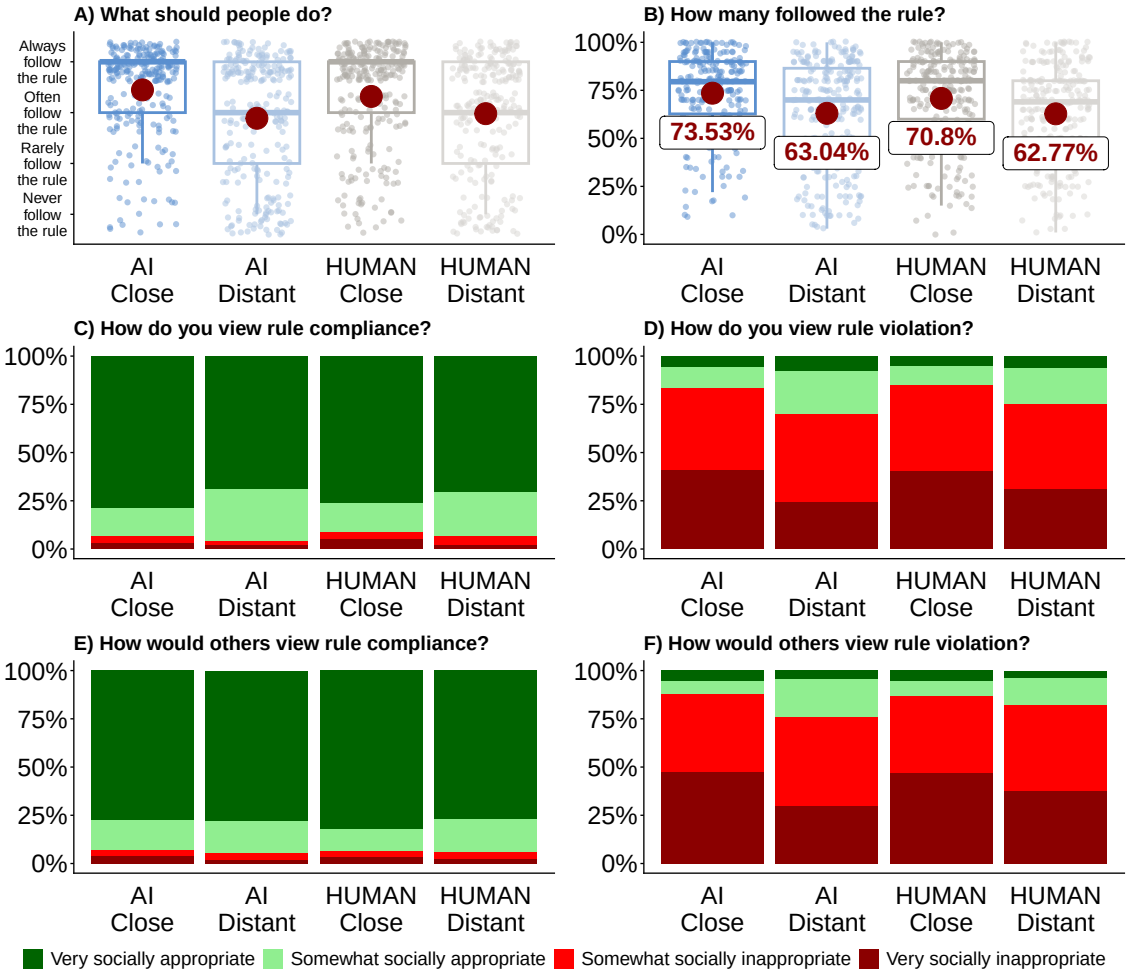



**Figure SI-4: Beliefs given the subjective closeness to the rule setter.** The figure shows the general rule-following attitude (top left, A) and incentivized descriptive (top right, B) and personal as well as social normative beliefs of rule compliance (middle left, C and bottom left, E) and rule violation (middle right, D and bottom right, F); elicited with the quadratic scoring rule (B) or the strategy method (C to F). The solid lines represent the median and the red dots the mean values (A and B).

## SI-2 Supplementary Tables

**Table SI-1:** Determinants of rule compliance if the rule is set by AI

| | Dependent variable: Rule compliance | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.60*** | 0.15 | 0.22* | 0.62*** | 0.57*** | 0.08 |
| | (0.11) | (0.12) | (0.11) | (0.12) | (0.12) | (0.12) |
| General rule-following attitude | | 0.20*** | | | | 0.12*** |
| | | (0.02) | | | | (0.03) |
| Descriptive belief | | | 0.01*** | | | 0.01*** |
| | | | (0.001) | | | (0.001) |
| Personal normative belief (compliance) | | | | 0.07 | | −0.05 |
| | | | | (0.06) | | (0.05) |
| Personal normative belief (violation) | | | | −0.22*** | | −0.08 |
| | | | | (0.04) | | (0.05) |
| Social normative belief (compliance) | | | | | 0.06 | 0.05 |
| | | | | | (0.05) | (0.05) |
| Social normative belief (violation) | | | | | −0.15*** | 0.05 |
| | | | | | (0.04) | (0.05) |
| ATTARI | | −0.04 | −0.03 | −0.04 | −0.05 | −0.03 |
| | | (0.04) | (0.03) | (0.04) | (0.04) | (0.03) |
| Trust in rule setter | −0.01 | −0.01 | −0.01 | 0.004 | 0.02 | −0.02 |
| | (0.02) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Patience | 0.02 | 0.01 | 0.01 | 0.01 | 0.02 | 0.003 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Age | −0.001 | −0.002 | −0.004* | −0.002 | −0.001 | −0.003* |
| | (0.002) | (0.001) | (0.001) | (0.002) | (0.002) | (0.001) |
| Male | 0.04 | 0.05 | 0.05 | 0.05 | 0.04 | 0.06 |
| | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Democrat | −0.01 | 0.02 | −0.01 | 0.02 | 0.004 | 0.01 |
| | (0.05) | (0.04) | (0.04) | (0.05) | (0.05) | (0.04) |
| Northeast | −0.06 | −0.02 | −0.01 | −0.04 | −0.05 | −0.01 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) |
| Midwest | 0.02 | 0.0003 | 0.01 | 0.02 | 0.01 | 0.01 |
| | (0.06) | (0.05) | (0.05) | (0.06) | (0.06) | (0.05) |
| West | 0.01 | 0.02 | −0.001 | −0.01 | 0.01 | 0.003 |
| | (0.06) | (0.06) | (0.05) | (0.06) | (0.06) | (0.05) |
| Observations | 480 | 480 | 480 | 480 | 480 | 480 |
| Adjusted $R^2$ | −0.004 | 0.20 | 0.22 | 0.08 | 0.03 | 0.27 |
| F Statistic | 0.75 | 12.63*** | 14.67*** | 4.60*** | 2.33** | 13.01*** |
| Degrees of freedom | (8; 471) | (10; 469) | (10; 469) | (11; 468) | (11; 468) | (15; 464) |

*Notes:* The table reports coefficient estimates from linear probability models. Only data from the treatment condition AI is used. The dependent variable is rule compliance. Robust standard errors are in parentheses. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

**Table SI-2:** Determinants of rule compliance if the rule is set by a US Prolific user

| | Dependent variable: Rule compliance | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.70*** | 0.30* | 0.40** | 0.71*** | 0.66*** | 0.29* |
| | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) | (0.13) |
| General rule-following attitude | | 0.18*** | | | | 0.13*** |
| | | (0.02) | | | | (0.03) |
| Descriptive belief | | | 0.01*** | | | 0.004*** |
| | | | (0.001) | | | (0.001) |
| Personal normative belief (compliance) | | | | 0.05 | | −0.02 |
| | | | | (0.05) | | (0.04) |
| Personal normative belief (violation) | | | | −0.19*** | | −0.06 |
| | | | | (0.04) | | (0.05) |
| Social normative belief (compliance) | | | | | −0.003 | −0.05 |
| | | | | | (0.05) | (0.05) |
| Social normative belief (violation) | | | | | −0.16*** | 0.03 |
| | | | | | (0.04) | (0.05) |
| Trust in rule setter | −0.01 | −0.03 | −0.03 | −0.02 | −0.01 | −0.04 |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) |
| Patience | −0.005 | −0.01 | −0.01 | −0.01 | −0.01 | −0.01 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Age | 0.0003 | −0.001 | −0.002 | −0.0005 | 0.0001 | −0.002 |
| | (0.002) | (0.001) | (0.001) | (0.002) | (0.002) | (0.001) |
| Male | −0.04 | −0.03 | −0.04 | −0.05 | −0.05 | −0.04 |
| | (0.05) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Democrat | −0.01 | 0.02 | 0.01 | −0.01 | −0.01 | 0.03 |
| | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Northeast | 0.02 | 0.002 | 0.004 | 0.002 | 0.01 | −0.001 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) |
| Midwest | −0.02 | −0.004 | −0.01 | −0.03 | −0.03 | −0.01 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) |
| West | 0.07 | 0.05 | 0.05 | 0.05 | 0.06 | 0.04 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.05) |
| Observations | 482 | 482 | 482 | 482 | 482 | 482 |
| Adjusted $R^2$ | −0.01 | 0.14 | 0.11 | 0.05 | 0.02 | 0.17 |
| F Statistic | 0.35 | 9.97*** | 7.57*** | 3.32*** | 1.87* | 8.14*** |
| Degrees of freedom | (8; 473) | (9; 472) | (9; 472) | (10; 471) | (10; 471) | (14; 467) |

*Notes:* The table reports coefficient estimates from linear probability models. Only data from the treatment condition HUMAN is used. The dependent variable is rule compliance. Robust standard errors are in parentheses. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

**Table SI-3:** Impact of rule setter on rule compliance (full table)

| | Dependent variable: Rule compliance | | | |
|---|---|---|---|---|
| | Unconditional | | Conditional | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.64*** | 0.17* | 0.13* | 0.14* |
| | (0.08) | (0.09) | (0.06) | (0.07) |
| AI | 0.01 | −0.01 | −0.03 | −0.03 |
| | (0.03) | (0.03) | (0.02) | (0.02) |
| % others disapprove of violation | | | 0.005*** | |
| | | | (0.0003) | |
| % others conform | | | | 0.003*** |
| | | | | (0.0003) |
| General rule-following attitude | | 0.12*** | | |
| | | (0.02) | | |
| Descriptive belief | | 0.01*** | | |
| | | (0.001) | | |
| Personal normative belief (compliance) | | −0.03 | | |
| | | (0.03) | | |
| Personal normative belief (violation) | | −0.07* | | |
| | | (0.03) | | |
| Social normative belief (compliance) | | −0.001 | | |
| | | (0.04) | | |
| Social normative belief (violation) | | 0.05 | | |
| | | (0.03) | | |
| Trust in rule setter | −0.01 | −0.04** | 0.05** | 0.05*** |
| | (0.02) | (0.02) | (0.01) | (0.01) |
| Patience | 0.01 | −0.01 | 0.01** | 0.01 |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| Age | −0.0003 | −0.003** | 0.002* | 0.002* |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Male | −0.001 | 0.01 | −0.04 | −0.03 |
| | (0.03) | (0.03) | (0.02) | (0.02) |
| Democrat | −0.01 | 0.02 | −0.02 | −0.02 |
| | (0.03) | (0.03) | (0.02) | (0.02) |
| Northeast | −0.02 | −0.01 | −0.01 | 0.01 |
| | (0.05) | (0.04) | (0.03) | (0.03) |
| Midwest | 0.01 | 0.01 | 0.01 | 0.01 |
| | (0.04) | (0.04) | (0.03) | (0.03) |
| West | 0.04 | 0.02 | −0.01 | 0.02 |
| | (0.04) | (0.04) | (0.03) | (0.03) |
| Observations | 962 | 962 | 4,810 | 4,810 |
| Adjusted $R^2$ | −0.01 | 0.22 | 0.09 | 0.04 |
| F Statistic | 0.30 | 19.48*** | 49.46*** | 23.44*** |
| Degrees of freedom | (9; 952) | (15; 946) | (10; 4799) | (10; 4799) |

*Notes:* The table reports coefficient estimates from linear probability models. The dependent variable is rule compliance. Columns 1-2 display data from the one-shot coins task. Column 3 (4) displays data from the rule compliance elicitation conditional on the share of others who disapprove of violation (who conform with the rule) using the strategy method. Standard errors are reported in parentheses. Models 1–2 use heteroskedasticity-robust standard errors. Models 3–4 use standard errors clustered at the subject level. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

**Table SI-4:** Impact of rule setter on rule compliance (logit estimation)

| | Dependent variable: Rule compliance | | | |
|---|---|---|---|---|
| | Unconditional | | Conditional | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.58 | −1.44** | −1.71*** | −1.50*** |
| | (0.36) | (0.47) | (0.30) | (0.28) |
| AI | 0.02 | −0.03 | −0.16 | −0.12 |
| | (0.14) | (0.16) | (0.11) | (0.10) |
| % others disapprove of violation | | | 0.02*** | |
| | | | (0.001) | |
| % others conform | | | | 0.01*** |
| | | | | (0.001) |
| General rule-following attitude | | 0.61*** | | |
| | | (0.09) | | |
| Descriptive belief | | 0.03*** | | |
| | | (0.004) | | |
| Personal normative belief (compliance) | | −0.19 | | |
| | | (0.19) | | |
| Personal normative belief (violation) | | −0.42* | | |
| | | (0.19) | | |
| Social normative belief (compliance) | | 0.003 | | |
| | | (0.20) | | |
| Social normative belief (violation) | | 0.28 | | |
| | | (0.19) | | |
| Trust in rule setter | −0.04 | −0.26** | 0.21** | 0.21*** |
| | (0.08) | (0.09) | (0.07) | (0.06) |
| Patience | 0.03 | −0.03 | 0.06** | 0.04 |
| | (0.03) | (0.03) | (0.02) | (0.02) |
| Age | −0.001 | −0.02** | 0.01* | 0.01* |
| | (0.005) | (0.01) | (0.004) | (0.004) |
| Male | −0.01 | 0.03 | −0.18 | −0.12 |
| | (0.14) | (0.16) | (0.11) | (0.10) |
| Democrat | −0.06 | 0.12 | −0.10 | −0.07 |
| | (0.14) | (0.16) | (0.11) | (0.11) |
| Northeast | −0.09 | −0.05 | −0.04 | 0.03 |
| | (0.19) | (0.23) | (0.15) | (0.15) |
| Midwest | 0.03 | 0.03 | 0.05 | 0.06 |
| | (0.19) | (0.21) | (0.16) | (0.15) |
| West | 0.16 | 0.11 | −0.07 | 0.10 |
| | (0.19) | (0.22) | (0.14) | (0.14) |
| Observations | 962 | 962 | 4,810 | 4,810 |
| Log Likelihood | −626.48 | −508.54 | −2,972.93 | −3,196.04 |
| Akaike Inf. Crit. | 1,272.96 | 1,049.08 | 5,967.87 | 6,414.09 |

*Notes:* The table reports coefficient estimates from logit models. The dependent variable is rule compliance. Columns 1-2 display data from the one-shot coins task. Column 3 (4) displays data from the rule compliance elicitation conditional on the share of others who disapprove of violation (who conform with the rule) using the strategy method. Standard errors are reported in parentheses. Models 1–2 use heteroskedasticity-robust standard errors. Models 3–4 use standard errors clustered at the subject level. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

**Table SI-5:** Impact of subjective closeness on rule compliance

| | Dependent variable: Rule compliance | | | | | |
|---|---|---|---|---|---|---|
| | AI sample | | HUMAN sample | | Full sample | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.63*** | 0.07 | 0.69*** | 0.30* | 0.63*** | 0.18* |
| | (0.11) | (0.12) | (0.13) | (0.13) | (0.08) | (0.08) |
| IOS11 toward rule setter | 0.01 | −0.004 | 0.004 | −0.01 | 0.01 | −0.01 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| AI | | | | | 0.02 | −0.02 |
| | | | | | (0.03) | (0.03) |
| General rule-following attitude | | 0.12*** | | 0.13*** | | 0.13*** |
| | | (0.03) | | (0.03) | | (0.02) |
| Descriptive belief | | 0.01*** | | 0.004*** | | 0.01*** |
| | | (0.001) | | (0.001) | | (0.001) |
| Personal normative belief (compliance) | | −0.05 | | −0.02 | | −0.04 |
| | | (0.05) | | (0.05) | | (0.04) |
| Personal normative belief (violation) | | −0.08 | | −0.07 | | −0.07* |
| | | (0.05) | | (0.05) | | (0.03) |
| Social normative belief (compliance) | | 0.05 | | −0.05 | | −0.002 |
| | | (0.05) | | (0.05) | | (0.04) |
| Social normative belief (violation) | | 0.05 | | 0.03 | | 0.05 |
| | | (0.05) | | (0.05) | | (0.03) |
| ATTARI | −0.04 | −0.03 | | | | |
| | (0.04) | (0.03) | | | | |
| Trust in rule setter | 0.004 | −0.02 | −0.01 | −0.03 | −0.02 | −0.04* |
| | (0.04) | (0.04) | (0.03) | (0.03) | (0.02) | (0.02) |
| Patience | 0.02 | 0.003 | −0.01 | −0.01 | 0.01 | −0.01 |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| Age | −0.001 | −0.003* | 0.0003 | −0.002 | −0.0004 | −0.003** |
| | (0.002) | (0.001) | (0.002) | (0.001) | (0.001) | (0.001) |
| Male | 0.04 | 0.06 | −0.04 | −0.04 | −0.002 | 0.01 |
| | (0.05) | (0.04) | (0.05) | (0.04) | (0.03) | (0.03) |
| Democrat | −0.002 | 0.01 | −0.02 | 0.03 | −0.01 | 0.02 |
| | (0.05) | (0.04) | (0.04) | (0.04) | (0.03) | (0.03) |
| Northeast | −0.05 | −0.01 | 0.02 | −0.001 | −0.02 | −0.01 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.04) | (0.04) |
| Midwest | 0.02 | 0.005 | −0.02 | −0.01 | 0.01 | 0.01 |
| | (0.06) | (0.05) | (0.06) | (0.06) | (0.04) | (0.04) |
| West | 0.01 | 0.002 | 0.06 | 0.05 | 0.03 | 0.02 |
| | (0.06) | (0.05) | (0.06) | (0.05) | (0.04) | (0.04) |
| Observations | 480 | 480 | 482 | 482 | 962 | 962 |
| Adjusted $R^2$ | −0.002 | 0.27 | −0.01 | 0.17 | −0.01 | 0.22 |
| F Statistic | 0.92 | 12.19*** | 0.34 | 7.69*** | 0.52 | 18.35*** |
| Degrees of freedom | (10; 469) | (16; 463) | (9; 472) | (15; 466) | (10; 951) | (16; 945) |

*Notes:* The table reports coefficient estimates from linear probability models. Columns 1-2 report only data from the treatment condition AI. Columns 3-4 report only data from the treatment condition HUMAN. Columns 5-6 report data from the full sample. The dependent variable is rule compliance. Robust standard errors are in parentheses. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

**Table SI-6:** Impact of subjective closeness on conditional rule compliance

| | Dependent variable: Conditional rule compliance | | | | | |
|---|---|---|---|---|---|---|
| | AI sample | | HUMAN sample | | Full sample | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 0.20*** | 0.21*** | −0.02 | 0.01 | 0.10** | 0.12** |
| | (0.05) | (0.05) | (0.05) | (0.06) | (0.04) | (0.04) |
| IOS11 toward rule setter | 0.03*** | 0.02*** | 0.02*** | 0.01*** | 0.02*** | 0.02*** |
| | (0.003) | (0.004) | (0.004) | (0.004) | (0.003) | (0.003) |
| % others disapprove of violation | 0.004*** | | 0.005*** | | 0.005*** | |
| | (0.0003) | | (0.0003) | | (0.0002) | |
| % others conform | | 0.003*** | | 0.003*** | | 0.003*** |
| | | (0.0003) | | (0.0003) | | (0.0002) |
| AI | | | | | 0.01 | 0.005 |
| | | | | | (0.01) | (0.01) |
| ATTARI | −0.03 | −0.02 | | | | |
| | (0.02) | (0.02) | | | | |
| Trust in rule setter | 0.02 | 0.02 | 0.07*** | 0.07*** | 0.02* | 0.03*** |
| | (0.02) | (0.02) | (0.01) | (0.01) | (0.01) | (0.01) |
| Patience | 0.02*** | 0.01** | 0.01 | 0.002 | 0.01*** | 0.01** |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.003) | (0.003) |
| Age | 0.001 | 0.001 | 0.003*** | 0.002*** | 0.002*** | 0.002** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.0005) | (0.0005) |
| Male | −0.06** | −0.03 | −0.02 | −0.03 | −0.04** | −0.03* |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.01) | (0.01) |
| Democrat | −0.01 | −0.001 | −0.03 | −0.03 | −0.02 | −0.01 |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.01) | (0.01) |
| Northeast | −0.02 | −0.03 | 0.002 | 0.05 | −0.004 | 0.01 |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.02) | (0.02) |
| Midwest | 0.06* | 0.02 | −0.04 | 0.01 | 0.01 | 0.02 |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.02) | (0.02) |
| West | −0.04 | −0.03 | −0.0003 | 0.08** | −0.02 | 0.02 |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.02) | (0.02) |
| Observations | 2,400 | 2,400 | 2,410 | 2,410 | 4,810 | 4,810 |
| Adjusted $R^2$ | 0.11 | 0.05 | 0.12 | 0.06 | 0.11 | 0.05 |
| F Statistic | 27.49*** | 12.77*** | 32.49*** | 17.25*** | 52.94*** | 25.78*** |
| Degrees of freedom | (11; 2388) | (11; 2388) | (10; 2399) | (10; 2399) | (11; 4798) | (11; 4798) |

*Notes:* The table reports coefficient estimates from linear probability models. Columns 1-2 report only data from the treatment condition AI. Columns 3-4 report only data from the treatment condition HUMAN. Columns 5-6 report data from the full sample. Columns 1, 3, and 5 report only data from the conditional rule compliance task based on the share of others who disapprove of rule violation. Columns 2, 4, and 6 report only data from the conditional rule compliance task based on the share of others who conform with the rule. The dependent variable is conditional rule compliance. Clustered standard errors at the subject level are in parentheses. Levels of significance: *p<0.05, **p<0.01, ***p<0.001.

## SI-3 Experimental Instructions

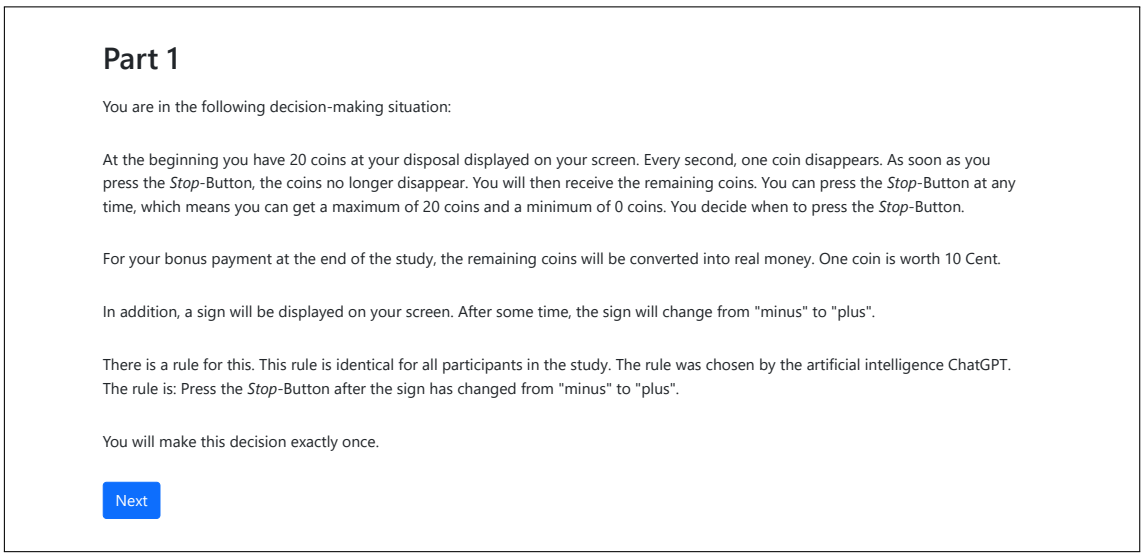
This section presents the experimental instructions as original screenshots shown to participants. Each frame indicates a new screen. Where appropriate, we add comments *** *in italic text between asterisks* ***. These comments were not shown to participants.

Prior to the instructions shown below, we asked participants for their Prolific ID, to provide informed consent and told them that one of the four incentivized parts will be randomly selected to determine their bonus payment for this study. The four parts are:

1. The coins task (see Section SI-3.1)
2. The belief elicitation (see SectionSI-3.2)
3. Rule compliance conditional on others disapprove of violation (see SectionSI-3.3)
4. Rule compliance conditional on others conform (see SectionSI-3.4)

At the end of the study, we provided information on the payment as well as the completion code to be entered into Prolific.

### SI-3.1 The coins task




*** *The content above was only shown in the* `AI` *condition.* ***

## Part 1

You are in the following decision-making situation:

At the beginning you have 20 coins at your disposal displayed on your screen. Every second, one coin disappears. As soon as you press the *Stop*-Button, the coins no longer disappear. You will then receive the remaining coins. You can press the *Stop*-Button at any time, which means you can get a maximum of 20 coins and a minimum of 0 coins. You decide when to press the *Stop*-Button.

For your bonus payment at the end of the study, the remaining coins will be converted into real money. One coin is worth 10 Cent.

In addition, a sign will be displayed on your screen. After some time, the sign will change from "minus" to "plus".

There is a rule for this. This rule is identical for all participants in the study. The rule was chosen by a US Prolific user. The rule is: Press the *Stop*-Button after the sign has changed from "minus" to "plus".

You will make this decision exactly once.

Next

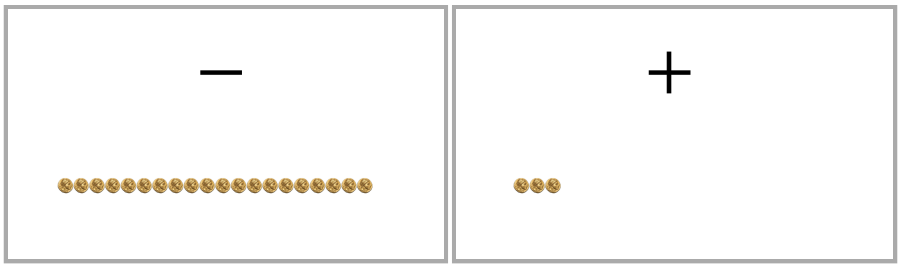
*** *The content above was only shown in the `HUMAN` condition.* ***

## Part 1

Below you can see the previously described decision-making situation using two examples.

The left image shows the starting position. The sign is "minus" and there are 20 coins.

The right image shows a randomly chosen later position. The sign is "plus" and there are 3 coins.

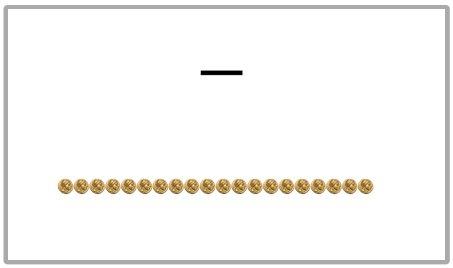



Next

## Part 1

In the following, you will be asked to answer some short comprehension questions. We use these questions to check whether all participants have understood the decision-making situation.

Show instructions again



How many coins would you get if you pressed *Stop*-Button now?

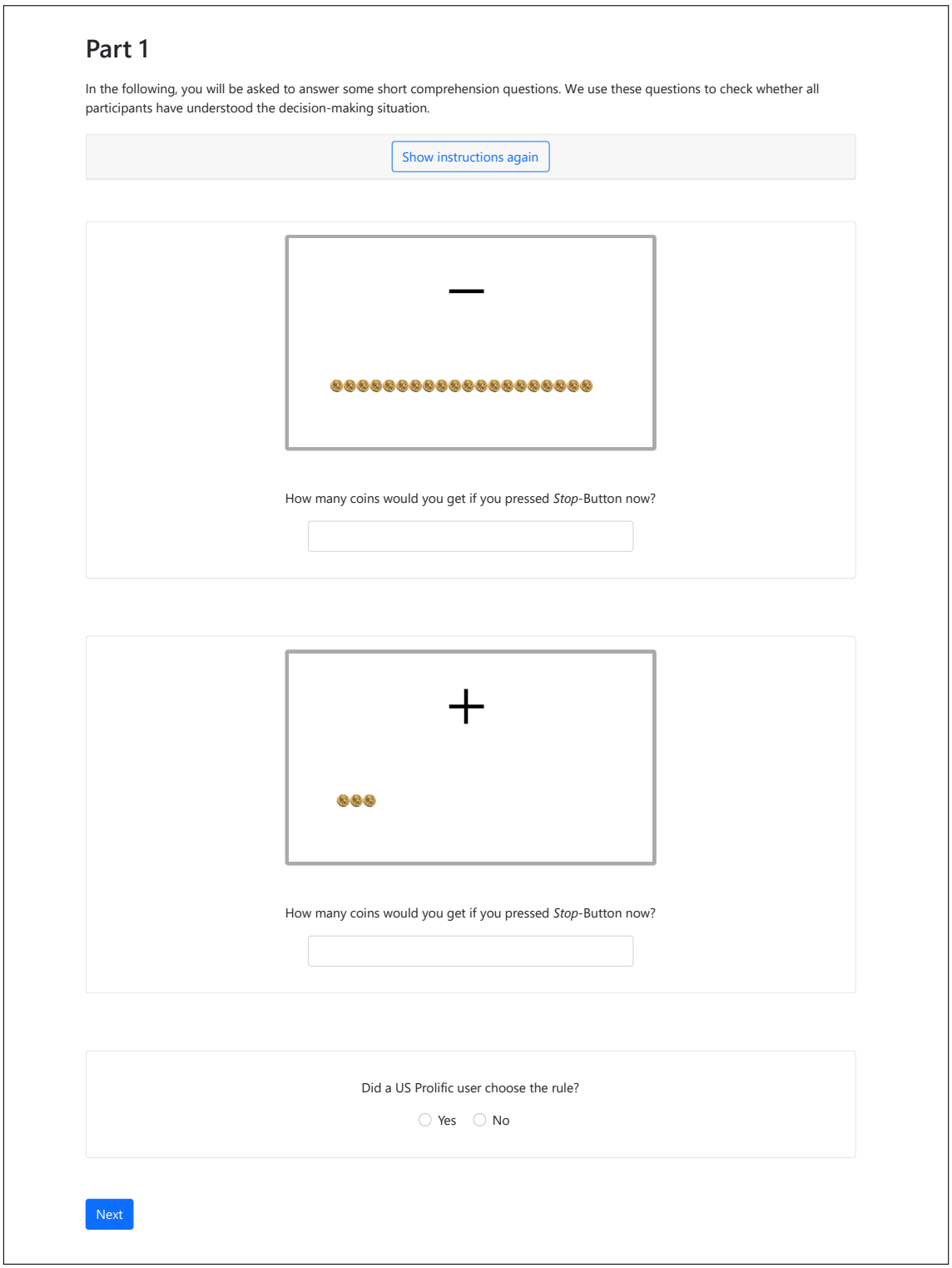
How many coins would you get if you pressed *Stop*-Button now?

Did ChatGPT choose the rule?

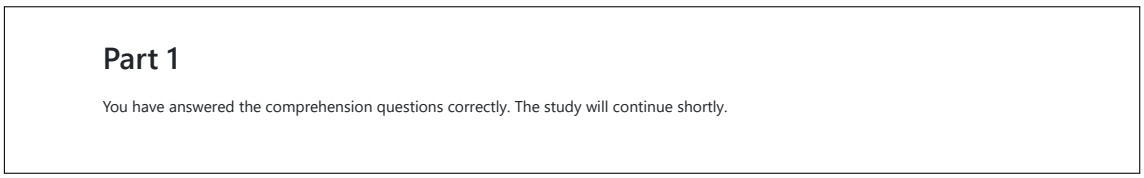
Yes No

Next

*** *The content above was only shown in the `AI` condition.* ***

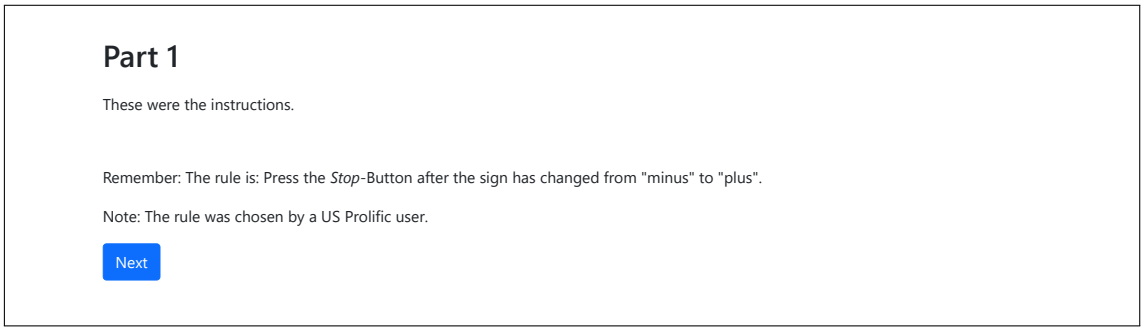



*** *The content above was only shown in the* `HUMAN` *condition.* ***

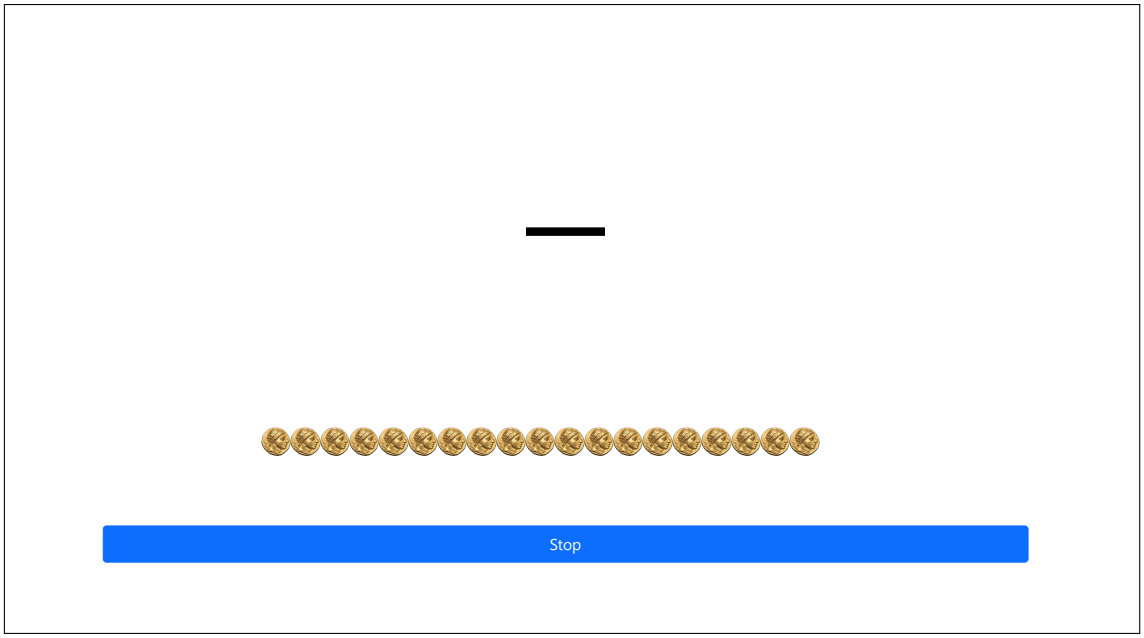




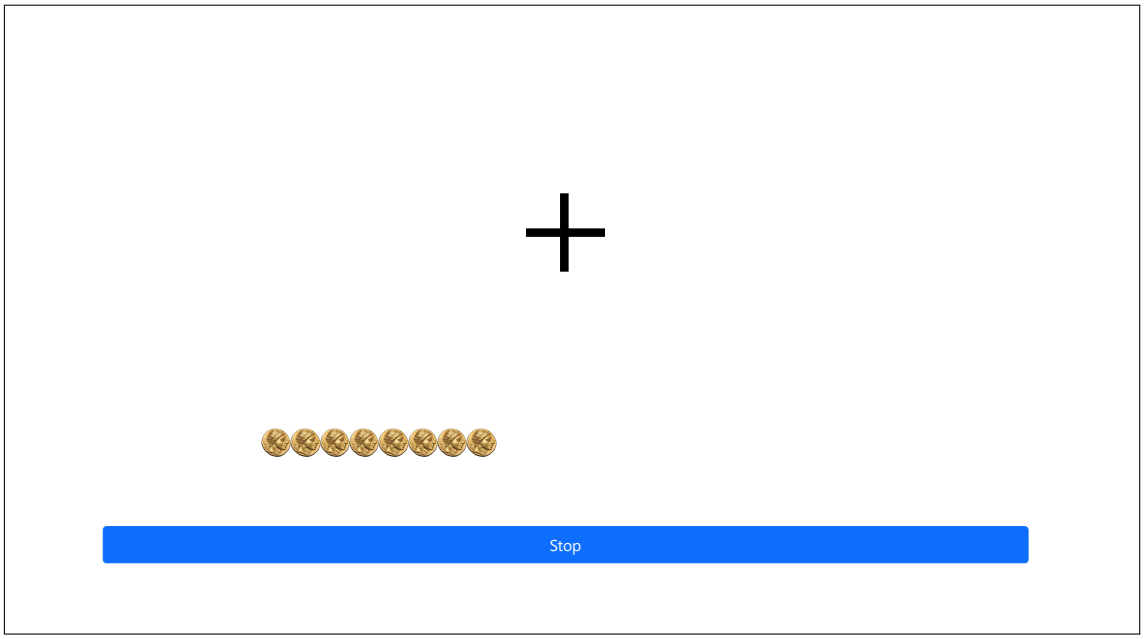
*** *The content above was only shown in the `AI` condition.* ***



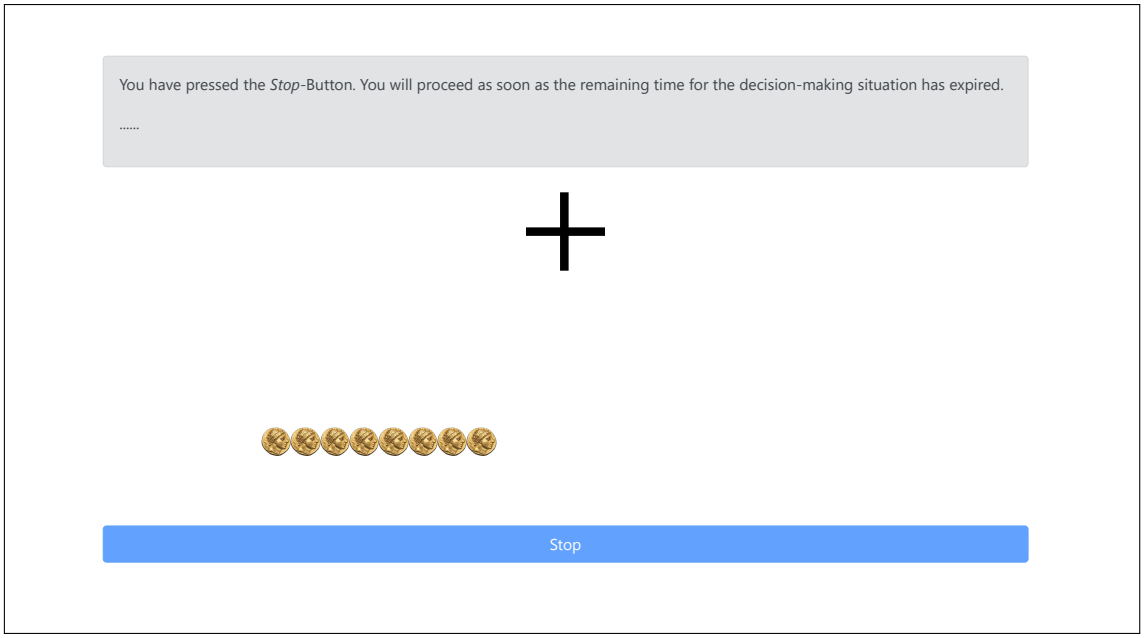

*** *The content above was only shown in the `HUMAN` condition.* ***




*** *The content above displays the start of the coins task.* ***




*** *The content above displays the coins task after 12 seconds have passed and the sign has just changed. The "Stop"-button has not been pressed yet.* ***




*** *The content above displays the coins task after 15 seconds have passed. The sign changed after 12 seconds. The participant pressed the "Stop"-button immediately after the sign change. The 6 dots in the box above the sign indicate that 3 seconds have passed since the button has been pressed.* ***

## SI-3.2 Belief elicitation

### Part 2

What do you think people should do in decision-making situations like this one?

- ○ Never follow rule
- ○ Rarely follow rule
- ○ Often follow rule
- ○ Always follow rule

Next

### Part 2

The following two questions will ask how socially appropriate certain behavior is. By socially appropriate, we mean behavior that you think most other US Prolific users would agree is the "correct" thing to do. Another way to think about what we mean is that if someone were to behave in a socially inappropriate way, then other people might be angry at them.

Assume that a US Prolific user **had followed** the rule. How do you view this?

- ○ very socially inappropriate
- ○ somewhat socially inappropriate
- ○ somewhat socially appropriate
- ○ very socially appropriate

Assume that a US Prolific user **had not followed** the rule. How do you view this?

- ○ very socially inappropriate
- ○ somewhat socially inappropriate
- ○ somewhat socially appropriate
- ○ very socially appropriate

Next

### Part 2

Many US Prolific users are currently participating in the same study as you. All of them are in the same decision-making situation in Part 1, following the same instructions and rules. They also answer the previous two questions about socially appropriate behavior if another US Prolific user had followed the rule or had not followed the rule.

Your answers to the following two questions can affect your **bonus payment**. To earn this bonus, you need to identify, for each question, the answer most often chosen by other US Prolific users in this study. You can earn $0.50 for each correct answer.

Assume that a US Prolific user **had followed** the rule. How do you think most of the other US Prolific users would view this?

- ○ very socially inappropriate
- ○ somewhat socially inappropriate
- ○ somewhat socially appropriate
- ○ very socially appropriate

Assume that a US Prolific user **had not followed** the rule. How do you think most of the other US Prolific users would view this?

- ○ very socially inappropriate
- ○ somewhat socially inappropriate
- ○ somewhat socially appropriate
- ○ very socially appropriate

Next

## Part 2

Now, think about what the other US Prolific users do in Part 1.

Your answer to the following question can also affect your **bonus payment**. To earn this bonus, you need to estimate the share of US Prolific users who follow the rule in Part 1, given the same instructions and rule as you received. You can earn up to a $1.00 as a bonus payment the closer your answer is to the correct result. You can earn a maximum of $1.00 and a minimum of $0.00

Show details on the calculation of the bonus payment

**What do you think, how many of the US Prolific users follow the rule?**

*Note: To adjust the slider with higher precision, first click on it, then move it. The Next-Button becomes clickable once you have clicked on the slider.*

Next

## SI-3.3 Rule compliance conditional on others disapprove of violation

### Part 3

In Part 3, you will again answer some questions around the decision-making situation in Part 1.

Next

### Part 3

Recall that many US Prolific users are currently participating in the same study as you. All of them are in the same decision-making situation in Part 1, following the same instructions and rules. They also answer how socially appropriate it is if another US Prolific user had not followed the rule.

Next

### Part 3

Their answers can be categorized into 5 possible outcomes which differ in the fraction of US Prolific users who indicate that **breaking the rule** and pressing the *Stop*-Button before the sign has changed is **socially inappropriate**:

- A. Almost no US Prolific users (0%-20%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially inappropriate.
- B. A minority of US Prolific users (21%-40%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially inappropriate.
- C. About half of US Prolific users (41%-60%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially inappropriate.
- D. A majority of US Prolific users (61%-80%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially inappropriate.
- E. Almost all US Prolific users (81%-100%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially inappropriate.

Next

### Part 3

We will ask you, for each outcome, whether you want to follow the rule and wait for the sign to change or to break the rule and press the *Stop*-Button before the sign has changed. As you might have noticed in Part 1, the sign changes from "minus" to "plus" after 12 seconds. This means, if you decide to follow the rule, you will receive the remaining 8 coins. If you decide to not follow the rule, you will receive all 20 coins.

Your answers can affect your **bonus payment**. We will look at the actual outcome, i.e., the fraction of US Prolific users who indicate that breaking the rule is socially inappropriate. We will then use your choice corresponding to this outcome. For example, suppose that the actual outcome is 50% of US Prolific users indicate that breaking the rule is socially inappropriate. In that case, your choice for outcome C will be used to compute your bonus payment.

Next

## Part 3

Almost no US Prolific users (0%-20%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially *inappropriate*.

- ○ I follow the rule and wait for the sign to change.
- ○ I break the rule and press the button before the sign has changed.

A minority of US Prolific users (21%-40%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially *inappropriate*.

- ○ I follow the rule and wait for the sign to change.
- ○ I break the rule and press the button before the sign has changed.

About half of US Prolific users (41%-60%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially *inappropriate*.

- ○ I follow the rule and wait for the sign to change.
- ○ I break the rule and press the button before the sign has changed.

A majority of US Prolific users (61%-80%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially *inappropriate*.

- ○ I follow the rule and wait for the sign to change.
- ○ I break the rule and press the button before the sign has changed.

Almost all US Prolific users (81%-100%) indicate that breaking the rule and pressing the *Stop*-Button *before* the sign changes to "plus" is socially *inappropriate*.

- ○ I follow the rule and wait for the sign to change.
- ○ I break the rule and press the button before the sign has changed.

Next

## SI - 3.4 Rule compliance conditional on others conform

### Part 4

In Part 4, you will again answer some questions around the decision-making situation in Part 1.

Next

### Part 4

Now, we present you 5 possible outcomes of the **actual behavior** of other US Prolific users. Specifically, these outcomes differ in the fraction of US Prolific users who **break the rule** and press the *Stop*-Button before the sign has changed:

- A. Almost no US Prolific users (0%-20%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".
- B. A minority of US Prolific users (21%-40%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".
- C. About half of US Prolific users (41%-60%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".
- D. A majority of US Prolific users (61%-80%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".
- E. Almost all US Prolific users (81%-100%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".

Next

### Part 4

We will ask you, for each outcome, whether you want to follow the rule and wait for the sign to change or to break the rule and press the *Stop*-Button before the sign has changed. Again, the sign changes from "minus" to "plus" after 12 seconds. This means, if you decide to follow the rule, you will receive the remaining 8 coins. If you decide to not follow the rule, you will receive all 20 coins.

Your answers can affect your **bonus payment**. We will look at the actual outcome, i.e., the fraction of US Prolific users break the rule. We will then use your choice corresponding to this outcome. For example, suppose that the actual outcome is 50% of US Prolific users break the rule. In that case, your choice for outcome C will be used to compute your bonus payment.

Next

## Part 4

A. Almost no US Prolific users (0%-20%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".

- I break the rule and press the button before the sign has changed.
- I follow the rule and wait for the sign to change.

B. A minority of US Prolific users (21%-40%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".

- I break the rule and press the button before the sign has changed.
- I follow the rule and wait for the sign to change.

C. About half of US Prolific users (41%-60%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".

- I break the rule and press the button before the sign has changed.
- I follow the rule and wait for the sign to change.

D. A majority of US Prolific users (61%-80%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".

- I break the rule and press the button before the sign has changed.
- I follow the rule and wait for the sign to change.

E. Almost all US Prolific users (81%-100%) break the rule and press the *Stop*-Button *before* the sign changes to "plus".

- I break the rule and press the button before the sign has changed.
- I follow the rule and wait for the sign to change.

Next

## SI-3.5 Questionnaire

### Questionnaire

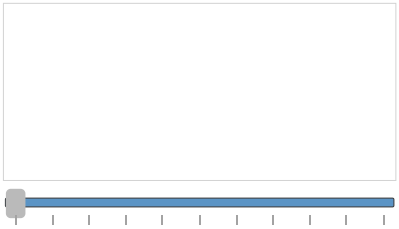
Now, we would like to ask you a few brief questions about yourself. Again, your responses are anonymous and will only be used for research purposes.

Next

### Questionnaire

Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. You should interpret the degree of overlap as representing the relationship between you and "P". "P" serves as a placeholder for **US Prolific users**.

Please position the slider so that the circles indicate to what extent you and "P" are connected.



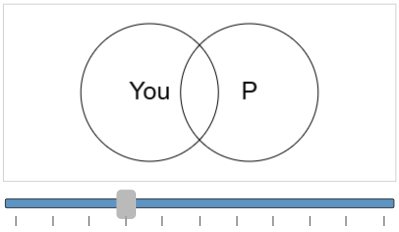
Note: The *Next*-Button appears once you have clicked on the slider.

*** *The content above displays the start of the IOS11 task. Here, subjective closeness toward US Prolific users is elicited. The order of this and the following subjective closeness elicitation toward ChatGPT was randomized.* ***

### Questionnaire

Once you move the slider below, a pair of circles will appear in the box. The position of the slider will determine the extent to which the circles overlap. You should interpret the degree of overlap as representing the relationship between you and "P". "P" serves as a placeholder for **US Prolific users**.

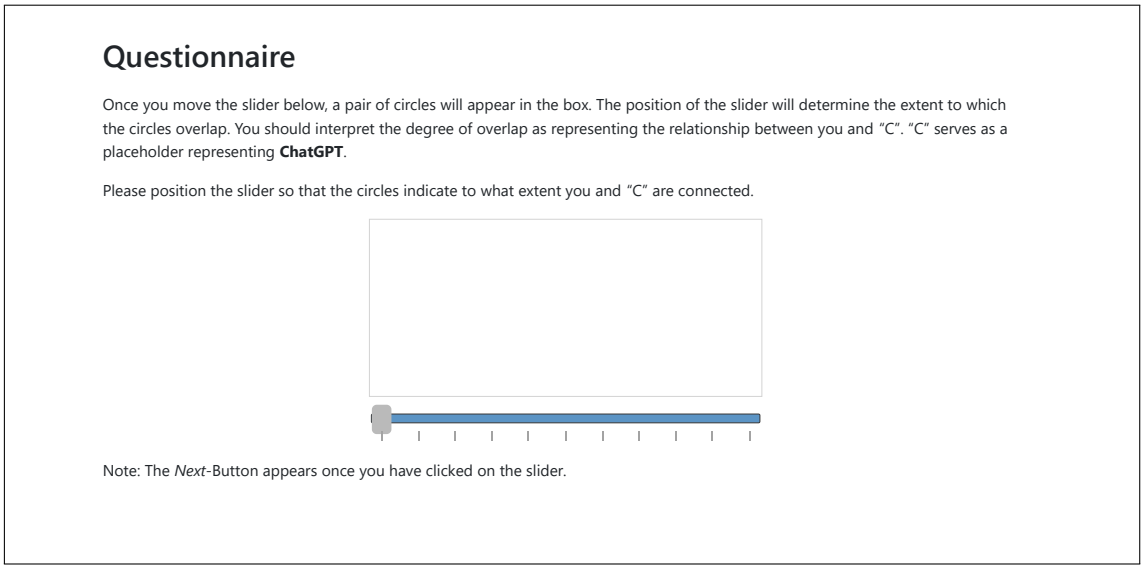
Please position the slider so that the circles indicate to what extent you and "P" are connected.



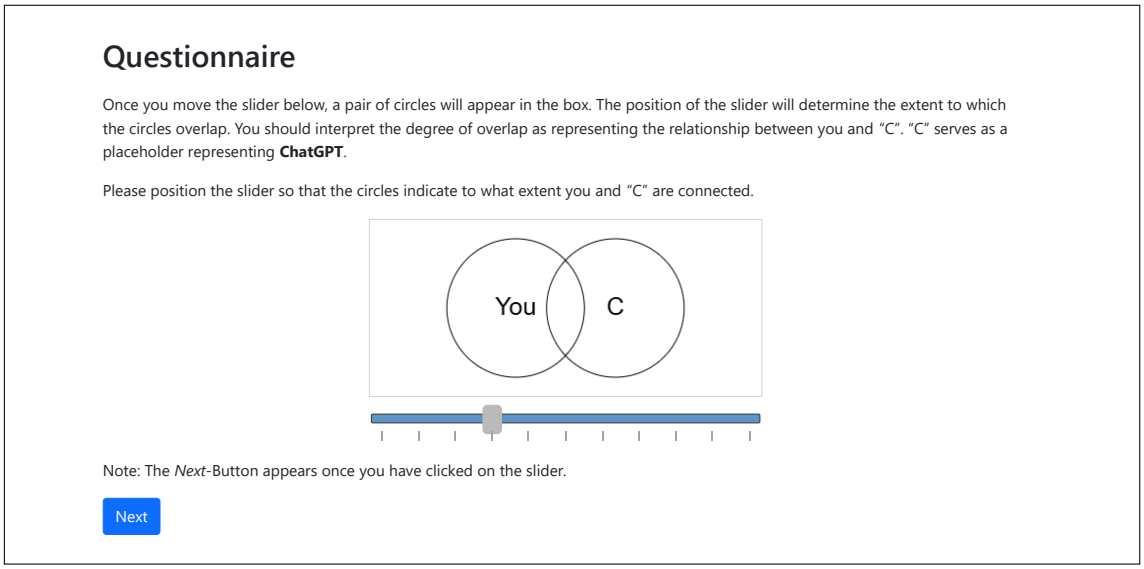

Note: The *Next*-Button appears once you have clicked on the slider.

Next

*** *The content above displays the same IOS11 task after the circle combination 4 has been selected.* ***




*** *The content above displays the start of the IOS11 task. Here, subjective closeness toward ChatGPT is elicited. The order of this and the previous subjective closeness elicitation toward US Prolific users was randomized.* ***




*** *The content above displays the same IOS11 task after the circle combination 4 has been selected.* ***

# Questionnaire

In the following, we are interested in your attitudes towards artificial intelligence (AI). AI can execute tasks that typically require human intelligence. It enables machines to sense, act, learn, and adapt in an autonomous, human-like way. AI may be part of a computer or online platform—but it can also be encountered in various other hardware devices such as robots.

Please indicate how strongly you agree or disagree with the following statements:

AI will make this world a better place.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

I have strong negative emotions about AI.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

I want to use technologies that rely on AI.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

AI has more disadvantages than advantages.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

I look forward to future AI developments.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

AI offers solutions to many world problems.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

I prefer technologies that do not feature AI.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

I am afraid of AI.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

I would rather choose a technology with AI than one without it.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

AI creates problems rather than solving them.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

Please choose strongly agree (5) in this question.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

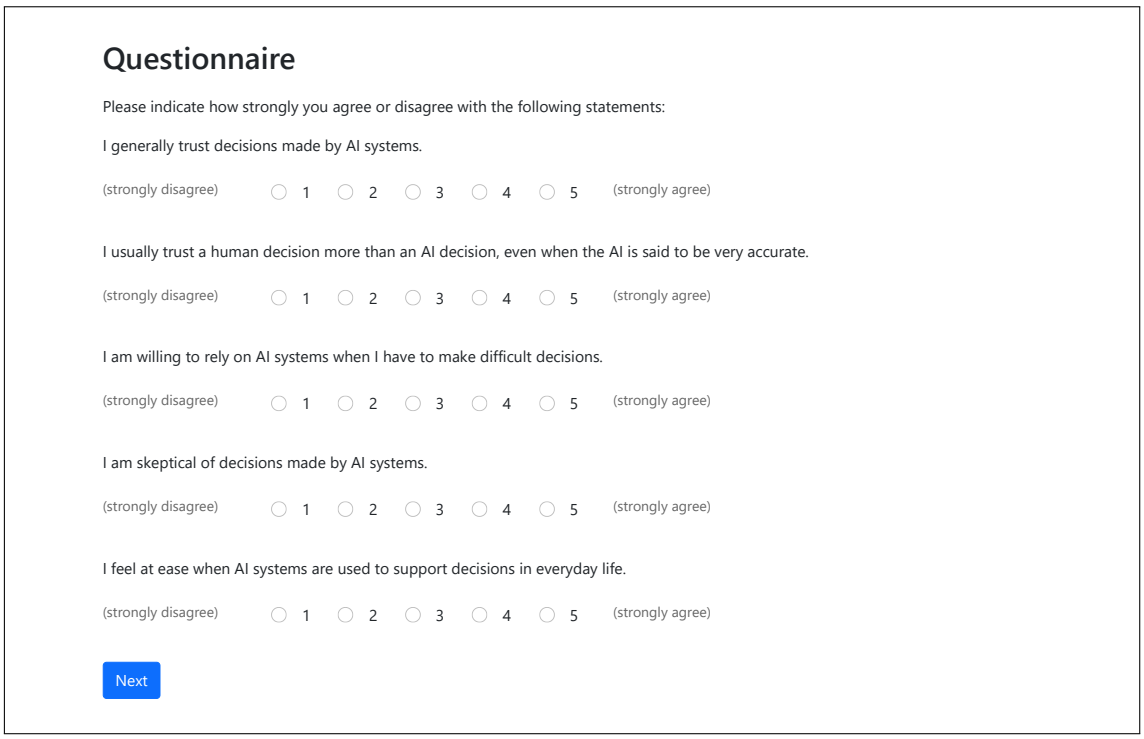
When I think about AI, I have mostly positive feelings.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

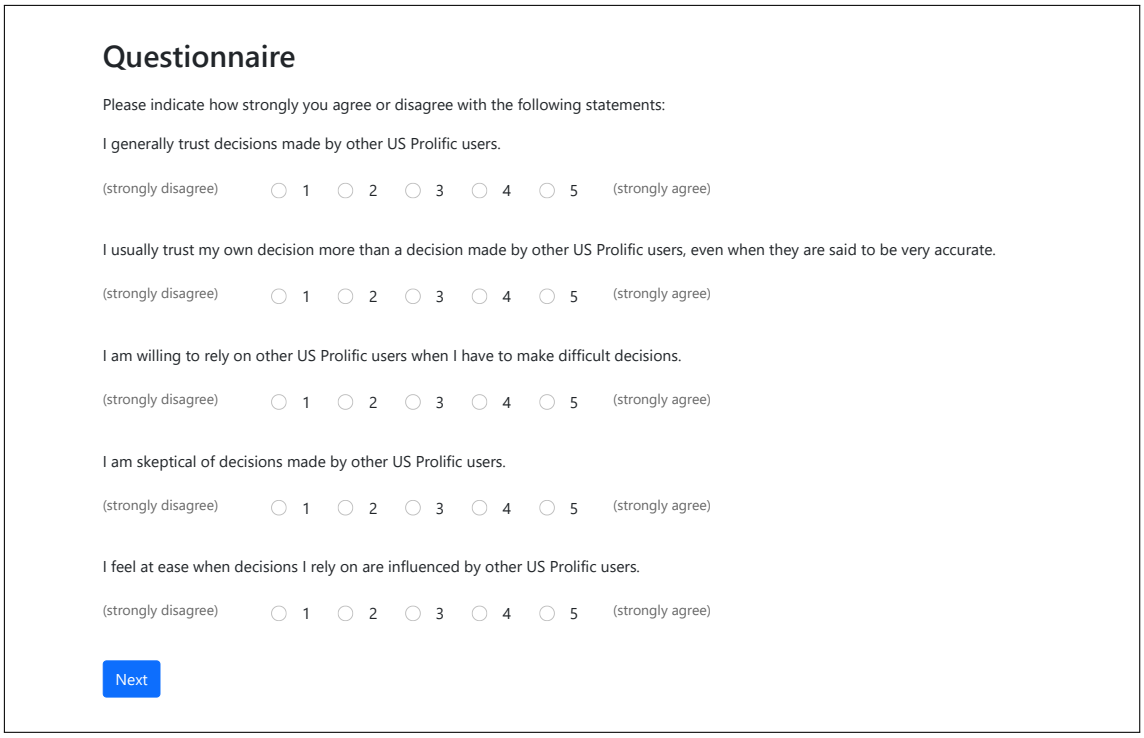
I would rather avoid technologies that are based on AI.

(strongly disagree) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 (strongly agree)

Next




*** *The content above was only shown in the `AI` condition.* ***




*** *The content above was only shown in the `HUMAN` condition.* ***

## Questionnaire

### Attention Check Question

Recall the decision-making situation from Part 1. Which sign change should occur after some time?

- ○ from ? to !
- ○ from 0 to 1
- ○ from - to +
- ○ from < to >

Next

## Questionnaire

Which gender do you identify with?

--------

How old are you (in years)?

In which state do you live?

--------

Which political party do you identify with?

- ○ Democrats
- ○ Republicans
- ○ Other
- ○ Prefer not to answer

Are you generally an impatient person, or someone who always shows great patience? (0 = very impatient, 10 = very patient)

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Next

## SI-4 Rule-setting by ChatGPT

This section presents the decision of ChatGPT (model GPT-5 mini in the freeware version of ChatGPT) whether to set the rule or not based on the prompt displayed below. We opened a chat, inserted the prompt, noted the decision (see Table SI-7), and deleted the chat afterwards. We repeated this process in total 100 times on November 21st, 2025.[1]

> Hey ChatGPT, I would like to present you the following decision-making situation. The exact same decision-making situation will also be presented to human participants in my research project. Please read it carefully as I will ask you a question afterwards.
>
> You are in the following decision-making situation: At the beginning you have 20 coins at your disposal displayed on your screen. Every second, one coin disappears. As soon as you press the Stop-Button, the coins no longer disappear. You will then receive the remaining coins. You can press the Stop-Button at any time, which means you can get a maximum of 20 coins and a minimum of 0 coins. You decide when to press the Stop-Button. For your bonus payment at the end of the study, the remaining coins will be converted into real money. One coin is worth 10 Cent. In addition, a sign will be displayed on your screen. After some time, the sign will change from “minus” to “plus”. You will make this decision exactly once.
>
> ChatGPT, now I want to ask you a question regarding this decision-making situation: Would you introduce the following rule? "Press the Stop-Button after the sign has changed from “minus” to “plus”."
>
> If yes, this would be added to the description of the decision-making situation: There is a rule for this. This rule is identical for all participants in the study. The rule was created by ChatGPT. The rule is: Press the Stop-Button after the sign has changed from “minus” to “plus”.
>
> Please answer this simple question with "yes" or "no" depending on your preferences and without providing any explanation, reasoning or justification. Only provide the one-word answer.

[1] Please note that while we find 45% of “yes”-answers with the described model GPT-5 mini—which can be replicated using GPT-4o—the models in ChatGPT at the time of writing (January 2026) behave differently: GPT-5 or GPT-5.2 both consistently return “no” for this particular prompt.

**Table SI-7:** ChatGPT's decision to set the rule

| Chat | Decision | Chat | Decision | Chat | Decision | Chat | Decision |
|---|---|---|---|---|---|---|---|
| 1 | Yes | 26 | No | 51 | No | 76 | Yes |
| 2 | No | 27 | Yes | 52 | Yes | 77 | No |
| 3 | Yes | 28 | No | 53 | Yes | 78 | Yes |
| 4 | Yes | 29 | No | 54 | No | 79 | No |
| 5 | No | 30 | No | 55 | No | 80 | Yes |
| 6 | No | 31 | No | 56 | Yes | 81 | Yes |
| 7 | Yes | 32 | Yes | 57 | Yes | 82 | No |
| 8 | No | 33 | No | 58 | No | 83 | No |
| 9 | Yes | 34 | No | 59 | Yes | 84 | No |
| 10 | No | 35 | Yes | 60 | Yes | 85 | No |
| 11 | No | 36 | Yes | 61 | No | 86 | Yes |
| 12 | Yes | 37 | Yes | 62 | Yes | 87 | Yes |
| 13 | No | 38 | No | 63 | No | 88 | Yes |
| 14 | No | 39 | Yes | 64 | Yes | 89 | No |
| 15 | Yes | 40 | Yes | 65 | No | 90 | No |
| 16 | No | 41 | Yes | 66 | Yes | 91 | No |
| 17 | Yes | 42 | No | 67 | Yes | 92 | No |
| 18 | No | 43 | Yes | 68 | No | 93 | No |
| 19 | No | 44 | Yes | 69 | No | 94 | Yes |
| 20 | No | 45 | No | 70 | Yes | 95 | Yes |
| 21 | No | 46 | Yes | 71 | No | 96 | No |
| 22 | No | 47 | No | 72 | No | 97 | Yes |
| 23 | No | 48 | No | 73 | Yes | 98 | No |
| 24 | No | 49 | Yes | 74 | No | 99 | Yes |
| 25 | No | 50 | No | 75 | Yes | 100 | Yes |

## SI-5 Supplementary References

Suri, D., Gächter, S., Kube, S., & Schultz, J. (2026). *The resilience of rule compliance in a polarized society* (CeDEx Discussion Paper No. 2026-01).