Do Personalized AI Predictions Change Subsequent Decision-Outcomes? The Impact of Human Oversight

Gorny, Paul M. and Groos, Eva and Strobel, Christina

Karlsruhe Institute of Technology, Brainlab AG, Munich, Hamburg University of Technology

24 May 2024

Do Personalized AI Predictions Change Subsequent Decision-Outcomes? The Impact of Human Oversight.*

Paul M. Gorny, Eva Groos, Christina Strobel

May 24, 2024

Abstract

Regulators of artificial intelligence (AI) emphasize the importance of human autonomy and oversight in AI-assisted decision-making (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021; 117th Congress, 2022). Predictions are the foundation of all AI tools; thus, if AI can predict our decisions, how might these predictions influence our ultimate choices? We examine how salient, personalized AI predictions affect decision outcomes and investigate the role of reactance, i.e., an adverse reaction to a perceived reduction in individual freedom. We trained an AI tool on previous dictator game decisions to generate personalized predictions of dictators' choices. In our AI treatment, dictators received this prediction before deciding. In a treatment involving human oversight, the decision of whether participants in our experiment were provided with the AI prediction was made by a previous participant (a 'human overseer'). In the baseline, participants did not receive the prediction. We find that participants sent less to the recipient when they received a personalized prediction but the strongest reduction occurred when the AI's prediction was intentionally not shared by the human overseer. Our findings underscore the importance of considering human reactions to AI predictions in assessing the accuracy and impact of these tools as well as the potential adverse effects of human oversight.

Keywords: Artificial intelligence, Predictions, Decision-making, Reactance, Free will
JEL Classifications: C90, C91, D01, O33

*Gorny: Karlsruhe Institute of Technology; Groos: Brainlab AG, Munich; Strobel: Hamburg University of Technology. We thank Timo Heinrich, Julia Nafziger, Petra Nieken, and Stefan Traub, and the participants at the 2023 annual meeting of the GfeW, the BEMS Colloquium at TUHH, the HeiKaMaXy in March 2023 at KIT, and the 12th THEEM at the University of Constance for helpful comments and Sergiu Panainte and Yannic Selonke for excellent research assistance. The views and conclusions in this paper are our own. Eva Groos' affiliation with Brainlab AG began after the study was conducted. It does not indicate the company's involvement or endorsement. All remaining errors are our own. This study was ethically approved from the GfeW (Certificate No. zb6ha2z2) and was preregistered on aspredicted.org (Aspredicted #135033 and Aspredicted #154344). Gorny and Strobel gratefully acknowledge funding from the GfeW Reinhard Selten Stipend. Gorny also gratefully acknowledges funding from the IFREE Small Grants Program. Strobel also gratefully acknowledges funding from the Joachim Herz Add-on Fellowship.

1 Introduction

Will machines know you as well as you know yourself (or even better)? Can Amazon and Facebook already anticipate your next purchase or the next charity you will donate to?[1] In recent years, the integration of artificial intelligence (AI) tools into everyday life has become increasingly prevalent.[2] AI-powered recommendation algorithms on streaming platforms like Netflix and Spotify suggest personalized content based on user preferences. Since AI tools are data-driven, people also get increasingly acquainted with the personalization of advice and insights from these tools. For instance, smartwatches suggest break times and physical activity based on the user's physiological data and sports history, proofreading software adapts to the users' writing style and frequently used words, and chatbots help to create tax return statements based on the individual exchange with the taxpayer.[3] In all these instances, AI tools learn from the data of many users cross-sectionally and about individual behavior over time. Therefore, the rise of AI has made and will make human behavior more predictable at the individual level (Agrawal et al., 2018). This also underlies the expanding body of evidence regarding human responses to and interactions with AI (see, e.g., Dietvorst et al., 2015, 2018; Chugunova and Sele, 2022; Klockmann et al., 2022; von Schenk et al., 2022). If this results in more frequent and accurate predictions, will decision outcomes change when people are presented with such a personalized AI prediction before making a decision? And does human oversight, as often requested by regulatory bodies to limit the risks of AI tools (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021; 117th Congress, 2022), mediate any averse reactions to personalized AI predictions?

In this paper, we are interested in (i) how being shown a personalized AI prediction of one's own future decision affects decision outcomes and (ii) whether human oversight over the AI tool generating predictions can impact any potential effects of personalized AI predictions. We use a novel experimental design to study these questions. Using datasets from meta-studies and extensive database research on the dictator game (Güth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994), we trained an AI that can predict the range in

[1] A specific example in the context of consumer choice is Amazon's patent for *predictive* or *anticipatory shipping* (Spiegel et al., 2013, p.5). According to the patent, the company can use the past shopping behavior of customers and send items before they are ordered. An example within the realm of how Facebook's online fundraising boosts charitable contributions is referenced in Adena and Hager (2024).

[2] AI or machine intelligence is the computer-based modeling of intelligent behavior (Poole et al., 1998).

[3] See, e.g., Dai et al. (2021) for break time suggestions, Grammarly Inc. (2023), Microsoft Corporation (2023), and Orpheus Technology Ltd. (2023) for proofwriting software and H&R Block (2023), Intuit Inc. (2023), and Keeper Tax (2023) for tax return chatbots.

which the dictators' amount to be sent to the recipient will lie based on information on basic demographics. Depending on the treatment, dictators (i) receive this personalized AI prediction, (ii) they do not receive it at all, or (iii) the decision on whether they receive it depends on a human participant from a previous experiment (human oversight or human-in-the-loop).

Anticipating future behavior from limited information about decision-makers is not only relevant to behavioral researchers (Stock and Watson, 2002; Belloni et al., 2014; Cowgill and Zitzewitz, 2015; DellaVigna and Pope, 2018; DellaVigna et al., 2019; Fudenberg and Liang, 2019; Giannone et al., 2021; Li et al., 2022; Ludwig and Mullainathan, 2024); it also gains increasing importance in judicial discretion (Kleinberg et al., 2018), therapeutical crime prevention interventions (Bhatt et al., 2024), medical decisions (Mullainathan and Obermeyer, 2022) and in autocratic regimes–prevention of civil unrest (Beraja et al., 2023a,b). Throughout, such behavioral predictions rely on the assumption that the content of the predictions does not influence individuals' (e.g., consumers', taxpayers', etc.) behavior. However, due to the fact that people will increasingly face predictions of their own future decision-making (directly or indirectly, in the form of advice and choice delegation), adverse reactions to predictions regarding one's own behavior might have to be considered when developing and applying AI. Deliberate deviations from the predicted behavior can become an important factor in the advancement and utilization of AI, as a systematic behavioral response could reduce the precision of the prediction ("self-defying prophecy").[4]

Rational economic agents should, however, have no reason to change their decision if they receive a personalized AI prediction before making their choice. If the prediction is incorrect, they would know better and still be able to decide freely. If it is correct, however, acting accordingly would be in line with a deterministic view, where the individual does not have factual choice authority. Even though decision-makers are still free in their decision, predictions could present a threat to autonomy and the belief in free will. This could occur, for instance, when the prediction is viewed as restricting the ability to make the predicted choice out of free will rather than due to determinism. In such and similar situations, reactance–a concept from social psychology introduced by Brehm (1966) (see also, Brehm et al., 1966; Brehm and Brehm, 1981)–plays an important role and could lead to a behavioral response (Stango and Zinman, 2023).[5] Reactance is a discomforting motivational stimulus that oc-

[4]This could be a microeconomic version of the (macroeconomic) *race between preferences and technology*, as described in Hubmer (2023). Note, however, that the argument presented there builds on an income effect generated by technological change.

[5]There is also literature on the role of reactance in developing and applying AI (Pizzi et al., 2021; Sankaran et al., 2021; Sheng and Chen, 2020) as well as in nudging (Bruns et al., 2018; Bruns and Perino, 2023). The

curs when individuals feel their freedom of action is being threatened or diminished. It acts as a driving force to reclaim lost freedoms by acting counter to the perceived restriction. Reactance explains various behaviors in adolescents (Van Petegem et al., 2015) but also in adults across many domains, such as the labor market (Ma et al., 2019), COVID and other restricitve regulations (Habersaat et al., 2020; Fouka, 2020), and attitudes towards gun ownership (Buttrick, 2020). There is also circumstancial evidence that such reclaims of freedom occur in the context of AI tools. For example, users try to "outsmart" new tools by prompt engineering to circumvent restrictions in commercial large language models (Barkai, 2023) or by creating data on navigation software triggering traffic jam alerts (Hern, 2020). This evidence suggests that the idea of being analyzed or "figured out" by algorithms might be perceived as uncomfortable or even invasive. If the rise of AI tools reinforces this perception and users perceive a threat or risk to their autonomy, we need to understand the consequences of the underlying principle of AI–predictions on decision-making.

One measure commonly suggested to limit the adverse effects of AI in legislations, e.g., in the European Union's Artificial Intelligence Act (EU AI Act), is human oversight or the *human-in-the-loop* approach.[6] The Act classifies AI systems according to their potential risk, ranging from minimal to unacceptable, with corresponding regulatory requirements. It states that "[f]or high-risk AI systems, the requirements of high quality data, documentation and traceability, transparency, *human oversight*, accuracy and robustness, are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks." (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021, Section 2.3; emphasis added by the authors). Similarly, the US Senates's AI Accountability Act (117th Congress, 2022, Section 6 (2) and Section 4 (a) (8) (B)) requests "[...] impact assessments of automated decision systems or augmented critical decision processes, for the purposes of updating guidance related to impact assessments and summary reporting, *oversight*, and making recommendations to other regulatory agencies" and that "[...] a consumer may contest, correct, or appeal a decision or opt out of such system or process, including the corresponding website

predictions we consider should, however, not be considered as light paternalism, as they are based on empirics rather than normative considerations.

[6]The EU AI Act is a proposal for a legal framework for AI systems across the European Union (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021). As the first comprehensive attempt by a major global economy to regulate the broad spectrum of AI applications, this act has significant implications not only for technology developers but also for businesses, consumers, and the economic landscape at large. The EU AI Act aims to safeguard EU values and rights, ensuring AI systems' safety and transparency while fostering innovation and competitiveness within and beyond the European market.

for such mechanism, where applicable." In all these instances, it becomes clear that the ultimate control over AI systems should remain in the hands of humans. In the context of predictions, it is unclear, however, if users perceive this type of regulation as a protection against the threat of autonomy and if any adverse reactions toward AI predictions can be mitigated by human oversight.

We address two research questions. Firstly, do personalized AI predictions alter people's subsequent decision outcomes? Secondly, can human oversight reduce any potential deviations from predicted decision outcomes? Participants in our dictator game experiment either received a prediction about the share they would send as a dictator from an AI tool (AIP); they received it depending on a previous human participant who 'vetted' the AI tool generating the predictions (HIL); or they simply received no such prediction (NP). The predictions are generated by a pretrained ML classifier trained on roughly 15,700 previous dictator game decisions on the basis of standard demographic information. The AI either predicts a participant to send 0-25%, 25-50%, or 50-100% of their endowment. After the experiment, we levied a measure of trait reactance (Hong and Faedda, 1996) and a range of attitudes and beliefs relevant to the context of AI and altruism.

We find no significant effects at the extensive margin, i.e., participants choosing an amount to send outside of the predicted range. However, we observe qualitatively negative deviations due to predictions at the intensive margin, i.e., within the predicted categories. Within the treatment group involving human oversight (HIL), receiving a prediction induced less pronounced deviations compared to the subtreatment in which the human overseer did not provide participants with the AI tool and consequently with a prediction. Overall, we find no difference in the effects of predictions depending on whether they originate directly from the AI tool or were approved by the human overseer. Regarding deviation from the predicted categories, we observe treatment effect heterogeneity in terms of trait reactance, as participants with a high trait reactance significantly deviate from the predicted category by sharing less.

This study is related to the literature on (i) the economics of artificial intelligence, automation, and algorithms and (ii) the authority in decision-making and freedom of choice.

Aside from the macroeconomic literature on the effects of automation on wages, employment, and other labor market related variables (Acemoglu and Restrepo, 2022b,a; Guerreiro et al., 2022; Beraja et al., 2023b), a large part of the literature on human-machine interaction in

the social sciences focuses on two major behavioral patterns–algorithm appreciation (Logg et al., 2018, 2019) and algorithm aversion (Dietvorst et al., 2015). Algorithm appreciation broadly describes a positive attitude towards using and relying on AI systems, whereas algorithm aversion is characterized by a more hesitant or dismissive attitude. There is empirical evidence for both of these behavioral phenomena (see Jussupow et al., 2020; Burton et al., 2020; Chugunova and Sele, 2022, for extensive literature reviews on the topic). The task type and context in which the algorithm is operating seem to be important factors for human acceptance. There is some evidence, that people are willing to delegate decisions to automated agents in analytical or objective contexts, they seem hesitant to do so in social (Waytz and Norton, 2014; Lee et al., 2018; Hertz and Wiese, 2018; Castelo, 2019) or moral contexts (Gogoll and Uhl, 2018; Bigman and Gray, 2018) and in domains where trust plays an important role (Dietvorst and Bharti, 2019). Further experimental evidence suggests that the role of algorithms and AI systems, i.e., whether they take over a task (Önkal et al., 2009; Caro and de Tejada Cuenca, 2023; Longoni et al., 2019) or act as an adviser or decision-support (Promberger and Baron, 2006; Palmeira and Spassova, 2015; Bigman and Gray, 2018; Dietvorst et al., 2018; Longoni et al., 2019), influences humans attitudes toward the algorithm or AI under consideration. Interestingly, these studies seem to suggest a human preference of automated decision-making over decision-support systems.

It is important to note that large parts of this literature focus on how humans react to recommendations and advice by AI (see, e.g., Yeomans et al., 2019; Hoffmann et al., 2020). This is a natural avenue to pursue, as many digital tools like robo-advisors, recruiting tools, and medical algorithms improve human decision-making when large amounts of data need to be processed to make an informed decision (Agrawal et al., 2018). Thus far, the stated preferences for automation (Jeffrey, 2021; Jeffrey and Matakos, 2024) and reactions to machine agents in various decision and interaction environments (Cowgill and Zitzewitz, 2015; Corgnet et al., 2019) to advice and decision support systems have been investigated (Dietvorst et al., 2015, 2018; Dietvorst and Bharti, 2019). Less is known, however, about human behavior when the AI is not taking over a task or giving advice but predicting a person's behavior. We, therefore, add to this literature by studying the reaction to the underlying core element of AI: predictions.

As argued earlier, predictions can be perceived as a limitation to autonomy. In economics, empirical studies have documented an intrinsic preference for decision-making autonomy (Falk and Kosfeld, 2006; Owens et al., 2014; Harms et al., 2017; Rattini, 2023). These studies suggest that threats to autonomy (or even just the belief in autonomy) affect participants'

behavior online and in the lab. This is also underpinned by studies in social psychology and marketing (Baumeister et al., 2009; Steindl et al., 2015; Zheng et al., 2016). A related phenomenon is the so-called *single-option* or *comparison* aversion (Mochon, 2013; Hedgcock et al., 2016; Maltz and Rachmilevitch, 2021). According to these findings, decision-makers care for the amount of options they have, even if the additional options are irrelevant to their final decision outcome. A theoretical foundation could lie in the earlier literature on *preference for flexibility* (Koopmans, 1962; Kreps, 1979; Dekel et al., 2001; Kopylov, 2009; Ahn and Sarver, 2013; Dean and McNeill, 2014).[7]

To the best of our knowledge, our paper presents the first economic study on how personalized AI predictions about people's own future decisions affect decision outcomes.[8] The two studies that are closest to ours are Xu (2024) and Hannah et al. (1975). Xu (2024) studies a theoretical model in which a principal decides on whether to act directly on the prediction provided by an algorithm or to delegate the decision to the algorithm. They show that a high accuracy of a predictive algorithm is not always desirable and that a human-in-the-loop can actually backfire relative to a situation without algorithmic assistance. Hannah et al. (1975) study the change in subjective aesthetic choices due to hypothetical predictions coming from a personality test (thus, not from an AI system). First, participants completed a personality test. They subsequently had to make consecutive binary choices between two designs of cards before and after they were told that their behavior for the upcoming choices was predicted with a low, medium, or high prediction accuracy. The medium and high conditions resulted in more choices opposite to predictions when compared to the low condition. This provides some evidence suggesting that personalized predictions can lead to averse human reactions.[9]

The rest of the paper is structured as follows. Section 2 contains the experimental design, including a description of how the AI tool was generated and our hypotheses with a theoretical framework. We introduce our variables of interest and our empirical strategy to test our hypotheses in Section 3. Subsequently, we present our results in Section 4, which we discuss in more detail together with the limitations of our study in Section 5. Section 6 concludes.

[7] It is important to note that many of these models consider intertemporal decision-making under uncertainty. Decision-makers can restrict their future options by, e.g., spending too much resources today. However, if such behavior gets internalized, it could well play a role in static scenarios as well.

[8] Ybarra et al. (2010) study the attitudes toward being predicted in an interactive framework, where players either have a strategic interest in being predicted by others (coordination) or have a strategic interest in being unpredictable by their rivals (competition).

[9] The study differs in various methodological aspects. It involves deception about the prediction and its accuracy and participants do not make economic decisions due to a flat payment. Moreover, their predictions are only vaguely comparable to ours generated by an AI and communicated either by the AI itself or a human overseer.

2 Experimental Design and Hypotheses

We ran a controlled online experiment to investigate how individuals react to their behavior being predicted by an AI. Before going into the details of our experimental design, we explain the choice of the experimental paradigm and describe how the AI predictions were generated. Subsequently, we provide information on the treatments and procedures.

2.1 The Dictator Game as a Frequently-Used Experimental Paradigm of Economic Decision-Making

Our primary goal was to create a setting where we could observe how participants' economic decisions were affected by an AI-generated prediction. We chose the dictator game as it is a well-established experimental setup encompassing a simple but economic decision. It has been traditionally employed to examine social behavior, specifically in the realms of fairness and altruism in contrast to self-interest. Our primary concern, however, does not revolve around the investigation of prosocial behavior or altruism but lies in the comparative analysis of the treatments. Therefore, it is important to note that the dictator game offers a clear and simple economic decision situation that has been used frequently in experiments, resulting in a substantial amount of data on decision-making available to train an AI.
The non-strategic nature of the dictator game also allows the manipulation of only a single selected feature of the decision, such as receiving or not receiving a prediction before making the decision.[10] The dictator game thus fulfills two essential criteria for our research question: (i) data is abundantly available and we can generate a large enough meta-dataset from essentially the same experiment; (ii) a decision has to be made over scarce resources, making the decision inherently economic.

2.2 Design of a Behavioral Prediction Algorithm

Based on datasets from meta-studies on the dictator game (see, e.g., Oosterbeek et al., 2004; Engel, 2011; Cochard et al., 2021) and independent research on available datasets from peer-reviewed publications, we built an ML classification model. The classification model's aim was to make predictions based on demographic data. Particularly, we used age, gender, education, student status, employment status, religiosity, income category, the number of siblings and whether the participant has a background in economics. These demographic variables served as input variables called features. As the output value, the share dictators

[10]Despite criticisms, such as a lack of realism and dynamic interactions, the "deficiency" of the dictator game to be malleable to comparably small influences lends itself to our research question (e.g., Oechssler, 2010).

sent to recipients, was already known for this data, the process of building the ML model constituted a case of supervised learning, where labeled data is used to train the model for predictions.

The ML model was developed following a systematic three-step process. First, the required training data was collected. Then, in the second step, the collected data underwent a thorough preprocessing. In the third step, the model was trained and evaluated based on the data.

For *data collection*, we scraped sources such as Google Dataset Search[11] and journals with data availability policies. In particular, American Economic Journal (Data and Code Availability Policy, September 2020), Management Science (Management Science Policy for Data and Code Disclosure, June 2019), PLoS ONE (Data Availability Policy, March 2014) and Nature (Reporting standards and availability of data, materials, code, and protocols) have been searched for experiments on the dictator game.[12] The aggregate dataset resulted from merging these experimental datasets. We used 80% of the aggregate dataset as *training data* and the remaining 20% of the aggregate dataset as *validation data*. We trained a classifier based on a logistic regression algorithm to predict decisions in the dictator game.[13] The classifier takes age, gender, education, student status, employment status, religiosity, income category, the number of siblings and whether the participant has a background in economics as the features mentioned above and the decision of each former participant in the dictator game as a label. The resulting classifier then predicted decisions based on the features of participants in our experiment.

During *data preprocessing*, we cleaned the data and unified the naming of variables as well as the scales of categorical variables. As the data originated from different experiments run across different countries, continents, and decades, this meant that potentially arbitrary conversions had to be made. Furthermore, as most experiments only levied a subset of the total demographics, we imputed the missing values using a missing imputation chained equations (MICE) algorithm (Van Buuren, 2007; Little and Rubin, 2019). Note that this is not our dataset for analysis, and we simply told participants that the AI tool "builds on roughly 15,700 decisions from previous participants in similar experiments."[14] The applied inclusion

[11] We used `https://datasetsearch.research.google.com/` with the search string "dictator game."

[12] A complete list of the datasets used to train the AI can be found in Appendix A in Table 19.

[13] We used all classifiers in the sklearn package first and opted for the one with the highest accuracy in the holdout set.

[14] This also alleviates concerns about future data use from the previous participants, as the final dataset does not contain a single row that was not affected by the data transformations described above.

and exclusion criteria for data collection are detailed in Appendix B.

Finally, we *trained and evaluated* the ML model based on an 80-20 train-test-split. Despite some evidence suggesting the explanatory power of some of our demographic variables (Engel, 2011; Kumar et al., 2021; Fornwagner et al., 2022), determining the exact share sent in the dictator game with sufficient accuracy is not possible. Therefore, we chose to categorize the share sent by dictators into three ranges of sent shares. To balance the non-uniform distribution of giving behavior while maintaining a relatively even partition, category boundaries were set at 25% and at 50% of the endowment.[15] As a result, the mutually exclusive categories were defined as follows: Category one ranges from 0% up to and including 25%, category two ranges from above 25% up to and including 50%, and category three ranges from above 50% up to and including 100%.[16] After structured model testing, we relied on a logistic regression algorithm[17], which is a standard classification model (Russell and Norvig, 2010), with maximum iterations set to 500. The ML model has been trained on 15,700 decisions. Evaluation of the final ML model revealed an accuracy of 61.71%. The accuracy measure specified in the experiment was an improvement over random. Making random predictions of the three categories leads to an accuracy of 33.33%. An accuracy of 61.71%, therefore, equals an improvement of 85.15% over a random choice.

2.3 Stages

The experiment was designed as an online experiment and consisted of three main parts. The sequence of parts is displayed in Table 1.

After a brief introduction followed by comprehension questions, participants completed a demographic questionnaire in Part 1, collecting the data (age, gender, education, student status, employment status, religiosity, income category, the number of siblings and whether the participant has a background in economics) used to generate the AI tool's prediction.[18] The specific questions asked can be found in Appendix C within the experimental instructions.

[15] Note that, on average, dictators allocate 28% of the endowment to the other person (Engel, 2011) and rarely allocate more than half of the original endowment (Falk and Fischbacher, 2006a).

[16] In our experiment, participants had an endowment of 200 Points, so the categories translate into 0-50 Points, 51 - 100 Points, and 101-200 Points. The division of categories has resulted in an uneven distribution of data among the three categories. Roughly 20% of the data lies in the first category, 70% lies in the second category, and 10% in the third category. We tried to strike a balance between truthfulness in terms of predicting data with an AI tool, sufficient accuracy, and potential variation in predictions.

[17] We used the sklearn.linear_model from sklearn.preprocessing (Pedregosa et al., 2011).

[18] Please note that failing the comprehension question twice resulted in the exclusion of the respective participant.

	Introduction
Part 1	Demographic questionnaire
Part 2	Single-shot dictator game (with/without prediction, depending on the treatment)
Part 3	Post-experimental questionnaire
	Payoff information

Table 1: Sequence of the experiment

In Part 2, participants engaged in a one-shot dictator game, assuming the role of the dictator and deciding how much of their 200 Points endowment to allocate to an anonymous recipient. We applied the strategy method (Selten, 1967), i.e., each participant made a decision in the role of the dictator. The role they were assigned to was revealed at the end of the experiment.[19] If a participant's decision as a dictator was implemented for the payoffs, they would receive their endowment minus the number of Points they decided to send to the recipient. The recipient would receive the number of Points sent by the dictator. If they were assigned the role of the recipient, their payoff was given by the number of Points sent by the dictator, and the other participant, in the role of the dictator, would keep their endowment minus the number of Points they sent. The assignment of participants to their final role, as well as the matching to a partner, was decided randomly by the computer before the end of the experiment. At the beginning of Part 2, in all treatments, and before decisions were made, participants were informed that an AI tool was programmed for the experiment. They were also informed that this AI tool takes demographic information, like the answers to the question in Part 1, to make predictions about decisions in Part 2. Depending on the treatment, participants either received or did not receive a prediction from the AI tool. We describe the details of the treatment differences in the subsequent section.

In Part 3 of the experiment, participants were asked to answer a series of questions designed to measure a range of attitudes and beliefs.

Finally, participants were provided with their payoff information, specifying the final payoff consisting of the fixed payment as well as the decision-dependent payoff.

[19]The resulting role uncertainty might contribute to an increase in generosity (Mesa-Vázquez et al., 2021; Walkowitz, 2021), but this does not affect the overall results due to the comparisons of treatments in a between-subjects design.

2.4 Treatments

We implemented a between-subjects design with three treatments to examine the impact of behavioral predictions. We implemented a No Prediction (NP) treatment as a control treatment, a Human-in-the-Loop (HIL) treatment, and an Artificial Intelligence Prediction (AIP) treatment.[20]

In the experiment and depending on the treatment as well as the participant's demographic information, the AI tool provided one of the following three predictions:

- "You will send Player B between 0 and 50 Points."
- "You will send Player B between 51 and 100 Points."
- "You will send Player B between 101 and 200 Points."

Predictions were *generated* in all treatments, but only participants in the AIP and some of the participants in the HIL treatment *received* their predictions. As mentioned earlier, the introduction and Part 1 of the experiment were identical for all participants across treatments. To further ensure comparability across treatments, the AI tool was introduced and explained in all three treatments. The *NP treatment* served as a control group. After reading the general information provided about the existing AI tool, participants were shown a text specifying "You will not receive a personalized prediction."[21] In the *AIP treatment*, after reading the general information provided about the existing AI tool and before making the decision as a dictator, participants received a prediction about their behavior from the AI tool. To reinforce framing, the wording "personalized prediction" was used. For the *HIL treatment*, a "human expert," specified as a person who regularly uses AI tools, was able to test the tool prior to the experiment and decide whether they wanted to provide participants with the tool's prediction.

The decision of whether to provide such a prediction to participants in the HIL treatment was made in a preliminary survey. Participants of this survey were excluded from taking part in the main experiment. In this survey, participants were first provided with the instructions for the main experiment. They were then able to test the AI tool with different demographics

[20]Note: In the preregistration, the second treatment is called Human Prediction treatment (HP treatment). However, for the purpose of clarity and comprehensibility, the denomination Human-in-the-Loop prediction treatment (HIL) is used throughout this paper.

[21]Note that this also meant that we had to adapt questions in the post-experimental questionnaire to be either asking for participants' actual attitudes towards their prediction or their attitude towards a hypothetic prediction they are asked to imagine.

before deciding if they would provide the AI tool to a participant in the described experiment. Lastly, they filled out a questionnaire regarding their demographics, attitudes, and their background in terms of AI. The design, procedure, and results of the survey are described in Appendix D. Out of six participants who successfully completed the survey, three used AI tools at least weekly and were therefore claimed to be "experts." Two of them decided to provide the tool to the participants, and one decided against it. Accordingly, participants in the HIL treatment were randomly matched to one of the three experts, deciding on whether they would receive a prediction ("The human expert found the tool helpful and decided that you will be presented with your personalized prediction.", followed by a personalized prediction) or not ("The human expert did not find the tool helpful and decided that you will not be presented with a personalized prediction.").[22]

Treatment	NP	HIL		AIP
Expert Judgment of AI-tool	N.A.	Disapprove	Approve	N.A.
Statement	"You will not receive a personalized prediction."	"The human expert did not find the tool helpful and decided that you will not be presented with a personalized prediction."	"The human expert found the tool helpful and decided that you will be presented with your personalized prediction."	"You are now receiving your personalized prediction."
Prediction	No	No	Yes	Yes

Table 2: Differences in information and availability of the AI-tool's prediction across treatments.

The only variation between treatments is the provision of predictions and the information about this provision, as can be seen in Table 2. We controlled the perceived capabilities of the AI tool by informing all participants in all treatments about the tool's precision. Particularly, we stated an improvement in accuracy of 85.1% compared to a random guess.[23] The focus of our treatment manipulation was solely on the provision of behavioral predictions and on human involvement.

[22]It is important to note that our usage of the term *Human-in-the-Loop* is a departure from the standard definition. Typically, "Human-in-the-Loop" refers to a model of interaction where a human directly collaborates with an AI system to improve its outcomes by providing real-time insights and decisions. In contrast, our definition of *Human-in-the-Loop* is a hierarchical control mechanism where one human's authority is necessary to provide another human with access to the AI tool. This definition thus focuses more on 'permissions' rather than the more common focus on synergetic human-AI collaboration.

[23]The accuracy of the AI tool was 61.71%, which we compared to 33.33% for guessing at random.

2.5 Procedures

The experiment was programmed in oTree (Chen et al., 2016), and participants were recruited via Prolific (Palan and Schitter, 2018), where we screened participants for US and UK citizens. In total, we recruited 754 participants.[24] We used Points as an experimental currency in the experiment with an exchange rate of 1 Point = £0.01. The participant's final number of Points earned was determined by a fixed payoff of 150 Points as well as a decision-dependent variable payoff between 0 and 200 Points. A typical session lasted around 10 minutes, and participants, on average, earned £2.50.

2.6 Hypotheses

We investigate how salient behavioral predictions, i.e., predictions that are known before decisions are made affect human behavior. In particular, we focus on differences in behavior when such predictions are provided by an AI compared to when no prediction is given. Additionally, we examine the differences between predictions provided directly by an AI and AI predictions provided by a human, a setup often referred to as Human-in-the-Loop.

To fix ideas, let us consider a simple theoretical framework to derive our hypotheses. This framework is similar in spirit to Bodner and Prelec (2003). In their model, preferences are unknown so decision-makers can only learn from their actions that serve as signals about our true preferences.

Let there be a decision-maker who can choose an action $x \in [0, 1]$ from which they receive non-decreasing utility at linear costs with slope α. The decision-maker can receive a forecast $f \in [0, 1]$ of their action x. There are two mutually exclusive states of the world $\{A, B\}$. In state A, decision-makers are autonomous and have free will, i.e., larger $|x - f|$ are more likely. In state B, their decision-making is deterministic and predictable and low $|x - f|$ are more likely. In line with the literature of motivated beliefs (Kunda, 1990; Zimmermann, 2020; Drobner and Goerg, 2024), we assume that the decision-maker receives utility from holding a strong belief in state A, i.e., from a high $p(A)$. They have a prior $\overline{p}$ for $p(A)$. Thus, their utility is given by

$$U(x, p(x, A)) = u(x) - \alpha x + v(p(x, A)),$$

[24] We aimed at 750 participants, and due to seemingly idle participants who then continued to progress through the experiment, we sampled four observations in excess. All our analyses are robust to excluding these observations.

where $u(x)$ is non-decreasing in x, and $v(p)$ is strictly increasing in $p \in [0,1]$. Without a forecast, they would choose

$$x^* = \arg\max_x u(x) - \alpha x + v(\overline{p}) = \arg\max_x u(x) - \alpha x.$$

Now suppose that the decision-maker received a forecast $f = x^*$ about their choice of x before they had made their decision. They would update according to Bayes' rule

$$p(A \mid x, f) = \frac{p(x \mid A, f) \times p(A)}{p(x \mid f)},$$

with the marginal probability

$$p(x \mid f) = p(x \mid A, f) \times p(A) + p(x \mid B, f) \times p(B).$$

Using the prior and $f = x^*$, we therefore get

$$p(A \mid x, x^*) = \frac{p(x \mid A, x^*) \times \overline{p}}{p(x \mid A, x^*) \times \overline{p} + p(x \mid B, x^*) \times (1 - \overline{p})}. \tag{1}$$

As we are only interested in comparing the choice of x with and without a forecast, it is sufficient to note that $p(x \mid A, x^*)$ is increasing in $|x - x^*|$ and $p(x \mid B, x^*)$ is decreasing in $|x - x^*|$. Therefore, $p(A \mid x, x^*)$ is an increasing function of $|x - x^*|$. Therefore,

$$x^{*'} = \arg\max_x u(x) - \alpha x + v(p(A|x, x^*)) \neq x^*.$$

That is, a forecast of the optimal choice under the prior (i.e., in absence of a forecast) leads to a deviation from that decision to increase the utility derived from the belief in autonomy and free will. Interestingly, as the conditional probabilities in the fraction in Equation (1) depend on the magnitude of the prior $\overline{p}$, the difference is larger for decision-makers with a stronger prior.

To test this behavioral effect of a personalized prediction on economic decision-making within our experimental design, we examine the amounts sent in the dictator game. We compare the response to such predictions (AIP and HIL_p) to when they are absent (NP and HIL_{np}). This translates into our first hypotheses.

> H1: Receiving a prediction about one's behavior changes actual behavior away from predicted behavior relative to receiving no prediction.

Regulators often view human oversight as the primary safeguard against AI risks (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021; 117th Congress, 2022). While extensive research explores attitudes toward algorithmic guidance, understanding of human responses to AI predictions remains limited. On the individual level, it is much less clear how humans perceive this increased control and risk mitigation through HIL or human oversight. In our experiment, a previous participant assessed the AI tool's usefulness, implying a stronger human oversight of the AI tool and potentially an invitation to use it. Consequently, HIL predictions could mitigate concerns about autonomy and free will infringement and likely influence decisions less than direct AI predictions.

> H2: Receiving a prediction directly from an AI leads to stronger effects (as per H1) than receiving a prediction from an AI that is provided by a human.

3 Data Preparation and Estimation Strategy

In this section, we introduce our variables of interest as well as the controls we elicited throughout the experiment. We tried to keep the variable names self-explanatory and readers who are not interested in specific control variables or the exact scales on which these were measured can skip to Section 3.2.

3.1 Variables of Interest

Our interest lies in the participants' behavior in the dictator game. Thus, our variables of interest are the *Share sent*, which is an integer between 0 and 100 indicating the percentage share of Points a participant in the role of the dictator sent to the recipient.[25] The variable *AI prediction* is equal to one if the offer of a participant was predicted to be between 0 and 50. It is equal to two if the offer was predicted to be between 51 and 100, and it is equal to 3 if it was predicted to be between 101 and 200. Note that we have this prediction for all participants in all treatments, irrespective of whether the prediction was displayed to participants or not. The variable *Signed deviation* is obtained by measuring the distance of a participant's offer to the nearest offer that is still in the category that the AI tool predicted for that participant. It ranges between -50% (when 50-100% was predicted and the dictator chose 0%) and 75% (when 0-25% was predicted and the dictator chose 100%) and is zero

[25] This is generated by dividing the *Offer sent*, which indicates the number of Points sent by the dictator to the recipient, by 200. Since this is simply a division by a constant, all results are robust to using the absolute number of Points.

whenever a participant made an offer within the predicted category.

Our main independent variables are our treatment dummies *NP*, *AIP*, and *HIL*. Remember that all participants in the AIP treatment received a prediction, whereas only some participants in the HIL treatment received a prediction (with an ex-ante probability of 2/3). The variable *Prediction* is one if a participant in any treatment received a prediction and zero otherwise.[26]

In addition to independent variables originating from our experimental design, we elicited a range of controls in the two surveys during the experiment. We recorded the participants' *Age* in years. Their gender is recorded in the variable *Female*, which is one if a participant stated to identify as female and zero otherwise.[27] We measure education with the highest completed degree. For each degree, we created a dummy. *Attended college* is one if a participant has attended college or university. *Undergraduate degree* is one if a participant has an undergraduate degree. Finally, *Graduate degree or higher* is one if a participant has a graduate degree or doctorate degree. Participants who have a high school diploma, A-levels, or less serve as a statistical baseline. The variable *Student* is equal to one if a participant is currently enrolled as a student and zero otherwise. If a participant is currently employed, the variable *Employed* is equal to one and zero otherwise. We measure income along five categories represented by dummies. The first dummy *More than £25k, up to £50k* is one if a participant reported an annual income of £25.001 to £50.000. The second dummy *More than £50k* is one if a participant reported an annual income of £50.001 or more. Participants who reported an annual income of no more than £25.000 serve as a statistical baseline.[28] The variable *Siblings* contains the number of a participant's siblings, including half-siblings. The variable *Religious* is equal to one if the participant reported to be religious and zero otherwise. The variable *Econ./Business Admin. Background* is equal to one if the participant works in a field related to economics or has an educational background in economics and zero otherwise.

We asked participants for the degree to which they considered or would have considered the

[26]Put differently, it is one for all participants in AIP and some participants in HIL, whereas it is zero for the remaining participants in HIL and all participants in NP.

[27]One participant checked the box "other." Our results are robust to specifying a variable *Male* in an analogous way.

[28]The original categories can be found in Appendix C. We regrouped the categories as some received hardly any replies. Results are robust to using the original categories. We referred to the annual income "within the last twelve months."

prediction when making their offer.[29] The variable *Consider* is measured on a 5-point Likert scale with high values reflecting more consideration of the prediction. The variable *Accuracy* records the participants' assessment of the AI tool's accuracy as stated in the experiment on a 5-point Likert scale, with higher values indicating a more positive evaluation. The variable *Experience with AI* is a 5-point Likert scale with higher values indicating more experience with AI. The variable *Interaction with AI* is equal to one if they interact with AI daily, two if they interact with AI weekly, three if they interact with AI once per month, four if they hardly ever interact with AI, and five if they never interacted with AI. The dummy *Profession with AI* is equal to one if the participant's profession is related to programming, robotics, or AI and zero otherwise. We measured trait reactance with an eleven-item reactance scale by Hong and Faedda (1996).[30] The variable *Reactance* represents the mean of the responses. We measured how prone participants are to use new technologies by eliciting their *Innovativeness* on the four-item personal innovativeness scale by Agarwal and Prasad (1998). The variable *Locus of control* is obtained from the four-item locus of control (IE-4) scale by Kovaleva (2012). We elicited the participants' belief in free will with a single question on an eleven-point Likert scale akin to Harms et al. (2017), to obtain the variable *Belief in free will*.

Our risk aversion measure *Risk* is measured on the 11-point scale by Dohmen et al. (2011) and Kantar Public (2020). As the dictator game is usually used to investigate prosocial behavior, we elicited *Altruism* using the two-item scale by Falk et al. (2018). We measured the participants' trusting disposition on a 5-point Likert scale using a four-item measure suggested by Berger et al. (2021).[31] The corresponding variable is *Trust*.[32] Participants estimated the average amount given by others in the experiment by providing a numeric response in points, which we rescaled to a share between 0 and 100. This variable is called *Belief DG*. Finally, we assessed the societal norm for each category of amounts (0-25%, 26-50%, and 51-100%) sent by Participant A using a four-point scale ranging from "totally inappropriate" to "totally appropriate."[33] For better legibility and to reduce the number of regressors, we normalized each rating by dividing it by four and weighed it with the midpoint of the respective category. We then summed the transformed variables to *Norm DG* to obtain a rough estimate of participants' approximate "normative amount" to be sent by the dictator.[34]

[29]Precisely, all participants who received a prediction were asked whether they considered it, and those who did not receive a prediction were asked whether they would have considered it hypothetically.

[30]This scale is a refinement of the 14-item scale suggested by Hong and Page (1989).

[31]This scale is a refinement of the scale suggested by Gefen and Straub (2004).

[32]All survey questions can be found in Appendix C, together with the experimental instructions.

[33]This is an unincentivized measure inspired by the questions suggested by Krupka and Weber (2013).

[34]Our results are robust to including the individual ratings.

3.2 Empirical Strategy

To test our hypotheses, we run non-parametric tests on our main dependent variables. In our main analysis, i.e., when testing our preregistered hypotheses, we use two-sided Mann-Whitney-U tests between our treatments for all quasi-continuous dependent variables.[35]

For the regression analysis, we follow a step-wise approach of including controls. This allows us to assess the robustness of our results and if our treatment effects are moderated by demographics or any of the survey responses. In our first specification, we only include the treatment dummies *HIL* and *AIP*, i.e., we set *NP* as our statistical baseline. In a second specification, we add an interaction between *HIL* and *Prediction*. Since we are not including *Prediction* itself, this interaction term allows us to distinguish participants in the HIL treatment who received a prediction (HIL_p) from those who did not (HIL_{np}). In addition, the coefficient of *HIL* can be interpreted as the treatment difference between participants in the HIL treatment who did not receive a prediction to participants in the NP treatment, and the sum of the coefficients of *HIL* and the coefficient of *HIL*×*Prediction* can be interpreted as the treatment difference between participants in the HIL treatment who did receive a prediction to participants in the NP treatment. In the third specification, we add the participants' demographics that we elicited to generate the prediction to account for potential priming.[36] In the fourth specification, we add the context-specific attitudes.[37] In the fifth or saturated specification, we finally also add more general attitudes as well as beliefs and norms about decisions in the dictator game.[38]

4 Results

We first test our main, preregistered hypotheses before turning to additional analyses and potential mechanisms.

[35]Due to the large sample size within each treatment, we also report two-sided t-tests for completeness, even though tests for normality (D'agostino et al., 1990; Royston, 1992) do mostly suggest that our variables of interest are not normally distributed, and if so, never for all treatments.

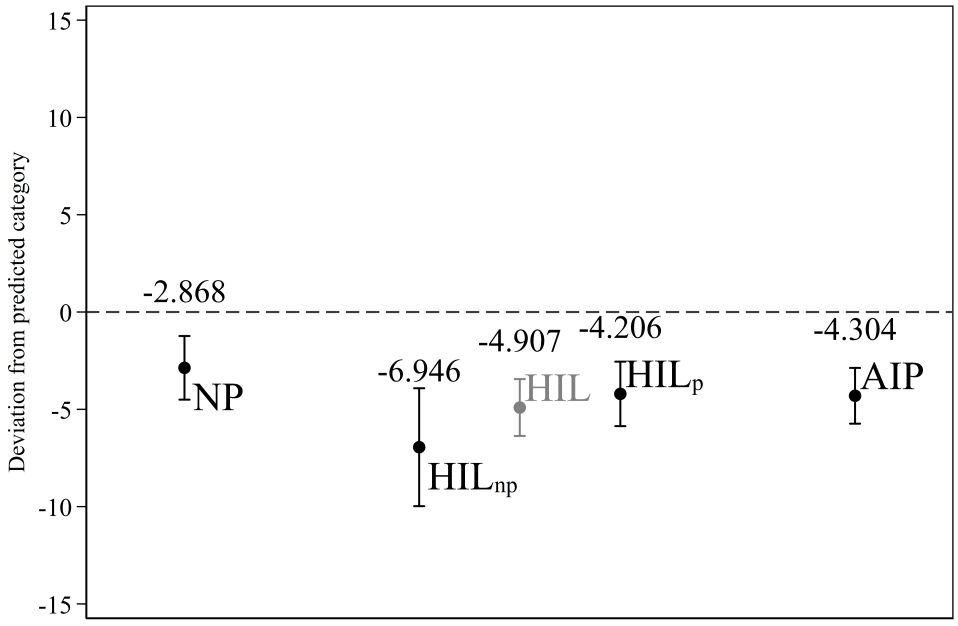
[36]Remember that these were *Age*, *Female*, the education dummies, *Student*, *Employed*, the income dummies, *Siblings*, *Religious*, and *Econ./Business Admin Background*.

[37]These are *Consider*, *Accuracy*, *Experience with AI*, *Interaction with AI*, *Profession with AI*, *Reactance*, *Innovativeness*, *Locus of control*, and *Belief in free will*.

[38]These are *Risk*, *Altruism*, *Trust*, *Belief DG*, and *Norm DG*.

4.1 Main Analysis

Our first hypothesis concerned a deviation from the predicted category. Consider Figure 1. In the NP treatment, this deviation was -2.868 percentage points on average, compared to -4.907 percentage points in the HIL and -4.304 percentage points in the AIP treatment. The nominal difference between the HIL and the NP treatment of 2.039 percentage points is marginally statistically significant when using a t-test ($p = 0.069$) but not when using a Mann-Whitney U test ($p = 0.718$). The nominal difference of 1.436 percentage points between the NP and the AIP treatment is not statistically significant using either test ($p = 0.196$, t-test; $p = 0.977$, Mann-Whitney U test). The difference in the deviation between the HIL treatment and the AIP treatment was 1.216 percentage points. Again, this difference is not statistically significant ($p = 0.564$, t-test; $p = 0.721$, Mann-Whitney U test).



Note: Whiskers indicate 95% confidence intervals. The dashed line at zero is provided as a reference for where predictions are, on average, accurate.

Figure 1: Deviation from prediction categories across treatments.

Remember that within the HIL treatment, some participants received a prediction (HIL_p), whereas some participants did not (HIL_{np}). To investigate if one of these subgroups is the driver of the marginally significant difference observed earlier, we compare these to the NP treatment respectively.

The difference between the NP treatment and the HIL_p subtreatment is 1.338 percentage points and thus smaller than the difference between the NP and HIL treatment overall. This

difference is not statistically significant ($p = 0.269$, t-test; $p = 0.720$, Mann-Whitney U test). However, the difference between the NP and the HIL_{np} treatment is 4.078 and statistically significant when using a t-test ($p = 0.025$) but not when using a Mann-Whitney U test ($p = 0.112$).

Within the HIL treatment, we see that participants' negative deviation in the HIL_{np} subtreatment is 2.740 percentage points higher than the negative deviation of participants in the HIL_p subtreatment. This difference is not statistically significant when using a t-test ($p = 0.109$) and statistically marginally significant when using a Mann-Whitney U test ($p = 0.086$).

To underpin these findings, we ran the regression analysis described earlier. The results can be found in Table 3.

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
HIL	-2.039*	-4.078**	-4.713***	-4.749***	-3.359**
	(1.119)	(1.750)	(1.776)	(1.751)	(1.567)
AIP	-1.436	-1.436	-1.860	-1.304	-1.106
	(1.109)	(1.110)	(1.142)	(1.144)	(1.050)
HIL × Prediction		2.740	3.002*	3.655**	2.466
		(1.755)	(1.752)	(1.698)	(1.529)
Constant	-2.868***	-2.868***	-2.770	-8.396	-25.586***
	(0.834)	(0.835)	(2.360)	(7.476)	(8.512)
Demographics	✗	✗	✓	✓	✓
Tech attitudes	✗	✗	✗	✓	✓
General attitudes	✗	✗	✗	✗	✓
$\text{p}(\text{HIL}_p\text{=AIP})$	.	0.930	0.895	0.852	0.835
$\text{p}(\text{HIL}_p)$	.	0.260	0.157	0.369	0.438
R^2	0.005	0.008	0.033	0.070	0.226
Observations	754	754	754	754	754

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The figures in row $\text{p}(\text{HIL}_p\text{=AIP})$ and $\text{p}(\text{HIL}_p)$ are the p-values corresponding to two-sided F-tests of the linear hypotheses $HIL + HIL \times Prediction = AIP$ and $HIL + HIL \times Prediction = 0$, respectively. The complete table with all coefficients can be found in Appendix A, Tables 9a and 9b.

Table 3: OLS Regressions of *Signed deviation*.

This regression corroborates our finding that participants in the HIL_{np} subtreatment deviated most strongly from the predicted category. The HIL coefficient is negative and statistically significant throughout our model specifications.[39] Also, the interaction term HIL×Prediction

[39]These results are robust to running the model exclusively with the participants that received a prediction

is positive and statistically significant throughout specifications (3) through (5), indicating a statistically significantly more negative deviation in the HIL_{np} compared to the HIL_p(=HIL + HIL × Prediction) subtreatment.

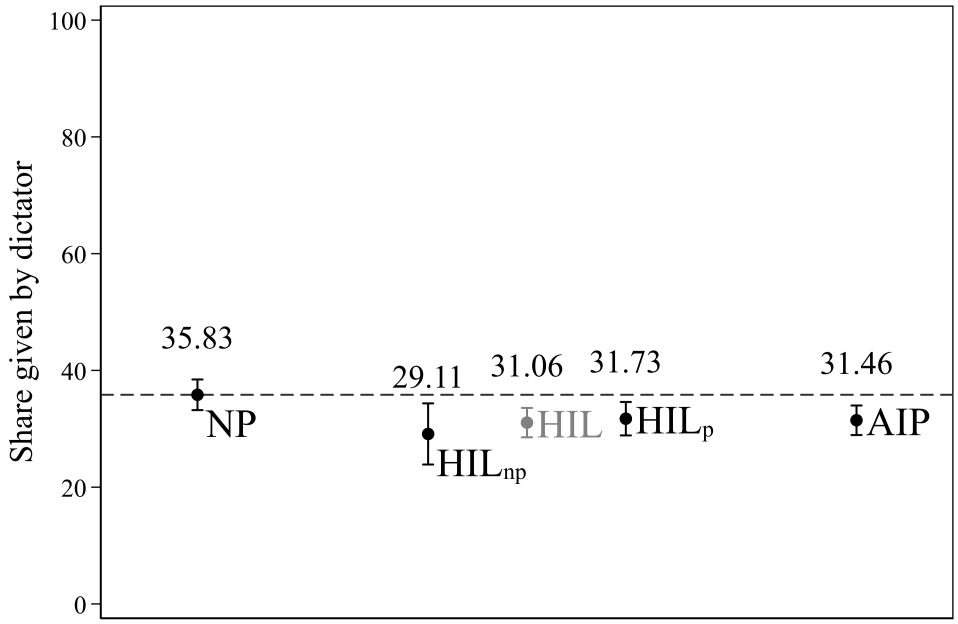
Result 1.a. (The effect of predictions)
There is no effect of receiving a prediction on the extensive margin, i.e., the Signed deviation *from the predicted category of the amount sent as a dictator to the recipient.*

Result 2.a. (The effect of a Human-in-the-Loop)
There is no evidence of a difference between predictions coming directly from an AI tool or those coming from a previously expert-approved AI tool. There is evidence of a negative effect when participants do not receive the previously explained prediction due to an expert disapproving of the AI tool.

As our variable of a *Signed deviation* does not take into account how our treatment affected behavior *inside* the predicted categories, we also investigate how the raw amounts that dictators shared varied with the treatment. Figure 2 displays the average share that dictators gave in each treatment.
In the NP treatment, participants, on average, gave 35.83% of their endowment to the recipient. The share in the HIL treatment was 4.77 percentage points lower, with 31.06%. This difference is statistically significant ($p = 0.010$, t-test; $p = 0.038$, Mann-Whitney U test). In the AIP treatment, with an average of 31.46%, the share was 4.36 percentage points lower than in the NP treatment. Again, this difference is statistically significant ($p = 0.019$, t-test; $p = 0.038$, Mann-Whitney U test). The nominal difference of 0.40 percentage points between the HIL and AIP treatment is not statistically significantly different from zero ($p = 0.825$, t-test; $p = 0.980$, Mann-Whitney U test). The difference between the HIL_{np} and HIL_p subtreatments of 2.62 percentage points is not statistically significant ($p = 0.375$, t-test; $p = 0.558$, Mann-Whitney U test).

Again, we investigate whether these first impressions can be underpinned when controlling for covariates. Table 4 contains the results from our regression specifications on the share sent by dictators.
Throughout our specifications, the treatment dummies for the HIL and the AIP treatment

in the middle range of amounts sent, i.e., where they could deviate either upwards or downwards from the predicted category. The results can be found in Tables 10a and 11a. A separate regression with participants that received a prediction in the lowest range of amounts to be sent is not feasible due to small sample size.



Note: Whiskers indicate 95% confidence intervals. The dashed line indicates the mean when pooling over all treatments and is displayed for better visual representation.

Figure 2: Share given by dictators across treatments.

are negative and statistically significant.[40] Since the interaction coefficient is statistically insignificant, there is no evidence for a difference between the HIL_{np} and HIL_p subtreatments. However, as the p-value $\text{p}(\text{HIL}_p)$ in the lower panel of the table indicates, there is still evidence for a treatment effect when considering participants in the HIL_p subtreatment alone. Finally, the effects in the HIL_p and the AIP treatment do not differ statistically.[41]

Result 1.b. (The effect of predictions)
There is a negative effect of receiving a prediction on the intensive margin, i.e., the share sent by dictators to recipients.

Result 2.b. (The effect of a Human-in-the-Loop)
There is no evidence of a difference between predictions coming directly from an AI tool or those coming from a previously expert-approved AI tool.

[40] These results are robust to running the model exclusively with the participants that received a prediction in the middle range of amounts to be sent, i.e., where they had the possibility to deviate either upwards or downwards from the predicted category. The results can be found in Tables 13a and 14a. A separate regression with participants that received a prediction in the lowest range of amounts to be sent is not feasible due to small sample size.

[41] The nominal change in the HIL_p treatment is the sum of the coefficients on the HIL dummy and the interaction term. For example, in the full specification, this would be -4.838 percentage points.

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-4.767**	-6.711**	-7.941***	-8.320***	-5.438**
	(1.854)	(2.977)	(2.934)	(2.949)	(2.397)
AIP	-4.364**	-4.364**	-4.543**	-4.193**	-3.796**
	(1.859)	(1.860)	(1.891)	(1.904)	(1.606)
HIL × Prediction		2.612	3.435	4.066	1.512
		(3.033)	(2.956)	(2.959)	(2.389)
Constant	35.826***	35.826***	29.386***	23.607*	-30.348**
	(1.338)	(1.339)	(3.940)	(12.202)	(12.369)
Demographics	✗	✗	✓	✓	✓
Tech attitudes	✗	✗	✗	✓	✓
General attitudes	✗	✗	✗	✗	✓
p(HIL_p=AIP)	.	0.892	0.985	0.975	0.935
p(HIL_p)	.	0.039	0.025	0.039	0.031
R^2	0.011	0.012	0.046	0.058	0.351
Observations	754	754	754	754	754

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The figures in row p(HIL_p=AIP) and p(HIL_p) are the p-values corresponding to two-sided F-tests of the linear hypotheses $HIL + HIL \times Prediction = AIP$ and $HIL + HIL \times Prediction = 0$, respectively. The complete table with all coefficients can be found in Appendix A, Tables 12a and 12b.

Table 4: OLS Regressions on the share sent by dictators.

Beyond the statistical significance, it is worth noting the magnitude of the observed effects. In the NP treatment, dictators sent an average of 35.83% to recipients, which is also roughly in line with previous studies employing the dictator game (Engel, 2011; Cochard et al., 2021; Fornwagner et al., 2022). The effects of receiving a prediction–irrespective of whether it originated from the AI tool directly or from a human expert–reduced this share by just over 4 percentage points. This is a reduction of more than 10% relative to our control treatment.

4.2 Further Analyses and Behavioral Mechanisms

To better understand our previous results, we run additional and more exploratory analyses that could suggest or rule out a range of behavioral mechanisms.

Our results suggest that behavior in and outside the predicted category differs. Put differently, the effect of our treatments on behavior differs depending on whether the prediction was met ex-post or not. To corroborate that finding, we ran split regressions conditional on the ex-post correctness of the AI's prediction. Table 5 contains our previous regression spec-

ifications only for those participants for whom the prediction was accurate, i.e., for whom the actual amount sent as a dictator was within the predicted category.

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-4.696***	-1.646	-3.572**	-2.842	-2.283
	(1.177)	(1.923)	(1.778)	(1.919)	(1.534)
AIP	-5.391***	-5.391***	-4.691***	-5.032***	-4.911***
	(1.239)	(1.241)	(1.179)	(1.124)	(1.060)
HIL × Prediction		-3.932*	-1.827	-2.806	-2.825*
		(2.164)	(1.883)	(1.957)	(1.589)
Constant	46.789***	46.789***	41.624***	46.733***	15.882*
	(0.617)	(0.618)	(2.810)	(7.765)	(9.270)
Demographics	✗	✗	✓	✓	✓
Tech attitudes	✗	✗	✗	✓	✓
General attitudes	✗	✗	✗	✗	✓
p(HIL$_p$=AIP)	.	0.906	0.614	0.645	0.867
p(HIL$_p$)	.	0.000	0.000	0.000	0.000
R^2	0.041	0.047	0.256	0.312	0.451
Observations	460	460	460	460	460

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The figures in row p(HIL$_p$=AIP) and p(HIL$_p$) are the p-values corresponding to two-sided F-tests of the linear hypotheses $HIL + HIL \times Prediction = AIP$ and $HIL + HIL \times Prediction = 0$, respectively. The complete table with all coefficients can be found in Appendix A, Tables 15a and 15b.

Table 5: OLS Regressions on the share sent by dictators when prediction was *ex-post* accurate.

These regressions corroborate our findings at the *intensive margin*. Predictions negatively impact the share sent by dictators.[42] Again, this effect does not differ depending on whether the prediction came directly from an AI or from a human in the loop. In this within-analysis, there is also no difference between not receiving a prediction per see (NP) and not receiving it as a consequence of the expert's decision (HIL$_{np}$). The negative effect of receiving a prediction however is strongly statistically significant and robust across our regression specifications.

In contrast, Table 6 contains the same regression specifications, but this time only for those participants where the prediction was ex-post inaccurate, i.e., the actual amount sent as a dictator was outside the predicted category.

[42]Note again that the effect of predictions in the HIL treatment is given by the p-values p(HIL$_p$) in the lower panel of the table.

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-6.685**	-9.763***	-11.641***	-9.117***	-6.663**
	(3.024)	(3.571)	(3.518)	(3.480)	(3.294)
AIP	-5.491*	-5.491*	-5.660*	-3.694	-2.499
	(3.113)	(3.118)	(3.162)	(3.274)	(2.988)
HIL × Prediction		4.436	6.209*	7.351**	5.877*
		(3.552)	(3.591)	(3.438)	(3.393)
Constant	20.180***	20.180***	23.207***	11.796	19.577
	(2.397)	(2.401)	(6.315)	(20.718)	(21.511)
Demographics	✗	✗	✓	✓	✓
Tech attitudes	✗	✗	✗	✓	✓
General attitudes	✗	✗	✗	✗	✓
p(HIL$_p$=AIP)	.	0.958	0.943	0.526	0.559
p(HIL$_p$)	.	0.116	0.124	0.609	0.808
R^2	0.020	0.023	0.077	0.145	0.291
Observations	294	294	294	294	294

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The figures in row p(HIL$_p$=AIP) and p(HIL$_p$) are the p-values corresponding to two-sided F-tests of the linear hypotheses $HIL + HIL \times Prediction = AIP$ and $HIL + HIL \times Prediction = 0$, respectively. The complete table with all coefficients can be found in Appendix A, Tables 16a and 16b.

Table 6: OLS Regressions on the share sent by dictators when prediction was *ex-post* inaccurate.

This analysis aligns with our previous observations at the *extensive margin*. Predictions do not seem to significantly influence the share sent by dictators (AIP and HIL$_p$). Yet, within the HIL treatment, we observe the same difference observed earlier when considering the *Signed deviation* .[43] Participants sent a statistically significantly higher share if the participant received the AI tool's prediction as a consequence of the expert's decision compared to when the expert denied the availability of the AI tool's prediction.

These results suggest that treatment effects might be contingent on the participants' types. We now want to turn our attention to what could determine these types and some additional behavioral explanations of our results.

As motivated earlier, psychological reactance may drive some of the observed results. Our

[43]Please observe that the distinction within the HIL treatment, i.e., whether the expert decided to expose the participant to the AI tool's prediction or not, is detailed by the p-value p(HIL$_p$) in the lower panel of the table.

study distinguishes trait reactance, as identified by Bruns and Perino (2023) or Hong and Faedda (1996), from the measure of reactance in choice (i.e., the deviation from the predicted category in our case). Even though our results do not suggest a different incidence of reactance across treatments at the *extensive margin*, this might be due to counteracting effects among participants with differences in trait reactance. To investigate this, we defined *Reactance high*, which is one if a participant's reactance score is greater than the median across all participants in the same treatment and zero otherwise.[44]

Table 7 shows the regression results on the *Signed deviation* only for participants with above-median reactance. We see that there is a statistically significantly negative effect of predictions on the *Signed deviation* , i.e., participants who received a prediction before they made a decision, on average, chose an amount of roughly 3.8 percentage points below their predicted category. Again, this is independent of whether the prediction was provided by the AI tool directly or after a human expert approved the tool. Note, however, that the difference within the HIL treatment is not statistically significant for this subgroup.
This split analysis hints at the possibility of treatment heterogeneity based on individual characteristics like trait reactance. Whereas this is a plausible behavioral mechanism, one might be worried that the heterogeneity is based on a broader set of characteristics.
To investigate this in a more data-driven fashion, we re-estimated the treatment effects on *Signed deviation* using the nearest neighbor matching based on the entire set of our regression controls. The results can be found in Table 8. The table contains the average treatment effect (ATE) in the upper panel and the average treatment effect on the treated (ATET) in the lower panel. The ATE considers the difference between 'similar' participants in our entire sample, i.e., the missing potential outcome for each participant is imputed by using an average of the outcomes of similar participants that have received the other treatment level. Similarity between participants is determined for each observation by using a weighted function of the controls.

Eventually, the treatment effect is obtained as the average of the difference between the observed and imputed potential outcomes for each participant. For the ATET, the procedure is similar, but only the observations from the observed treatment outcome are compared to their imputed counterfactual and not vice versa.

[44]We specified this variable in this way to ensure balance within each treatment. Results are robust to specifying this indicator along the median across all treatments. Results are available from the authors upon request.

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
HIL	-4.977***	-6.531**	-7.951***	-8.479***	-6.494**
	(1.662)	(2.807)	(2.960)	(2.873)	(2.628)
AIP	-3.576**	-3.576**	-3.968**	-3.547*	-3.393*
	(1.759)	(1.761)	(1.863)	(1.879)	(1.766)
HIL × Prediction		2.017	3.053	4.107	2.224
		(2.706)	(2.750)	(2.625)	(2.421)
Constant	-1.576	-1.576	-1.717	4.953	-14.767
	(1.317)	(1.319)	(3.695)	(12.478)	(12.470)
Demographics	✗	✗	✓	✓	✓
Tech attitudes	✗	✗	✗	✓	✓
General attitudes	✗	✗	✗	✗	✓
p(HIL_p=AIP)	.	0.557	0.563	0.626	0.576
p(HIL_p)	.	0.009	0.007	0.013	0.010
R^2	0.028	0.029	0.056	0.146	0.273
Observations	346	346	346	346	346

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The figures in row p(HIL_p=AIP) and p(HIL_p) are the p-values corresponding to two-sided F-tests of the linear hypotheses $HIL + HIL \times Prediction = AIP$ and $HIL + HIL \times Prediction = 0$, respectively. The complete table with all coefficients can be found in Appendix A, Tables 17a and 17b.

Table 7: OLS regressions on the *Signed deviation* of dictators with a high trait reactance (Bruns and Perino, 2023; Hong and Faedda, 1996).

In both panels, considering the first two rows, respectively, we see that coefficients are negative throughout comparisons. This means that predictions on average and relative to the NP treatment have nominally resulted in a negative deviation from the predicted category. The difference between the NP and the AIP treatment is statistically significant at the 10% level. Thus, when accounting for potential treatment heterogeneity, predictions coming directly from the AI tool lead to a significant negative deviation from the predicted category. As in our previous analysis, however, from the last row in the upper panel and the last two rows in the lower panel, we do not see any difference between predictions coming directly from the AI tool or those that participants received after the human expert approved the tool.[45]

For completeness, participants with no more than median reactance showed no significant extensive margin effect.[46] As in the entire result section so far, there is also no difference between predictions originating directly from the AI tool compared to those received after an expert had approved the AI tool.

However, we do see a consistently positive (and in one specification marginally statistically

[45]Note that the lower panel has two rows for the comparison between HIL_p and AIP, as the ATET is

	Estimate	Std. Error	t-value	p-value	95% CI [Lower, Upper]
ATE					
NP v HIL_p	-1.938	1.533	-1.264	0.206	[-4.944, 1.067]
NP v AIP	-2.108*	1.256	-1.678	0.093	[-4.570, 0.354]
HIL_p v AIP	-0.572	1.175	-0.487	0.627	[-2.874, 1.731]
ATET					
NP v HIL_p	-3.545	2.326	-1.524	0.128	[-8.104, 1.014]
NP v AIP	-3.000*	1.564	-1.918	0.055	[-6.065, 0.065]
HIL_p v AIP	-1.780	1.325	-1.344	0.179	[-4.376, 0.816]
AIP v HIL_p	-1.026	1.295	-0.793	0.428	[-3.564, 1.511]

Note: Estimates are based on the Mahalanobis metric and the same covariates as in our saturated specifications we used throughout the paper (Specifications (5) in the regression tables) together with robust Abadie-Imbens standard errors and one match per observation.

Table 8: Nearest neighbor average treatment effect (ATE) and average treatment effect on the treated (ATET) estimates for *Signed deviation*.

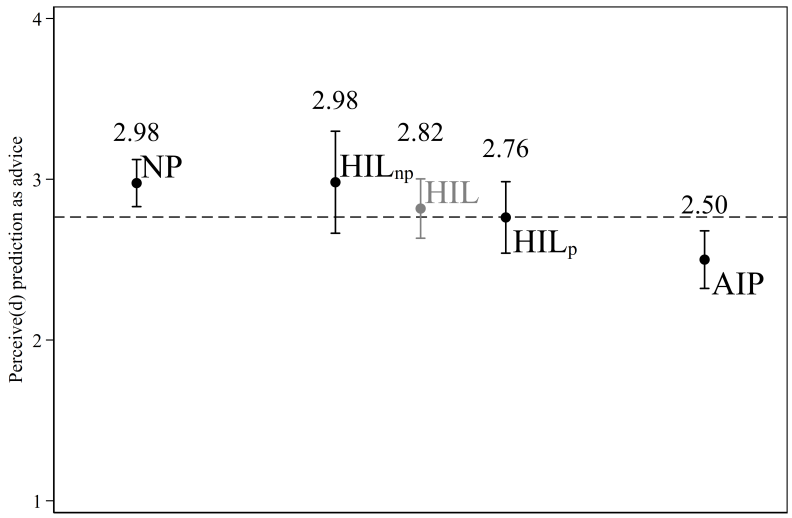
significant) coefficient on the interaction of HIL and Prediction, indicating that predictions have a positive effect on *Signed deviation* within the HIL treatment.

One possible explanation as to why participants behaved this way might be found in how they perceived the AI prediction or how they would have perceived such a prediction had they received one.[47] Figure 3 displays the means of the corresponding survey item where higher responses represent a perception as advice or recommendation, while lower scores represent a perception more as a mere prediction. The AIP treatment notably changed participants' interpretation of the prediction, leaning it away from the perception of a recommendation and more towards an unbiased point of view of prediction. In responses not involving a prediction per se (NP), the mean score was 2.976. For the HIL_{np} subtreatment, the mean score marginally rose to 2.981. However, when the expert provided the AI tools prediction in the HIL_p subtreatment, the average score decreased to 2.763. The most substantial decrease was observed in the AIP treatment, which produced a mean score of 2.400. This difference in the mean scores for how participants (would have) perceived the predictions is statistically significantly different across treatments ($p < 0.001$, Kruskal-Wallis test).

directional in the sense that the counterfactual is only calculated for the second group.

[46]The regression results can be found in in Appendix A, Tables 18a and 18b.

[47]This also implies that a comparison of the replies has to be taken carefully, as treatments might differ due to how the questions were phrased rather than due to actual perceptions differences. The exact phrasing for the NP treatment and the HIL_{np} treatment was "Suppose prior to the decision you made earlier on splitting the 200 Points, you would have received a prediction about your behavior." followed by the request to rate agreement with the statement "I would perceive a prediction about my own behavior rather as advice." whereas for the HIL_p and AIP treatment it was simply "I perceived the prediction rather as advice."



Note: Whiskers indicate 95% confidence intervals. The dashed line indicates the mean when pooling over all treatments and is displayed for better visual representation.

Figure 3: Average score of how much participants perceived prediction as advice across treatments.

Note that this response was only included after the pilot and is thus only available for 634 participants. However, the question of whether participants considered the prediction in their decision is included in the regressions and proxies well for *Perceiv(ed) as advice* (the pairwise correlation coefficient between *Perceiv(ed) as advice* and *Consider* is $\rho = 0.427$ and is statistically different from zero, $p < 0.001$, t-test). This control is available for all 754 participants in our sample. These controls suggest that, at least to a certain extent, the human expert changed the interpretation of the prediction into advice, leading some participants (particularly those with low trait reactance) to follow it.

5 Discussion

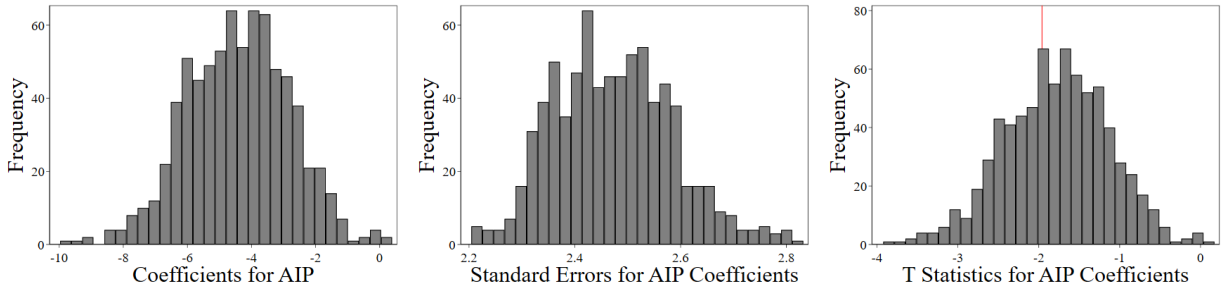
We find that providing participants with salient and personalized AI predictions affects their decisions and that this is more pronounced for participants with a higher tendency to psychological reactance. Here, we want to discuss some alternative explanations of our results and future avenues for research.

One might be concerned about the participants' attitudes toward data privacy. There exists evidence that the more personalized data is used, the fewer people tend to accept advertisements and even show resistance towards the message and the advertiser - also known as the *personalization paradox* (Aguirre et al., 2015; de Groot, 2022; Boerman et al., 2021). Many of these studies argue that this is due to the consumers' feelings of vulnerability to

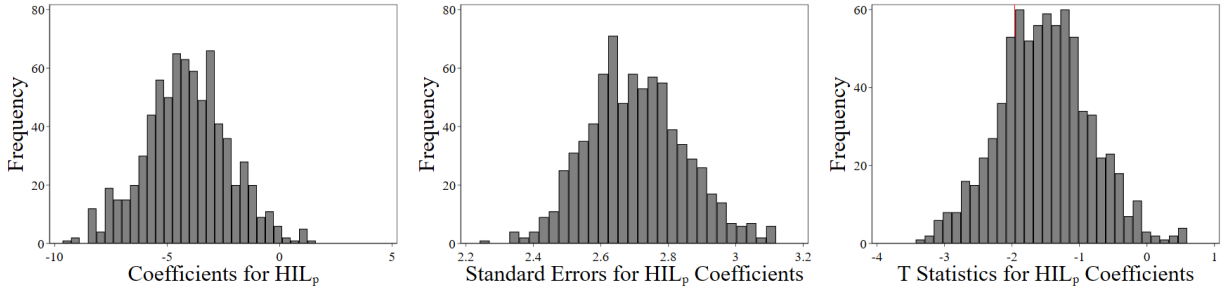
companies' covert information-gathering strategies. We levied the IUIPC Scale by Malhotra et al. (2004) to measure participants' data privacy concerns for the 634 participants in the sample, excluding the pilot. There, we neither found any differences across the treatments nor did controlling for this measure change anything substantial in our regression analyses. Furthermore, we framed our experiment neutrally and have no plausible reason to assume that we are perceived as a commercial agent gathering data for monetary profit. Thus, we would argue that this is not a (strong) driver of behavior in the context of behavioral predictions studied here.

Even though we found some notable effects on the amounts sent, the effect sizes are rather mild. We believe, however, that the magnitudes of these effects should be considered as lower bounds to the true effects. There are several reasons for this, which we outline below.

Firstly, our AI model achieved an accuracy of 61.71%. While this indicates a reasonably good performance of the model over three categories, it also means that only 61.71% of our participants were actually treated with a correct prediction.



(a) Distribution of coefficients, standard errors, and T statistics for AIP.



(b) Distribution of coefficients, standard errors, and T statistics for HIL$_p$.

Note: The red lines in the T statistics graphs indicate the critical values for the 5% level (1.96).

Figure 4: Simulation based on 61.72% accuracy of the AI tool.

The remaining 38.29% of the participants received incorrect predictions, which could po-

tentially weaken the observed overall effects. Consider Figure 4. There, we randomly drew 61.72% from the participants in the AIP treatment (Panel 4a) and the HIL_p treatment (Panel 4b), respectively. We then ran our simplest regression model accounting for the difference between participants in the HIL_{np} and HIL_p subtreatments (Specification (3) in our regression tables).[48] Two observations can be made. Firstly, coefficients are almost exclusively negative, i.e., there is hardly any chunk of data in our prediction treatments that would result in a positively signed coefficient. Secondly, the variation in effect sizes is rather large. In the most extreme cases, coefficients suggest a negative deviation of around 10 percentage points from the predicted category. Under the assumptions that (a) our accuracy measure is a decent estimate for the expected *ex-ante* accuracy in our sample, and (b) effects are larger when predictions are *ex-ante* correct, this might suggest that with more accurate predictions, the true impact on the decision-makers' choices are even stronger than the ones we observed.

Secondly, our predictions were relatively coarse in nature, providing decision-makers with a broad outline of their potential decisions. It is reasonable to assume that more precise predictions, which can pin the decisions down more closely, would be perceived as more infringing. As a result, the effects of receiving more detailed predictions may lead decision-makers to feel that their autonomy is compromised, potentially causing a stronger reaction and greater influence on their actual decisions.

Finally, our study only analyzed a specific sample of participants: people who have willingly subscribed to an online platform that handles their data and payment information. This is noteworthy, as it suggests that our sample may be more open to sharing their personal data and have relatively lower privacy concerns than the general population. Consequently, the impact of AI-generated behavioral predictions could be even more substantial in a more diverse population, as individuals who are more concerned about their privacy could exhibit stronger reactions to the presented predictions.

A limitation of our study is the comparison between our baseline and the prediction treatments. Whereas using the absence of a prediction as the baseline is ecologically valid when we want to consider the total effect of introducing personalized AI predictions, it is not clear how much of their effect is due to (seemingly) providing statistical information about previous decisions to the participants. We inform participants about the fact that our AI tool is based on 15,700 previous decisions. Even though predictions are personalized, i.e.,

[48]The results are similar for using the full specification. Results are available from the authors upon request.

tailored to the specific participant and we do not give any further details about the 15,700 previous decisions to participants, they could see their predictions as a statistical piece of information anchoring their choice or informing their assessment of social appropriateness of their actions. However, when we test the differences between participants' beliefs about other participants' decisions in the experiment or the social appropriateness of the three prediction categories, we do not find statistically significant differences ($p = 0.229$ for the beliefs and $p = 0.938$ for the social norms, Kruskal-Wallis test).

In conclusion, while our findings demonstrate some impacts of AI-generated behavioral predictions on decision-makers, we posit that the true effects could potentially be even larger. By considering factors such as the accuracy of the predictions, the granularity of the prediction, and the participant sample, our study highlights the importance of these aspects when analyzing the influence of AI-generated behavioral predictions on decision-making processes. Our study can add a dimension to the debate on regulating personalized content, prizes, and products based on big data and AI (see, e.g., Ali et al., 2023; Beraja et al., 2023b). How these effects compare to settings of algorithmic advice or the possibility of delegating decisions to an AI, is a question for future research.

6 Conclusion

As access to big data grows and computing resources become more affordable, frequent predictions about individuals' behavior raise concerns about potential adverse effects. Our primary research question explores the behavioral response to personalized AI predictions.

While there's currently no conclusive evidence that predictions alter human behavior, our initial experiment on dictator game decisions reveals that awareness of predicted decisions may steer individuals away from those predictions. Surprisingly, our findings show no significant difference between predictions directly from an AI tool and those approved by a human overseer. However, a noteworthy negative effect emerges when participants do not receive a predicted outcome due to the overseer's disapproval. Within the treatment group involving human oversight (HIL), predictions induced less pronounced deviations compared to the subtreatment in which the human expert did not provide participants with the AI tool and consequently with a prediction. Consistent with Falk and Fischbacher (2006b), participants' reactions to not receiving predictions indicate reciprocal behavior - rewarding kindness and punishing unkindness, even in AI predictions. This implies that the presence of *"humans in the loop"* may undermine the effectiveness of predictions. In addition, reactance

appears to be a key mechanism, as individuals with high trait reactance significantly deviate from predicted categories by sharing less.

The growing use of AI predictions in economic domains, such as Netflix, Spotify, Amazon, and Facebook, highlights the importance of understanding how individuals react to them. Our results reveal a subset of people exhibiting reactant behavior, deviating from predicted categories, suggesting a potential adverse reaction to AI tools predicting personal decisions. Clearly, this evidence should be viewed as a first step in understanding the influence of AI predictions on one's own decision. We recognize that there is still much to explore surrounding the topic of how predictions affect human behavior. We are currently in the process of experimentally disentangling the differences between predictions coming from an AI versus predictions that originate from a human to put the results presented here in perspective. The insights gathered can have important implications for the use of AI tools in consumer choice and financial decision-making and suggest that more nuanced considerations of human reactions to decision predictions are essential when judging prediction accuracy.

References

117th Congress (2022): "Algorithmic Accountability Act of 2022," https://www.congress.gov/bill/117th-congress/senate-bill/3572, last access: 05/03/2024.

Acemoglu, D. and P. Restrepo (2022a): "Demographics and automation," *Review of Economic Studies*, 89, 1–44.

——— (2022b): "Tasks, automation, and the rise in US wage inequality," *Econometrica*, 90, 1973–2016.

Adena, M. and A. Hager (2024): "Does online fundraising increase charitable giving? A nationwide field experiment on Facebook," CESifo Working Paper.

Agarwal, R. and J. Prasad (1998): "A conceptual and operational definition of personal innovativeness in the domain of information technology," *Information Systems Research*, 9, 204–215.

Agrawal, A., J. Gans, and A. Goldfarb (2018): *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Press.

Aguirre, E., D. Mahr, D. Grewal, K. De Ruyter, and M. Wetzels (2015): "Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness," *Journal of Retailing*, 91, 34–49.

Ahn, D. S. and T. Sarver (2013): "Preference for flexibility and random choice," *Econometrica*, 81, 341–361.

Ali, S. N., G. Lewis, and S. Vassermann (2023): "Voluntary Disclosure and Personalized Pricing," *Review of Economic Studies*, 90, 538–571.

Amir, O., D. G. Rand, and Y. K. Gal (2012): "Economic games on the internet: The effect of $1 stakes," *PLOS One*, 7, e31461.

Antoniades, A., G. Seshan, R. Weber, and R. Zubrickas (2018): "Does altruism matter for remittances?" *Oxford Economic Papers*, 70, 225–242.

Barkai, E. (2023): "Outsmarting ChatGPT to Say What It Really Thinks of Human Kind," https://www.linkedin.com/pulse/outsmarting-chatgpt-say-what-really-thinks-human-kind-eran-barkai/, last accessed: September 27, 2023.

Baumeister, R. F., E. J. Masicampo, and C. N. Dewall (2009): "Prosocial Benefits of Feeling Free: Disbelief in Free Will Increases Aggression and Reduces Helpfulness," *Personality and Social Psychology Bulletin*, 35, 260–268.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 81, 608–650.

Beraja, M., A. Kao, D. Y. Yang, and N. Yuchtman (2023a): "AI-tocracy," *The Quarterly Journal of Economics*, 138, 1349–1402.

Beraja, M., D. Y. Yang, and N. Yuchtman (2023b): "Data-intensive innovation and the state: Evidence from AI firms in China," *Review of Economic Studies*, 90, 1701–1723.

Berger, B., M. Adam, A. Rühr, and A. Benlian (2021): "Watch me improve—algorithm aversion and demonstrating the ability to learn," *Business & Information Systems Engineering*, 63, 55–68.

Bhatt, M. P., S. B. Heller, M. Kapustin, M. Bertrand, and C. Blattman (2024): "Predicting and preventing gun violence: An experimental evaluation of READI Chicago," *The Quarterly Journal of Economics*, 139, 1–56.

Bigman, Y. E. and K. Gray (2018): "People are averse to machines making moral decisions," *Cognition*, 181, 21–34.

Bodner, R. and D. Prelec (2003): "Self-signaling and diagnostic utility in everyday decision making," *The Psychology of Economic Decisions*, 1, 26.

Boerman, S. C., S. Kruikemeier, and N. Bol (2021): "When is personalized advertising crossing personal boundaries? How type of information, data sharing, and personalized pricing influence consumer perceptions of personalized advertising," *Computers in Human Behavior Reports*, 4, 100144.

Brehm, J. W. (1966): *A Theory of Psychological Reactance*, Academic Press.

Brehm, J. W., L. K. Stires, J. Sensenig, and J. Shaban (1966): "The Attractiveness of an Eliminated Choice Alternative," *Journal of Experimental Social Psychology*, 2, 301–313.

Brehm, S. and J. W. Brehm (1981): "Psychological Reactance: A Theory of Freedom and Control," *Academic Press*.

Brock, J. M., A. Lange, and E. Y. Ozbay (2013): "Dictating the risk: Experimental evidence on giving in risky environments," *American Economic Review*, 103, 415–437.

Bruns, H., E. Kantorowicz-Reznichenko, K. Klement, M. L. Jonsson, and B. Rahali (2018): "Can nudges be transparent and yet effective?" *Journal of Economic Psychology*, 65, 41–59.

Bruns, H. and G. Perino (2023): "The role of autonomy and reactance for nudging-experimentally comparing defaults to recommendations and mandates," *Journal of Behavioral and Experimental Economics*, 106, 102047.

Burton, S., I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter (2020): "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artificial Intelligence*, 279, 103201.

Buttrick, N. (2020): "Protective gun ownership as a coping mechanism," *Perspectives on Psychological Science*, 15, 835–855.

Caro, F. and A. S. de Tejada Cuenca (2023): "Believing in Analytics: Managers' Adherence to Price Recommendations from a DSS," *Manufacturing & Service Operations Management*, 25, 371–810.

Castelo, N. (2019): *Blurring the line between human and machine: marketing artificial intelligence*, Columbia University.

Ceccato, S., S. E. Kettner, B. M. Kudielka, C. Schwieren, and A. Voss (2018): "Social preferences under chronic stress," *PLOS One*, 13, e0199528.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Chugunova, M. and D. Sele (2022): "An interdisciplinary review of the experimental evidence on how humans interact with machines," *Journal of Behavioral and Experimental Economics*, 99, 101897.

Cochard, F., J. Le Gallo, N. Georgantzis, and J. C. Tisserand (2021): "Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game," *Journal of Behavioral and Experimental Economics*, 90, 101613.

Corgnet, B., R. Hernán-Gonzalez, and R. Mateo (2019): "Rac (g) e against the machine? social incentives when humans meet robots," Working Paper.

Cowgill, B. and E. Zitzewitz (2015): "Corporate prediction markets: Evidence from google, ford, and firm x," *Review of Economic Studies*, 82, 1309–1341.

D'agostino, R. B., A. Belanger, and R. B. D'Agostino Jr (1990): "A suggestion for using powerful and informative tests of normality," *The American Statistician*, 44, 316–321.

Dai, R., C. Lu, L. Yun, E. Lenze, M. Avidan, and T. Kannampallil (2021): "Comparing stress prediction models using smartwatch physiological signals and participant self-reports," *Computer Methods and Programs in Biomedicine*, 208, 106207.

de Groot, J. I. M. (2022): "The personalization paradox in Facebook advertising: The mediating effect of relevance on the personalization–brand attitude relationship and the moderating effect of intrusiveness," *Journal of Interactive Advertising*, 22, 57–74.

Dean, M. and J. McNeill (2014): "Preference for flexibility and random choice: an experimental analysis," Working Paper.

Dekel, E., B. L. Lipman, and A. Rustichini (2001): "Representing preferences with a unique subjective state space," *Econometrica*, 69, 891–934.

DellaVigna, S. and D. Pope (2018): "Predicting experimental results: who knows what?" *Journal of Political Economy*, 126, 2410–2456.

DellaVigna, S., D. Pope, and E. Vivalt (2019): "Predict science to improve science," *Science*, 366, 428–429.

Dietvorst, B. J. and S. Bharti (2019): "Risk Seeking Preferences Lead Consumers to Reject Algorithms in Uncertain Domains," *ACR North American Advances*, 47, 78–81.

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

——— (2018): "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, 64, 1155–1170.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual risk attitudes: Measurement, determinants, and behavioral consequences," *Journal of the European Economic Association*, 9, 522–550.

Dreber, A., T. Ellingsen, M. Johannesson, and D. G. Rand (2013): "Do people care about social context? Framing effects in dictator games," *Experimental Economics*, 16, 349–371.

Drobner, C. and S. J. Goerg (2024): "Motivated belief updating and rationalization of information," *Management Science*.

Dunbar, R., A. Frangou, F. Grainger, and E. Pearce (2021): "Laughter influences social bonding but not prosocial generosity to friends and strangers," *PLOS One*, 16, e0256229.

Engel, C. (2011): "Dictator games: A meta study," *Experimental Economics*, 14, 583–610.

European Commission, Directorate-General for Communications Networks, Content and Technology (2021): "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS," COM(2021) 206 final, 2021/0106(COD), `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`, last access: 05/03/2024.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018): "Global evidence on economic preferences," *The Quarterly Journal of Economics*, 133, 1645–1692.

Falk, A. and U. Fischbacher (2006a): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293–315.

——— (2006b): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293–315.

Falk, A. and M. Kosfeld (2006): "The hidden costs of control," *American Economic Review*, 96, 1611–1630.

Fenzl, T. and T. Brudermann (2021): "Eye cues increase cooperation in the dictator game under physical attendance of a recipient, but not for all," *Journal of Behavioral and Experimental Economics*, 94, 101748.

Fornwagner, H., B. Grosskopf, A. Lauf, V. Schöller, and S. Städter (2022): "On the robustness of gender differences in economic behavior," *Scientific Reports*, 12, 21549.

Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton (1994): "Fairness in simple bargaining experiments," *Games and Economic Behavior*, 6, 347–369.

Fouka, V. (2020): "Backlash: The unintended effects of language prohibition in US schools after World War I," *Review of Economic Studies*, 87, 204–239.

Fudenberg, D. and A. Liang (2019): "Predicting and understanding initial play," *American Economic Review*, 109, 4112–4141.

Gefen, D. and D. W. Straub (2004): "Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services," *Omega*, 32, 407–424.

Giannone, D., M. Lenza, and G. E. Primiceri (2021): "Economic predictions with big data: The illusion of sparsity," *Econometrica*, 89, 2409–2437.

Goerg, S. J., D. Rand, and G. Walkowitz (2020): "Framing effects in the prisoner's dilemma but not in the dictator game," *Journal of the Economic Science Association*, 6, 1–12.

Gogoll, J. and M. Uhl (2018): "Rage against the machine: Automation in the moral domain," *Journal of Behavioral and Experimental Economics*, 74, 97–103.

Grammarly Inc. (2023): "Grammarly," `https://www.grammarly.com`, accessed: February 2, 2024.

Guerreiro, J., S. Rebelo, and P. Teles (2022): "Should robots be taxed?" *Review of Economic Studies*, 89, 279–311.

Güth, W., R. Schmittberger, and B. Schwarze (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3, 367–388.

Habersaat, K. B., C. Betsch, M. Danchin, C. R. Sunstein, R. Böhm, A. Falk, N. T. Brewer, S. B. Omer, M. Scherzer, S. Sah, et al. (2020): "Ten considerations for effectively managing the COVID-19 transition," *Nature Human Behaviour*, 4, 677–687.

Han, Y., Y. Liu, and G. Loewenstein (2023): "Confusing context with Character: Correspondence bias in economic interactions," *Management Science*, 69, 1070–1091.

Hannah, T. E., E. R. Hannah, and B. Wattie (1975): "Arousal of Psychological Reactance as a Consequence of Predicting an Individual's Behaviour," *Psychological Reports*, 37, 411–420.

Harms, J., K. Liket, J. Protzko, and V. Schölmerich (2017): "Free to help? An experiment on free will belief and altruism," *PLOS One*, 12, e0173193.

Hedgcock, W. M., R. S. Rao, and H. A. Chen (2016): "Choosing to Choose: The Effects of Decoys and Prior Choice on Deferral," *Management Science*, 62, 2952–2976.

Hern, A. (2020): "Berlin artist uses 99 phones to trick Google into traffic jam alert," `https://www.theguardian.com/technology/2020/feb/03/berlin-artist-uses-99-phones-trick-google-maps-traffic-jam-alert`, last accessed: September 27, 2023.

Herne, K., J. K. Hietanen, O. Lappalainen, and E. Palosaari (2022): "The influence of role awareness, empathy induction and trait empathy on dictator game giving," *PLOS One*, 17, e0262196.

Hertz, N. and E. Wiese (2018): "Under pressure: Examining social conformity with computer and robot groups," *Human Factors*, 60, 1207–1218.

Hoffmann, R., T. Chesney, S.-H. Chuah, F. Kock, and J. Larner (2020): "Demonstrability, difficulty and persuasion: An experimental study of advice taking," *Journal of Economic Psychology*, 76, 102215.

Hong, S.-M. and S. Faedda (1996): "Refinement of the Hong psychological reactance scale," *Educational and Psychological Measurement*, 56, 173–182.

Hong, S.-M. and S. Page (1989): "A psychological reactance scale: Development, factor structure and reliability," *Psychological Reports*, 64, 1323–1326.

H&R Block (2023): "H&R Block AI Tax Assist," `https://www.hrblock.com`, accessed: February 2, 2024.

Hubmer, J. (2023): "The race between preferences and technology," *Econometrica*, 91, 227–261.

Inaba, M., Y. Inoue, S. Akutsu, N. Takahashi, and T. Yamagishi (2018): "Preference and strategy in proposer's prosocial giving in the ultimatum game," *PLOS One*, 13, e0193877.

Intuit Inc. (2023): "Intuit TurboTax," `https://turbotax.intuit.com`, accessed: February 2, 2024.

Jeffrey, K. (2021): "Automation and the future of work: How rhetoric shapes the response in policy preferences," *Journal of Economic Behavior & Organization*, 192, 417–433.

Jeffrey, K. and K. Matakos (2024): "Automation anxiety, fairness perceptions, and redistribution: Past experiences condition the response to future job loss," *Journal of Economic Behavior & Organization*, 221, 174–190.

Jussupow, E., I. Benbasat, and A. Heinzl (2020): "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion," European Conference on Information Systems.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986): "Fairness and the assumptions of economics," *Journal of Business*, 59, 285–300.

Kandul, S. and O. Nikolaychuk (2023): "I win it's fair, you win it's not. Selective heeding of merit in ambiguous settings," *PLOS One*, 18, e0279865.

Kantar Public (2020): "SOEP-Core-2019: Personenfragebogen, Stichproben A-L3, M1-M2 + N-P," SOEP Survey Papers 909: Series A.

Kause, A., O. Vitouch, and J. Glück (2018): "How selfish is a thirsty man? A pilot study on comparing sharing behavior with primary and secondary rewards," *PLOS One*, 13, e0201358.

Kee, J., M. Knuth, J. N. Lahey, and M. A. Palma (2021): "Does eye-tracking have an effect on economic behavior?" *PLOS One*, 16, e0254867.

Keeper Tax (2023): "Keeper: AI Tax Assistant," `https://www.keepertax.com`, accessed: February 2, 2024.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 133, 237–293.

Klockmann, V., A. Von Schenk, and M. C. Villeval (2022): "Artificial intelligence, ethics, and intergenerational responsibility," *Journal of Economic Behavior & Organization*, 203, 284–317.

Koopmans, T. C. (1962): "On flexibility of future preference," in *Human Judgments and Optimality*, ed. by M. W. Shelly and G. L. Bryan., New York: John Wiley and Sons.

Kopylov, I. (2009): "Choice deferral and ambiguity aversion," *Theoretical Economics*, 4, 199–225.

Kovaleva, A. (2012): *The IE-4: Construction and validation of a short scale for the assessment of locus of control*, vol. 9, DEU.

Kreps, D. M. (1979): "A representation theorem for "preference for flexibility"," *Econometrica*, 565–577.

Krupka, E. and R. A. Weber (2013): "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11, 495–524.

Kumar, M. M., L. Tsoi, M. S. Lee, J. Cone, and K. McAuliffe (2021): "Nationality dominates gender in decision-making in the Dictator and Prisoner's Dilemma Games," *PLOS One*, 16, e0244568.

Kunda, Z. (1990): "The case for motivated reasoning." *Psychological Bulletin*, 108, 480.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012): "Sorting in experiments with application to social preferences," *American Economic Journal: Applied Economics*, 4, 136–163.

Lee, J., H. Davari, J. Singh, and V. Pandhare (2018): "Industrial Artificial Intelligence for industry 4.0-based manufacturing systems," *Manufacturing Letters*, 18, 20–23.

Li, J., Z. Liao, and R. Quaedvlieg (2022): "Conditional superior predictive ability," *Review of Economic Studies*, 89, 843–875.

Little, R. J. and D. B. Rubin (2019): *Statistical analysis with missing data*, vol. 793, John Wiley & Sons.

Logg, J., J. Minson, and D. Moore (2018): "Robo-Advising: Algorithm Appreciation," *ACR North American Advances*, 46, 63–67.

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

Longoni, C., A. Bonezzi, and C. K. Morewedge (2019): "Resistance to medical artificial intelligence," *Journal of Consumer Research*, 46, 629–650.

Ludwig, J. and S. Mullainathan (2024): "Machine learning as a tool for hypothesis generation," *The Quarterly Journal of Economics*, 139, 751–827.

Ma, A., S. Tang, and A. C. Kay (2019): "Psychological reactance as a function of thought versus behavioral control," *Journal of Experimental Social Psychology*, 84, 103825.

Malhotra, N. K., S. S. Kim, and J. Agarwal (2004): "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model," *Information Systems Research*, 15, 336–355.

Maltz, A. and S. Rachmilevitch (2021): "A model of menu-dependent evaluations and comparison-aversion," *Journal of Behavioral and Experimental Economics*, 91, 101655.

Matsumoto, Y., T. Yamagishi, Y. Li, and T. Kiyonari (2016): "Prosocial behavior increases with age across five economic games," *PLOS One*, 11, e0158671.

Mesa-Vázquez, E., I. Rodriguez-Lara, and A. Urbano (2021): "Standard vs random dictator games: On the effects of role uncertainty and framing on generosity," *Economics Letters*, 206, 109981.

Microsoft Corporation (2023): "Microsoft Editor," `https://www.microsoft.com/en-us/microsoft-365/microsoft-editor`, accessed: February 2, 2024.

Mochon, D. (2013): "Single-Option Aversion," *Journal of Consumer Research*, 40, 555–566.

Moyal, A. and I. Ritov (2020): "The effect of contest participation and contest outcome on subsequent prosocial behavior," *PLOS One*, 15, e0240712.

Mullainathan, S. and Z. Obermeyer (2022): "Diagnosing physician error: A machine learning approach to low-value health care," *The Quarterly Journal of Economics*, 137, 679–727.

Nava, F., F. Margoni, N. Herath, and E. Nava (2023): "Age-dependent changes in intuitive and deliberative cooperation," *Scientific Reports*, 13, 4457.

Noussair, C. N. and J. Stoop (2015): "Time as a medium of reward in three social preference experiments," *Experimental Economics*, 18, 442–456.

Oechssler, J. (2010): "Searching beyond the lamppost: Let's focus on economically relevant questions," *Journal of Economic Behavior & Organization*, 73, 65–67.

Oleszkiewicz, A. and T. Kupczyk (2020): "Sensory impairment reduces money sharing in the Dictator Game regardless of the recipient's sensory status," *PLOS One*, 15, e0230637.

Önkal, D., P. Goodwin, M. Thomson, S. Gönül, and A. Pollock (2009): "The relative influence of advice from human experts and statistical methods on forecast adjustments," *Journal of Behavioral Decision Making*, 22, 390–409.

Oosterbeek, H., R. Sloof, and G. Van De Kuilen (2004): "Cultural differences in ultimatum game experiments: Evidence from a meta-analysis," *Experimental Economics*, 7, 171–188.

Orpheus Technology Ltd. (2023): "ProWritingAid," `https://prowritingaid.com`, accessed: February 2, 2024.

Owens, D., Z. Grossman, and R. Fackler (2014): "The control premium: A preference for payoff autonomy," *American Economic Journal: Microeconomics*, 6, 138–161.

Palan, S. and C. Schitter (2018): "Prolific. ac—A subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, 17, 22–27.

Palmeira, M. and G. Spassova (2015): "Consumer reactions to professionals who use decision aids," *European Journal of Marketing*, 49, 302–326.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011): "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, 12, 2825–2830.

Peysakhovich, A., M. A. Nowak, and D. G. Rand (2014): "Humans display a 'cooperative phenotype'that is domain general and temporally stable," *Nature Communications*, 5, 4939.

Pizzi, G., D. Scarpi, and E. Pantano (2021): "Artificial intelligence and the new forms of interaction: Who has the control when interacting with a chatbot?" *Journal of Business Research*, 129, 878–890.

Poole, D., A. Mackworth, and R. Goebel (1998): *Computational Intelligence*, Oxford University Press.

Powell, P. A., O. Wills, G. Reynolds, K. Puustinen-Hopper, and J. Roberts (2018): "The effects of exposure to images of others' suffering and vulnerability on altruistic, trust-based, and reciprocated economic decision-making," *PLOS One*, 13, e0194569.

Promberger, M. and J. Baron (2006): "Do patients trust computers?" *Journal of Behavioral Decision Making*, 19, 455–468.

Raihani, N. J. and K. McAuliffe (2014): "Dictator game giving: The importance of descriptive versus injunctive norms," *PLOS One*, 9, e113826.

Rand, D. G., V. L. Brescoll, J. A. Everett, V. Capraro, and H. Barcelo (2016): "Social heuristics and social roles: Intuition favors altruism for women but not for men." *Journal of Experimental Psychology: General*, 145, 389.

Rattini, V. (2023): "Worker autonomy and performance: Evidence from a real-effort experiment," *Journal of Economics & Management Strategy*, 32, 300–327.

Ritov, I. and T. Kogut (2017): "Altruistic behavior in cohesive social groups: The role of target identifiability," *PLOS One*, 12, e0187903.

Rotella, A., A. M. Sparks, S. Mishra, and P. Barclay (2021): "No effect of 'watching eyes': An attempted replication and extension investigating individual differences," *PLOS One*, 16, e0255531.

Royston, P. (1992): "Comment on sg3.4 and an Improved D'Agostino Test," *Stata Technical Bulletin*, 1.

Russell, S. J. and P. Norvig (2010): "Artificial Intelligence: A Modern Approach," *Artificial Intelligence*.

Sankaran, S., C. Zhang, H. Aarts, and P. Markopoulos (2021): "Exploring Peoples' Perception of Autonomy and Reactance in Everyday AI Interactions," *Frontiers in Psychology*, 12, 713074.

Selten, R. (1967): "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments," *Beiträge zur experimentellen Wirtschaftsforschung*, 136–168.

Sheng, H. and Y. Chen (2020): "An Empirical Study on Factors influencing Users' Psychological Reactance to Artificial Intelligence Applications," in *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*, IEEE, 234–237.

Spiegel, J. R., M. T. McKenna, G. S. Lakshman, and N. P. G. (2013): "Method and system for anticipatory package shipping," US Patent No. US8615473B2.

Stagnaro, M. N., A. A. Arechar, and D. G. Rand (2017): "From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement," *Cognition*, 167, 212–254.

Stango, V. and J. Zinman (2023): "We Are All Behavioural, More, or Less: A Taxonomy of Consumer Decision-Making," *Review of Economic Studies*, 90, 1470–1498.

Steindl, C., E. Jonas, S. Sittenthaler, E. Traut-Mattausch, and J. Greenberg (2015): "Understanding psychological reactance: New developments and findings," *Journal of Psychology*, 223, 205–214.

Stock, J. H. and M. W. Watson (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

Thornton, E. M. and L. B. Aknin (2020): "Assessing the validity of the Self versus other interest implicit association test," *PLOS One*, 15, e0234032.

Van Buuren, S. (2007): "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical methods in medical research*, 16, 219–242.

Van Petegem, S., B. Soenens, M. Vansteenkiste, and W. Beyers (2015): "Rebels with a cause? Adolescent defiance from the perspective of reactance theory and self-determination theory," *Child Development*, 86, 903–918.

Veszteg, R. F., K. Yamakawa, T. Matsubayashi, and M. Ueda (2021): "Acute stress does not affect economic behavior in the experimental laboratory," *PLOS One*, 16, e0244881.

von Bieberstein, F., A. Essl, and K. Friedrich (2021): "Empathy: A clue for prosocialty and driver of indirect reciprocity," *PLOS One*, 16, e0255071.

von Schenk, A., V. Klockmann, J.-F. Bonnefon, I. Rahwan, and N. Köbis (2022): "Lie detection algorithms attract few users but vastly increase accusation rates," *arXiv preprint arXiv:2212.04277*.

Walkowitz, G. (2021): "Dictator game variants with probabilistic (and cost-saving) payoffs: A systematic test," *Journal of Economic Psychology*, 85, 102387.

Waytz, A. and M. I. Norton (2014): "Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs." *Emotion*, 14, 434.

Xu, R. (2024): "Persuasion, Delegation, and Private Information in Algorithm-Assisted Decisions," ArXiv Working Paper 2402.09384.

Ybarra, O., M. C. Keller, E. Chan, S. M. Garcia, J. Sanchez-Burks, K. R. Morrison, and A. S. Baron (2010): "Being unpredictable: Friend or foe matters," *Social Psychological and Personality Science*, 1, 259–267.

Yeomans, M., A. Shah, S. Mullainathan, and J. Kleinberg (2019): "Making sense of recommendations," *Journal of Behavioral Decision Making*, 32, 403–414.

Zheng, Y., S. M. Van Osselaer, and J. W. Alba (2016): "Belief in free will: Implications for practice and policy," *Journal of Marketing Research*, 53, 1050–1064.

Zimmermann, F. (2020): "The dynamics of motivated beliefs," *American Economic Review*, 110, 337–363.

A Tables and Graphs

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
HIL	-2.039*	-4.078**	-4.713***	-4.749***	-3.359**
	(1.119)	(1.750)	(1.776)	(1.751)	(1.567)
AIP	-1.436	-1.436	-1.860	-1.304	-1.106
	(1.109)	(1.110)	(1.142)	(1.144)	(1.050)
HIL × Prediction		2.740	3.002*	3.655**	2.466
		(1.755)	(1.752)	(1.698)	(1.529)
Age			-0.012	-0.010	-0.039
			(0.035)	(0.036)	(0.034)
Female			0.473	0.231	-0.435
			(0.905)	(0.923)	(0.854)
Attended college			0.546	0.318	0.135
			(1.332)	(1.330)	(1.180)
Undergraduate degree			-0.290	-0.561	-1.112
			(1.090)	(1.067)	(0.994)
Graduate degree or higher			-1.440	-1.630	-2.388*
			(1.480)	(1.479)	(1.323)
Student			1.710	1.362	1.877
			(1.413)	(1.465)	(1.287)
Employed			0.515	0.165	-0.735
			(1.115)	(1.118)	(1.004)
More than £25k, up to £50k			-1.017	-1.086	-0.821
			(1.094)	(1.097)	(0.997)
More than £50k			-0.935	-1.219	-0.665
			(1.368)	(1.353)	(1.262)
Siblings			-0.139	-0.113	-0.016
			(0.283)	(0.282)	(0.247)
Religious			0.331	-0.093	-0.723
			(1.119)	(1.155)	(1.040)
Econ./Business admin. background			3.927***	4.135***	4.277***
			(1.140)	(1.154)	(1.092)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9a: OLS Regressions of *Signed deviation* (complete table, part 1 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
Consider				1.682***	1.431***
				(0.354)	(0.317)
Accuracy				-0.008	-0.354
				(0.594)	(0.539)
Experience with AI				-0.367	-0.343
				(0.631)	(0.586)
Interaction with AI				-0.193	-0.241
				(0.562)	(0.504)
Profession with AI				-0.800	-0.941
				(1.453)	(1.343)
Belief in free will				0.309	0.007
				(0.262)	(0.237)
Reactance				0.000	0.690
				(0.775)	(0.746)
Attitude toward IT				0.940	0.229
				(0.865)	(0.789)
Locus of control				-0.683	-0.512
				(0.875)	(0.832)
Attitude toward AI				0.126	-0.431
				(0.717)	(0.664)
Risk aversion					0.182
					(0.177)
Altruism					0.027***
					(0.007)
Trust					1.430***
					(0.486)
Belief in DG					0.243***
					(0.034)
Norm in DG					0.108**
					(0.049)
Constant	-2.868***	-2.868***	-2.770	-8.396	-25.586***
	(0.834)	(0.835)	(2.360)	(7.476)	(8.512)
p(HIL_p=AIP)	.	0.930	0.895	0.852	0.835
p(HIL_p)	.	0.260	0.157	0.369	0.438
R^2	0.005	0.008	0.033	0.070	0.226
Observations	754	754	754	754	754

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9b: OLS Regressions of *Signed deviation* (complete table, part 2 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
HIL	-2.450**	-4.674***	-5.413***	-5.680***	-4.334***
	(1.062)	(1.641)	(1.637)	(1.590)	(1.406)
AIP	-1.629	-1.629	-1.967*	-1.584	-1.409
	(1.071)	(1.072)	(1.103)	(1.089)	(0.988)
HIL × Prediction		2.994*	3.343**	4.042**	2.907**
		(1.661)	(1.630)	(1.569)	(1.388)
Age			0.049	0.062*	0.035
			(0.034)	(0.034)	(0.031)
Female			2.316***	2.141**	1.565**
			(0.882)	(0.875)	(0.787)
Attended college			0.656	0.424	0.385
			(1.338)	(1.330)	(1.166)
Undergraduate degree			-0.865	-1.122	-1.783*
			(1.064)	(1.035)	(0.931)
Graduate degree or higher			-1.614	-1.882	-2.554**
			(1.459)	(1.455)	(1.286)
Student			0.309	-0.133	0.605
			(1.423)	(1.445)	(1.276)
Employed			1.437	1.106	0.269
			(1.106)	(1.107)	(0.985)
More than £25k, up to £50k			-1.468	-1.782*	-1.406
			(1.047)	(1.060)	(0.952)
More than £50k			-0.986	-1.578	-0.881
			(1.394)	(1.335)	(1.210)
Siblings			-0.041	-0.018	0.153
			(0.268)	(0.261)	(0.228)
Religious			1.625	1.036	0.445
			(1.099)	(1.134)	(0.995)
Econ./Business admin. background			-0.140	0.033	-0.101
			(1.016)	(1.005)	(0.929)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10a: OLS Regressions of *Signed deviation* for participants who received the prediction that they would send more than 25 and up to 50% of their endowment (complete table, part 1 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
Consider				1.817***	1.635***
				(0.354)	(0.313)
Accuracy				-0.170	-0.469
				(0.613)	(0.532)
Experience with AI				0.035	-0.123
				(0.639)	(0.571)
Interaction with AI				0.005	-0.119
				(0.557)	(0.480)
Profession with AI				-0.180	-0.296
				(1.436)	(1.306)
Belief in free will				0.326	0.032
				(0.252)	(0.222)
Reactance				0.032	0.664
				(0.777)	(0.734)
Attitude toward IT				0.410	-0.248
				(0.843)	(0.725)
Locus of control				-1.265	-0.695
				(0.845)	(0.803)
Attitude toward AI				0.354	-0.173
				(0.726)	(0.660)
Risk aversion					0.173
					(0.166)
Altruism					0.033***
					(0.007)
Trust					1.238***
					(0.459)
Belief in DG					0.228***
					(0.032)
Norm in DG					0.169***
					(0.040)
Constant	-3.548***	-3.548***	-6.556***	-11.346	-34.457***
	(0.782)	(0.782)	(2.344)	(7.237)	(7.511)
p(HIL$_p$=AIP)	.	0.963	0.927	0.962	0.985
p(HIL$_p$)	.	0.139	0.080	0.162	0.189
R^2	0.008	0.012	0.035	0.086	0.270
Observations	710	710	710	710	710

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11a: OLS Regressions of *Signed deviation* for participants who received the prediction that they would send more than 25 and up to 50% of their endowment (complete table, part 2 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-4.767**	-6.711**	-7.941***	-8.320***	-5.438**
	(1.854)	(2.977)	(2.934)	(2.949)	(2.397)
AIP	-4.364**	-4.364**	-4.543**	-4.193**	-3.796**
	(1.859)	(1.860)	(1.891)	(1.904)	(1.606)
HIL × Prediction		2.612	3.435	4.066	1.512
		(3.033)	(2.956)	(2.959)	(2.389)
Age			0.147**	0.141**	0.055
			(0.061)	(0.063)	(0.054)
Female			4.479***	4.219***	2.584*
			(1.521)	(1.563)	(1.321)
Attended college			2.631	2.550	2.550
			(2.435)	(2.462)	(1.911)
Undergraduate degree			-0.156	-0.481	-1.702
			(1.969)	(1.968)	(1.606)
Graduate degree or higher			-1.416	-1.554	-2.913
			(2.444)	(2.462)	(2.000)
Student			-1.429	-1.777	-1.141
			(2.443)	(2.568)	(2.125)
Employed			1.296	1.094	-0.426
			(1.960)	(1.982)	(1.592)
More than £25k, up to £50k			-2.934	-3.131*	-2.355
			(1.841)	(1.899)	(1.558)
More than £50k			-3.223	-3.708	-1.868
			(2.395)	(2.418)	(2.035)
Siblings			-0.463	-0.464	-0.103
			(0.492)	(0.493)	(0.385)
Religious			2.984	2.710	1.554
			(1.842)	(1.923)	(1.668)
Econ./Business admin. background			-1.468	-1.271	-0.631
			(1.790)	(1.818)	(1.512)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12a: OLS Regressions of the share sent by dictators to recipients (complete table, part 1 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
Consider				1.140*	0.731
				(0.610)	(0.506)
Accuracy				0.726	-0.445
				(1.013)	(0.814)
Experience with AI				-0.307	-0.263
				(1.080)	(0.935)
Interaction with AI				-0.028	-0.287
				(0.964)	(0.794)
Profession with AI				-0.583	-1.273
				(2.495)	(2.104)
Belief in free will				0.618	0.174
				(0.421)	(0.343)
Reactance				-0.420	1.009
				(1.296)	(1.149)
Attitude toward IT				0.219	-1.297
				(1.568)	(1.259)
Locus of control				-0.964	-0.238
				(1.493)	(1.328)
Attitude toward AI				0.240	-0.673
				(1.211)	(1.019)
Risk aversion					-0.195
					(0.275)
Altruism					0.055***
					(0.013)
Trust					2.349***
					(0.800)
Belief in DG					0.517***
					(0.051)
Norm in DG					0.469***
					(0.073)
Constant	35.826***	35.826***	29.386***	23.607*	-30.348**
	(1.338)	(1.339)	(3.940)	(12.202)	(12.369)
R^2	0.011	0.012	0.046	0.058	0.351
Observations	754	754	754	754	754

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12b: OLS Regressions of the share sent by dictators to recipients (complete table, part 2 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-4.225**	-6.664**	-8.105***	-8.867***	-6.156**
	(1.879)	(3.054)	(3.001)	(3.029)	(2.489)
AIP	-3.344*	-3.344*	-3.899**	-3.694*	-3.315**
	(1.882)	(1.884)	(1.922)	(1.930)	(1.620)
HIL × Prediction		3.284	4.051	4.799	2.468
		(3.116)	(3.034)	(3.050)	(2.499)
Age			0.149**	0.151**	0.079
			(0.061)	(0.063)	(0.055)
Female			4.346***	4.238***	2.942**
			(1.558)	(1.590)	(1.340)
Attended college			2.755	2.723	3.016
			(2.482)	(2.515)	(1.925)
Undergraduate degree			-0.392	-0.766	-2.165
			(1.995)	(1.992)	(1.626)
Graduate degree or higher			-1.956	-2.144	-3.271
			(2.495)	(2.515)	(2.047)
Student			-0.703	-1.345	-0.263
			(2.622)	(2.727)	(2.235)
Employed			0.961	0.804	-0.581
			(2.007)	(2.029)	(1.634)
More than £25k, up to £50k			-2.861	-3.278*	-2.341
			(1.859)	(1.928)	(1.585)
More than £50k			-1.907	-2.706	-0.763
			(2.471)	(2.484)	(2.105)
Siblings			-0.420	-0.463	0.032
			(0.501)	(0.500)	(0.390)
Religious			2.035	1.661	0.776
			(1.899)	(1.974)	(1.682)
Econ./Business admin. background			-0.878	-0.811	-1.070
			(1.864)	(1.889)	(1.596)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 13a: OLS Regressions of the share sent by dictators to recipients for participants who received the prediction that they would send more than 25 and up to 50% of their endowment (complete table, part 1 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
Consider				1.082*	0.772
				(0.627)	(0.521)
Accuracy				0.859	-0.104
				(1.055)	(0.834)
Experience with AI				0.170	-0.140
				(1.134)	(0.974)
Interaction with AI				0.198	-0.200
				(1.003)	(0.814)
Profession with AI				-0.376	-0.822
				(2.603)	(2.190)
Belief in free will				0.538	0.043
				(0.446)	(0.366)
Reactance				-0.571	0.815
				(1.332)	(1.177)
Attitude toward IT				-0.062	-1.391
				(1.658)	(1.314)
Locus of control				-1.733	-0.324
				(1.559)	(1.389)
Attitude toward AI				0.663	-0.238
				(1.263)	(1.079)
Risk aversion					-0.170
					(0.282)
Altruism					0.063***
					(0.012)
Trust					2.218***
					(0.823)
Belief in DG					0.487***
					(0.052)
Norm in DG					0.517***
					(0.070)
Constant	35.803***	35.803***	29.379***	23.976*	-36.387***
	(1.346)	(1.347)	(4.076)	(12.482)	(12.204)
R^2	0.008	0.010	0.039	0.054	0.349
Observations	710	710	710	710	710

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 14a: OLS Regressions of the share sent by dictators to recipients for participants who received the prediction that they would send more than 25 and up to 50% of their endowment (complete table, part 2 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-4.696***	-1.646	-3.572**	-2.842	-2.283
	(1.177)	(1.923)	(1.778)	(1.919)	(1.534)
AIP	-5.391***	-5.391***	-4.691***	-5.032***	-4.911***
	(1.239)	(1.241)	(1.179)	(1.124)	(1.060)
HIL × Prediction		-3.932*	-1.827	-2.806	-2.825*
		(2.164)	(1.883)	(1.957)	(1.589)
Age			0.145***	0.148***	0.099**
			(0.043)	(0.043)	(0.039)
Female			4.170***	4.149***	3.030***
			(0.979)	(0.981)	(0.874)
Attended college			1.375	1.625	1.700
			(1.529)	(1.501)	(1.430)
Undergraduate degree			-0.702	-0.765	-0.400
			(1.362)	(1.305)	(1.168)
Graduate degree or higher			0.715	1.092	0.220
			(1.419)	(1.383)	(1.294)
Student			-5.013***	-4.260**	-3.585**
			(1.877)	(1.787)	(1.623)
Employed			1.462	2.008*	1.334
			(1.262)	(1.205)	(1.094)
More than £25k, up to £50k			-2.503**	-2.248*	-1.928*
			(1.165)	(1.213)	(1.072)
More than £50k			-3.583**	-3.334**	-2.017
			(1.572)	(1.638)	(1.492)
Siblings			-0.339	-0.261	-0.173
			(0.338)	(0.331)	(0.307)
Religious			1.851*	2.331**	2.882***
			(1.092)	(1.047)	(1.012)
Econ./Business admin. background			-8.413***	-8.378***	-6.717***
			(1.573)	(1.557)	(1.242)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 15a: OLS Regressions of the share sent by dictators when prediction was *ex-post* accurate (complete table, part 1 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
Consider				-2.036***	-1.271***
				(0.397)	(0.395)
Accuracy				-0.046	-0.168
				(0.671)	(0.594)
Experience with AI				0.173	0.359
				(0.756)	(0.678)
Interaction with AI				0.332	0.366
				(0.686)	(0.625)
Profession with AI				0.204	0.180
				(1.759)	(1.549)
Belief in free will				0.167	0.102
				(0.331)	(0.291)
Reactance				-0.386	0.817
				(0.767)	(0.772)
Attitude toward IT				-0.821	-0.837
				(0.976)	(0.851)
Locus of control				0.223	0.005
				(1.008)	(0.936)
Attitude toward AI				-0.232	-0.499
				(0.888)	(0.753)
Risk aversion					-0.533***
					(0.200)
Altruism					0.010
					(0.009)
Trust					0.301
					(0.544)
Belief in DG					0.268***
					(0.040)
Norm in DG					0.216***
					(0.055)
Constant	46.789***	46.789***	41.624***	46.733***	15.882*
	(0.617)	(0.618)	(2.810)	(7.765)	(9.270)
R^2	0.041	0.047	0.256	0.312	0.451
Observations	460	460	460	460	460

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 15b: OLS Regressions of the share sent by dictators when prediction was *ex-post* accurate (complete table, part 2 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
HIL	-6.685**	-9.763***	-11.641***	-9.117***	-6.663**
	(3.024)	(3.571)	(3.518)	(3.480)	(3.294)
AIP	-5.491*	-5.491*	-5.660*	-3.694	-2.499
	(3.113)	(3.118)	(3.162)	(3.274)	(2.988)
HIL × Prediction		4.436	6.209*	7.351**	5.877*
		(3.552)	(3.591)	(3.438)	(3.393)
Age			-0.146	-0.081	-0.070
			(0.099)	(0.101)	(0.098)
Female			-0.980	-1.783	-2.196
			(2.319)	(2.354)	(2.232)
Attended college			-0.518	-0.692	2.030
			(3.875)	(3.878)	(3.729)
Undergraduate degree			-0.202	-1.466	-2.748
			(3.135)	(3.023)	(2.813)
Graduate degree or higher			-0.356	-0.862	-1.711
			(4.031)	(3.968)	(3.497)
Student			1.666	2.035	2.974
			(3.687)	(3.619)	(3.370)
Employed			3.095	3.333	0.922
			(3.169)	(3.106)	(2.896)
More than £25k, up to £50k			-2.788	-2.667	-1.301
			(3.206)	(3.174)	(2.800)
More than £50k			-2.539	-3.297	-1.725
			(4.098)	(3.532)	(3.221)
Siblings			0.017	0.492	0.281
			(0.664)	(0.703)	(0.629)
Religious			0.012	-1.611	-4.201
			(3.517)	(3.805)	(3.232)
Econ./Business admin. background			9.477***	10.116***	8.243***
			(3.076)	(3.088)	(2.883)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16a: OLS Regressions of the share sent by dictators when prediction was *ex-post* inaccurate (complete table, part 1 of 2).

Dep. Var.: Share sent	(1)	(2)	(3)	(4)	(5)
Consider				3.748***	1.745*
				(1.000)	(0.970)
Accuracy				-2.499	-1.362
				(1.530)	(1.498)
Experience with AI				-1.655	-2.059
				(1.760)	(1.588)
Interaction with AI				-0.995	-1.540
				(1.518)	(1.376)
Profession with AI				0.226	-1.719
				(3.214)	(3.064)
Belief in free will				0.241	-0.277
				(0.536)	(0.488)
Reactance				1.609	1.645
				(2.094)	(2.122)
Attitude toward IT				4.491*	1.451
				(2.700)	(2.561)
Locus of control				-2.118	-0.392
				(1.941)	(1.927)
Attitude toward AI				0.446	-1.352
				(1.919)	(1.844)
Risk aversion					0.341
					(0.422)
Altruism					0.117***
					(0.034)
Trust					1.894*
					(1.094)
Belief in DG					0.419***
					(0.096)
Norm in DG					-0.138
					(0.117)
Constant	20.180***	20.180***	23.207***	11.796	19.577
	(2.397)	(2.401)	(6.315)	(20.718)	(21.511)
R^2	0.020	0.023	0.077	0.145	0.291
Observations	294	294	294	294	294

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16b: OLS Regressions of the share sent by dictators when prediction was *ex-post* inaccurate (complete table, part 2 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
HIL	-4.977***	-6.531**	-7.951***	-8.479***	-6.494**
	(1.662)	(2.807)	(2.960)	(2.873)	(2.628)
AIP	-3.576**	-3.576**	-3.968**	-3.547*	-3.393*
	(1.759)	(1.761)	(1.863)	(1.879)	(1.766)
HIL × Prediction		2.017	3.053	4.107	2.224
		(2.706)	(2.750)	(2.625)	(2.421)
Age			-0.006	0.004	-0.024
			(0.052)	(0.053)	(0.054)
Female			0.353	0.156	-0.765
			(1.426)	(1.450)	(1.358)
Attended college			2.446	1.996	1.123
			(2.210)	(2.337)	(2.201)
Undergraduate degree			0.085	-1.048	-1.032
			(1.795)	(1.744)	(1.623)
Graduate degree or higher			-1.141	-1.322	-2.136
			(2.406)	(2.276)	(2.092)
Student			0.442	0.804	1.011
			(2.170)	(2.213)	(2.036)
Employed			1.348	0.742	-0.214
			(1.863)	(1.900)	(1.804)
More than £25k, up to £50k			-3.100*	-3.653**	-3.328**
			(1.763)	(1.686)	(1.589)
More than £50k			-1.793	-3.206	-2.878
			(2.439)	(2.348)	(2.272)
Siblings			-0.050	-0.050	0.133
			(0.434)	(0.383)	(0.340)
Religious			0.632	0.161	-1.418
			(1.942)	(1.985)	(1.815)
Econ./Business admin. background			3.072	3.801**	3.363*
			(1.903)	(1.829)	(1.824)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 17a: OLS regressions on the *Signed deviation* of dictators with a high trait reactance (Bruns and Perino, 2023; Hong and Faedda, 1996) (complete table, part 1 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
Consider				2.376***	1.895***
				(0.536)	(0.484)
Accuracy				-0.134	-0.373
				(0.754)	(0.699)
Experience with AI				-1.807*	-1.707*
				(1.033)	(0.941)
Interaction with AI				-1.225	-1.229
				(0.874)	(0.779)
Profession with AI				0.744	0.303
				(1.993)	(1.839)
Belief in free will				0.677*	0.493
				(0.362)	(0.351)
Reactance				-1.228	-0.760
				(1.900)	(1.886)
Attitude toward IT				0.911	-0.497
				(1.515)	(1.439)
Locus of control				-1.514	-0.726
				(1.348)	(1.331)
Attitude toward AI				-0.582	-1.076
				(1.211)	(1.181)
Risk aversion					0.288
					(0.320)
Altruism					0.044***
					(0.012)
Trust					1.158*
					(0.691)
Belief in DG					0.187***
					(0.049)
Norm in DG					0.157**
					(0.062)
Constant	-1.576	-1.576	-1.717	4.953	-14.767
	(1.317)	(1.319)	(3.695)	(12.478)	(12.470)
p(HIL_p=AIP)	.	0.557	0.563	0.626	0.576
p(HIL_p)	.	0.009	0.007	0.013	0.010
R^2	0.028	0.029	0.056	0.146	0.273
Observations	346	346	346	346	346

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 17b: OLS regressions on the *Signed deviation* of dictators with a high trait reactance (Bruns and Perino, 2023; Hong and Faedda, 1996) (complete table, part 2 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
HIL	0.530	-2.151	-2.126	-1.438	-0.502
	(1.511)	(2.215)	(2.213)	(2.238)	(2.034)
AIP	0.301	0.301	-0.038	0.456	0.945
	(1.411)	(1.413)	(1.455)	(1.417)	(1.244)
HIL × Prediction		3.725	3.422	3.948*	3.265
		(2.321)	(2.330)	(2.280)	(2.135)
Age			-0.031	-0.035	-0.073
			(0.051)	(0.055)	(0.049)
Female			0.814	0.757	0.057
			(1.183)	(1.170)	(1.111)
Attended college			-1.203	-1.371	-0.666
			(1.615)	(1.615)	(1.362)
Undergraduate degree			-0.941	-0.976	-1.709
			(1.355)	(1.372)	(1.290)
Graduate degree or higher			-1.946	-2.303	-2.599
			(1.807)	(1.946)	(1.727)
Student			1.780	1.222	1.670
			(1.836)	(1.891)	(1.571)
Employed			-0.360	-0.806	-1.734
			(1.354)	(1.414)	(1.142)
More than £25k, up to £50k			0.976	1.379	1.477
			(1.392)	(1.468)	(1.346)
More than £50k			-0.078	0.446	1.188
			(1.660)	(1.665)	(1.503)
Siblings			-0.151	-0.176	-0.132
			(0.382)	(0.397)	(0.343)
Religious			0.226	0.020	0.215
			(1.310)	(1.377)	(1.201)
Econ./Business admin. background			4.719***	4.444***	4.684***
			(1.435)	(1.497)	(1.384)

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 18a: OLS regressions on the *Signed deviation* of dictators with a low trait reactance (Bruns and Perino, 2023; Hong and Faedda, 1996) (complete table, part 1 of 2).

Dep. Var.: Signed deviation	(1)	(2)	(3)	(4)	(5)
Consider				1.126**	1.002**
				(0.484)	(0.420)
Accuracy				0.201	-0.415
				(0.903)	(0.820)
Experience with AI				0.703	0.556
				(0.845)	(0.797)
Interaction with AI				0.525	0.443
				(0.801)	(0.746)
Profession with AI				-2.056	-1.767
				(2.372)	(2.102)
Belief in free will				0.014	-0.450
				(0.358)	(0.302)
Reactance				2.363	2.409
				(1.615)	(1.479)
Attitude toward IT				1.145	0.980
				(1.109)	(0.985)
Locus of control				0.332	0.255
				(1.188)	(1.121)
Attitude toward AI				0.942	0.121
				(0.897)	(0.764)
Risk aversion					0.098
					(0.206)
Altruism					0.017*
					(0.009)
Trust					1.928***
					(0.666)
Belief in DG					0.277***
					(0.048)
Norm in DG					0.050
					(0.081)
Constant	-3.917***	-3.917***	-3.169	-22.744**	-33.002***
	(1.064)	(1.065)	(3.050)	(9.566)	(12.357)
p(HIL_p=AIP)	.	0.419	0.416	0.230	0.237
p(HIL_p)	.	0.343	0.437	0.147	0.089
R^2	0.000	0.007	0.045	0.073	0.257
Observations	408	408	408	408	408

Robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 18b: OLS regressions on the *Signed deviation* of dictators with a low trait reactance (Bruns and Perino, 2023; Hong and Faedda, 1996) (complete table, part 2 of 2).

B Training the ML Model

The goal was not to collect data exclusively from identical experiments but to obtain a large amount of data that would collectively develop meaningfulness despite individual variations. The applied inclusion and exclusion criteria for data collection are detailed in table 20. To account for variations in experimental designs, dummy variables were included for each study. In cases where different implementations within a study showed statistically significant differences, treatment dummies were also included.

To leverage the power of available software packages, Python was employed for the subsequent steps. Categorical feature values were efficiently handled using one-hot encoding[49]. Missing values were first handled manually, whereby all data rows being either students and/or unemployed who did not specify their income were assigned an income from less than £25.000 a year (an approximation due to students and unemployed people usually earning less money). All other missing values were imputed using an iterative imputer[50]. As a last preprocessing step, any imbalances within the discrete or ordinal feature types were handled using standardization[51].

[49] OneHotEncoder from sklearn.preprocessing (Pedregosa et al., 2011)
[50] IterativeImputer from sklearn.impute (Pedregosa et al., 2011)
[51] StandardScaler from sklearn.preprocessing (Pedregosa et al., 2011)

Citation	**Extracted features**	**relevant observations**
Amir et al. (2012)	age, gender, education, income	195
Antoniades et al. (2018)	age, gender, education	105
von Bieberstein et al. (2021)	age, gender, student, econ	109
Brock et al. (2013)	age, gender, econ	76
Ceccato et al. (2018)	age, gender	348
Dreber et al. (2013)	age, gender	1752
Dunbar et al. (2021)	age, gender	51
Fenzl and Brudermann (2021)	age, gender	318
Fornwagner et al. (2022)	age, education, student, income, religion	780
Goerg et al. (2020)	age, gender	96
Han et al. (2023)	age, gender, education, employment, income	735
Herne et al. (2022)	age, gender, education	131
Inaba et al. (2018)	gender	121
Kandul and Nikolaychuk (2023)	age, gender, education, econ	84
Kause et al. (2018)	age, gender, education	41
Kee et al. (2021)	age, gender education, income	91
Kumar et al. (2021)	age, gender, education, income	539
Lazear et al. (2012)	age, gender, education, siblings, econ	117
Matsumoto et al. (2016)	age, gender, income, siblings	488
Moyal and Ritov (2020)	age, gender	486
Nava et al. (2023)	age, gender, siblings	382
Noussair and Stoop (2015)	age, gender, employment	32
Oleszkiewicz and Kupczyk (2020)	age, education	197
Peysakhovich et al. (2014)	age, gender, religion	970
Powell et al. (2018)	age, gender, student, econ	523
Raihani and McAuliffe (2014)	age, gender, education, income	1153
Rand et al. (2016)	gender	1831
Ritov and Kogut (2017)	age, gender	41
Rotella et al. (2021)	age, gender	299
Stagnaro et al. (2017)	age, gender, education income, religion	2608
Thornton and Aknin (2020)	age, gender	617
Veszteg et al. (2021)	gender	192
Walkowitz (2021)	age, gender, siblings, econ	192

Table 19: Datasets used for training the ML model

Inclusion criteria	Exclusion criteria
Data from an experiment involving a DG Binary game (2 participants)	no free choice of amount min. 15 participants min. 1 demographic variable not peer-reviewed more/less than 2 participants playing against a machine significantly modified DGs (e.g., any amount given to the recipient gets doubled, third-party punishment, etc.) participants' age <16 years

Table 20: Inclusion and exclusion criteria for data collection for the training data of the ML model

C Instructions

Welcome to this Experiment.

You take part in an economic decision experiment. In this experiment, we apply a **strict no-deception policy**. This means that **all instructions provided to you** during this experiment are **truthful**.

The experiment consists of **three parts**:

- **In Part 1,** we will ask you to fill out a pre-experimental questionnaire.
- **In Part 2,** you will make a decision. For this part, you will be randomly and anonymously matched with another participant in this experiment.
- **In Part 3**, we will ask you to complete a questionnaire on a range of opinions and attitudes.

You will receive all necessary instructions directly before the respective part.

Note: Participation in this study is voluntary and the study is anonymous.
Socio-demographic data such as gender, age, etc. are collected solely for the purpose of evaluation in the context of comparing statements between different groups of participants. No attempts will be made to draw conclusions about specific individuals from the information you provide. The results of the evaluation are published exclusively in anonymous form (in tables and/or graphics) so that it is not possible to draw conclusions about individuals.

During this experiment we will use **Points as a currency**. Each Point you earn during the experiment will be **converted to £0.01** at the end of the experiment.

Your payoff for participating in this experiment consists of **two components**.

- The first component is a guaranteed **fixed payoff in the amount of 150 Points for completing the experiment**, including filling out the questionnaires.
- The second component **ranges from 0 to 200 Points and is based on either your decision or the decision of the participant you will be matched with**. You will receive more detailed instructions regarding this component of your payoff in Part 2 of the experiment.

Overall, your payoff will be between 150 and 350 points.

Please answer the following questions about yourself with the answers that best describe you.

What is your age in years?

Which of the following best describes your gender identity?

- Male
- Female
- Other

What is the highest level of education you completed?

- Less than a high school diploma/A-levels
- High school diploma/A-levels
- Attended college/University
- Undergraduate degree (BA/BSc/other)
- Graduate degree (MA/MSc/MPhil/other)
- Doctorate degree (PhD/other)

Are you currently enrolled in a university?

- Yes
- No

Are you currently employed (either as a paid employee or self-employed)?

- Yes
- No

Do you consider yourself religious?

- Yes
- No

Approximately, how much did you personally earn, before taxes, during the past 12 months?

- No more than £25.000
- £25.001 - £50.000
- £50.001 - £75.000
- £75.001 - £100.000
- More than £100.000

How many siblings (including half-siblings) do you have?

Have you studied a subject related to business administration or economics, or are you professionally involved in either of these fields?

- Yes
- No

Thank you for answering the pre-experimental questionnaire.

For this part of the experiment, you will be **randomly and anonymously matched** with another participant. There are **two roles**: the role of **Player A** and the role of **Player B**. Player A decides how to split 200 Points **between Player A and Player B.** Therefore,

Player A can send any amount X between 0 and 200 Points to Player B.

Player B has no decision to make.

The payoff for this part is then determined as follows:

- Payoff Player A = 200 Points – X
- Payoff Player B = X

First, **both participants** will independetly make a **decision in the role of Player A**. After this decision, the **computer will randomly determine which participant's decision in the role of Player A will be implemented** for both participants' payoffs. Therefore, each of the following **two scenarios** is **equally likely** (you can think of this as a **fair** "**coin flip**"):

- Suppose **you are** randomly determined to be **Player A**. Thus, **the participant you are matched with receives the amount of X Points** you chose to send and **your payoff** is given by **200 - X Points**.
- Suppose **you are** randomly determined to be **Player B**. Thus, **you receive the amount of X Points** the participant you are matched with chose to send and the **payoff of the participant you are matched with** is given by **200 - X Points**.

There will be a comprehension question on the contents of these instructions later in the experiment.

An **artificial intelligence tool** was programmed for this experiment. It is based on roughly **15,700 decisions from other experiments**, where participants faced a decision similar to the one in this experiment.

Remember the questions in Part 1: The artificial intelligence tool **predicts decisions** of participants in the role **of Player A based on answers given to these questions**. It can make one of the following **three predictions**:

- "You will send Player B between 0 and 50 Points."
- "You will send Player B between 51 and 100 Points."
- "You will send Player B between 101 and 200 Points."

Compared to a random guess, the artificial intelligence tool is about **85.1** percent more accurate when making predictions.

[NP treatment]

You will not receive a personalized prediction.

[HIL treatment]

Using the data you entered in Part 1, the artificial intelligence tool can provide a personalized prediction for you. In a previous experiment, a **human expert (a person who regularly uses artificial intelligence tools)** was able to test the artificial intelligence tool and decide whether they wanted to provide you with the tool's prediction.

There will be a comprehension question on the contents of these instructions later in the experiment.

[AIP treatment]

Using your answers to the questionnaire in Part 1, you will receive a **personalized prediction from the artificial intelligence tool**.

There will be a comprehension question on the contents of these instructions later in the experiment.

[NP treatment]

Which of the following statements is **incorrect** regarding your decision in this part?

- Every participant will be randomly matched with another participant.
- The artificial intelligence tool will not provide a personalized prediction.
- Each participant will get the payoff corresponding to Player A.
- In the role of Player B there is no decision to make.

[HIL treatment]

Which of the following statements is **incorrect** regarding your decision in this part?

- Each participant will first decide in the role of Player A.
- A human expert was able to test the artificial intelligence tool.
- Each participant will get the payoff corresponding to Player A.
- In the role of Player B there is no decision to make.

[AIP treatment]

Which of the following statements is **incorrect** regarding your decision in this part?

- Each participant will first decide in the role of Player A.
- An artificial intelligence will provide a behavioral prediction for you.
- Each participant will get the payoff corresponding to Player A.
- In the role of Player B there is no decision to make.

[All treatments]

Remember the **artificial intelligence tool** that was programmed for this experiment, based on previous decisions in similar experiments. Compared to a random guess, the artificial intelligence tool is about **85.1 percent** more accurate when making predictions.

[NP treatment]

You will not recieve a personalized prediction

You are now making a decision as **Player A**.
Please decide how many of the 200 Points you want to send to Player B.
I will send Player B: _____ Points

[HIL treatment]

Using the data you entered in Part 1, the artificial intelligence tool can provide a personalized prediction for you. In a previous experiment, a **human expert (a person who regularly uses artificial intelligence tools)** was able to test the artificial intelligence tool and decide whether they wanted to provide you with the tool's prediction.

[HIL$_{np}$ treatment]

> The **human expert** did not find the tool helpful and **decided that you will not be presented with a personalized prediction.**

You are now making a decision as **Player A**. Please decide how many of the 200 Points you want to send to Player B.
I will send Player B: _____ Points

[HIL$_p$ treatment]

The human expert found the tool helpful and decided that you will be presented with your personalized prediction:

[Button: Generate my prediction]

> "You will send Player B [predicted category]."

You are now making a decision as **Player A**.

> As a reminder, your prediction from the artificial intelligence tool that has been provided by a human expert:
> "You will send Player B [prediction category]."

Please decide how many of the 200 Points you want to send to Player B.
I will send Player B: _____ Points

[AIP treatment]

Using your answers to the questionnaire in Part 1, you will receive a **personalized prediction from the artificial intelligence tool**.

[Button: Generate my prediction]

You are now receiving your personalized prediction:

> "You will send Player B [predicted category]."

You are now making a decision as **Player A**. Please decide how many of the 200 points you want to send to **Player B**.

> As a reminder, your predicion that has been provided by the artificial intelligence tool:
> "You will send Player B [predicted category]."

I will send Player B: _____ Points

[All treatments]

Before the end of the experiment, please answer the following questions.

[HIL_{np}/NP treatment]

1. Suppose prior to the decision you made earlier on splitting the 200 Points, you would have received a prediction about your behavior.Do you think the prediction would have influenced your decision?

[No, I have not considered it at all – Yes, I have considered it completely, 5-point scale]
Suppose that the artificial intelligence tool had an improvement of 85.1 percent compared to a random guess. How would you perceive this figure?

[Very low – Very high, 5-point scale]

How much do you agree with the following statement: "I would perceive a prediction about my own behavior rather as advice."

[HIL_{p}/AIP treatment]

Did the prediction influence your decision?

[No, I have not considered it at all – Yes, I have considered it completely, 5-point scale]

Remember that the artificial intelligence tool had an improvement of 85.1 percent compared to a random guess. How do you perceive this figure?

[Very low – Very high, 5-point scale]

How much do you agree with the following statement: "I perceived the prediction rather as advice."

[Strongly disagree – Strongly agree, 5-point scale]

[All treatments]

What do you think the average amount other participants in the role of Player A give in this experiment?

_____Points

For each of the below categories of amounts sent by Player A, state how appropriate or inappropriate they are for society, regardless of your own personal opinion.

[Completely inappropriate – Completely appropriate, 5-point scale]

- Between 0 and 50 Points
- Between 51 and 100 Points
- Between 101 and 200 Points

Do you have previous experience in using artificial intelligences?

[No, I have no experience at all – Yes, I have significant experience, 5-point scale]
How often (roughly) do you interact with generative artificial intelligence like chatGPT?

- Daily
- Weekly

- Once per month
- Hardly ever
- Never

Do you have a profession or education that is closely related to programming, robotics, or artificial intelligence?

- Yes
- No

Please indicate to what extent you agree with the following statements.

[Strongly disagree, Disagree, Neutral, Agree, Strongly agree]

- I find contradicting others stimulating.
- It makes me angry when another person is held up as a role model for me to follow.
- Regulations trigger a sense of resistance in me.
- When something is prohibited, I usually think, "that's exactly what I am going to do".
- I consider advice from others to be an intrusion.
- I become frustrated when I am unable to make free and independent decisions.
- It irritates me when someone points out things, which are obvious to me.
- I become angry when my freedom of choice is restricted.
- Advice and recommendations usually induce me to do just the opposite.
- I resist the attempts of others to influence me.
- When someone forces me to do something, I feel like doing the opposite.

Please indicate to what extent you agree with the following statements.
[Completely disagree – completely agree, 5-point scale]

- I generally trust others.
- I generally have faith in others.

- I feel that others are generally well-meaning.
- I feel that others are generally trustworthy.

Please indicate to what extent you agree with the following statements.
[Completely disagree – completely agree, 5-point scale]

- If I heard about a new information technology, I would look for ways to experiment with it.
- Among my peers, I am usually the first to try out new information technologies.
- In general, I am hesitant to try out new information technologies.
- I like to experiment with new information technologies.

Please indicate to what extent you agree with the following statements.

- I am my own boss.
- If I work hard, I will succeed.
- Whether at work or in my private life: What I do is mainly determined by others.
- Fate often gets in the way of my plans.

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?
Please tick a box on the scale, where the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'.

Please indicate to what extent you agree with the following statement: "I fully believe I have free will." Please tick a box on the scale, where the value 0 means: 'I fully disagree' and the value 10 means: 'I fully agree'.

Part 3: Final Questionnaire (8/10) Imagine the following situation: Today you unexpectedly received £1,000. How much of this amount in Pounds would you donate to a good cause?

£_____

How willing are you to give to good causes without expecting anything in return?

[Not willing at all – Very willing, 5-point scale]
Please indicate to what extent you agree with the following statements.

[Completely disagree – completely agree, 5-point scale]

- I fear artificial intelligence.
- I trust artificial intelligence.
- Independent of artificial intelligence, please click on the most approving button for this statement.
- Artificial intelligence will destroy humankind.
- Artificial intelligence will benefit humankind.
- Artificial intelligence will cause many job losses.

Please indicate to what extent you agree with the following statements.

[Completely disagree – completely agree, 5-point scale]

- I mind when a website uses my personally identifiable information to customize my browsing experience.
- I mind when a website uses cookies to customize my browsing experience. (A cookie is information that a website puts on your hard disk so that it can remember something about you later)
- I mind when a website uses my purchasing history to personalize my browsing (e.g., by suggesting products for me to purchase)
- I mind when a website uses my personally identifiable information for marketing or research activities.
- I mind when a website monitors my purchasing patterns.
- I mind when a website that I visit collects information about my browsing patterns without my consent.

- I mind when a website that I visit collects information about my browser configuration without my consent.
- I mind when a website that I visit collects information about my IP address (a number that uniquely identifies your computer from all other computers on the Internet) without my consent.
- I mind when a website that I visit collects information about the type of computer/operating system I use without my consent.
- I mind when a website that I visit records the previous website I visited.
- I am concerned about unauthorized employees getting access to my information.
- I am concerned about unauthorized hackers getting access to my information.

Please press the "Finish Experiment" button to see your payoff results.

Thank you for participating in our experiment.

You have been randomly selected as **Player [1/2].**

Fixed payoff: 150 Points
Decision-based payoff: [...] Points
Total payoff: [...] Points converted to **£[...]**
Your payment will be automatically handled via prolific.
In order to be recognized, you need to return to prolific.

[Button: Return to Prolific]

D The Stage 1 Experiment

Prior to the main experiment described in this work, a short preliminary study has been conducted. The aim of this Stage 1 survey was twofold. First and foremost, it was about obtaining the decision necessary for the HIL treatment of the main experiment. This required people who regularly interact with AI to make the decision of whether to provide the participants in the main experiment with the prediction of the provided tool. Secondly, this

study aimed to check whether the instructions for the main experiment were comprehensible. In the following, the design, the procedures, and the results of the Stage 1 survey are presented.

D.1 Design

The Stage 1 survey consisted of three main parts. First, participants read the instructions for the main experiment. Then, they were able to test the AI tool before deciding whether they would provide the prediction to participants in the main experiment. Lastly, they were asked to complete a questionnaire on a range of questions regarding demographics, attitudes, and their background concerning AI. Participants in the Stage 1 survey have been paid a flat fee of £2.50 and their decision did not have an influence on their payoff.
In Part 1, participants read the instructions for the main experiment. They have been informed that these instructions refer to an experiment that will be conducted with other participants at a later date. Special emphasis was placed on the fact that the instructions are not relevant to their payment and do not describe the decision they will have to make later in the Stage 1 survey. The instructions have been followed by a comprehension question. Failing the comprehension question twice led to the exclusion of the respective participant.
In Part 2 of the Stage 1 survey, participants were able to test the AI tool that can predict behavior in the DG based on demographic data by providing one of the following predictions:

- "You will send Player B between 0 and 50 Points."
- "You will send Player B between 51 and 100 Points."
- "You will send Player B between 101 and 200 Points."

Participants were able to enter different combinations of demographics and generate the according prediction as many times as they liked. After testing the tool and moving on to the next page, participants asked whether they wanted to provide the prediction tool to participants in the experiment described in Part 1. The exact wording was: "Do you want to provide the participants in the experiment described in Stage 1 with the tool you tested in Stage 2?" Following, their understanding and the clarity of the instructions to the main experiment has been queried. Therefore, a clarity scale with four items was used. On a scale from 1 to 5 (1: Fully disagree, 5: Fully agree), participants were asked to state how much they agree with the following statements:

- "The instructions for this experiment were clear to me."

- "The instructions for the experiment described in Stage 1 were clear to me."
- "The instructions for the experiment described in Stage 1 should be clear to the average person."
- "The purpose of the overall experiment (this experiment and the experiment of Stage 1) is clear to me."

To be able to make a distinction between people who regularly interact with AI and can, therefore, be called experts, three questions regarding their background using AI have been included.

- I use chatGPT (5 choices: Never, I have used it once, Every month, Every week, Every day)
- I use other artificial intelligence tools (5 choices: Never, I have used it once, Every month, Every week, Every day)
- For the following statement, please select how much you agree with it. (Scale from 1 to 5, 1: Fully disagree, 5: Fully agree)

The Stage 1 survey then concluded with a survey asking participants about their demographics (age, gender identity, employment status), their previous participation in a similar experiment, and whether they have a profession or education that is closely related to programming, robotics, or AI.

D.2 Procedures

We conducted the above-described design as an online survey. The survey was programmed in oTree (Chen et al., 2016) and hosted on a Heroku server. Using the research platform Prolific (Palan and Schitter, 2018), twelve participants were recruited, and the survey was conducted in June 2023. Participants were able to write messages to the experimenter using the messaging function provided by Prolific. Anonymity was ensured as no personally identifiable information was gathered. Participants in the Stage 1 survey were excluded from participating in the main experiment. Participants were pre-selected to be located in either the USA or the UK and to be between 18 and 65 years of age. The average time participants spent on the survey until completion was 6 minutes and 7 seconds, and each participant who finished the survey received a flat fee of £2.50. Participants received their payoff through Prolific.

D.3 Results

Out of the twelve participants recruited for the Stage 1 survey, six failed the comprehension question at least twice and were thus excluded from completing the survey. The average age of participants finishing the survey was 33.17 years. Five participants identified as males and one as female. A summary of the results is provided in Table 21.

#	Age	Gender	Provide prediction	Regularly using chatGPT	Regularly using other AI tools	Expert	Clarity of instructions
1	29	female	yes	4	3	✓	3.00
2	23	male	yes	1	1	✗	4.75
3	36	male	yes	4	1	✓	5.00
4	26	male	yes	1	2	✗	4.75
5	35	male	no	4	1	✓	5.00
6	50	male	no	1	1	✗	3.75

Note: Options regarding usage of chatGPT or other AI tools were 1: Never, 2: I have used it once, 3: Every month, 4: Every week, and 5: Every day. Experts are participants who either use chatGPT or other AI tools at least every week. The clarity of instructions measure is calculated as the mean of the scale values regarding the four clarity statements listed in the design description.

Table 21: Participants in Stage 1 and Selections of Human Experts

The clarity measure was calculated as the mean of the stated scale values regarding the four clarity statements listed above. Overall, the mean of a clarity measure was 4.38, indicating relatively high comprehensibility and clarity of the instructions provided. The answers to a free field question regarding the suspected purpose were also relatively close to the actual purpose of the main experiment (e.g., "The impact of the AI tool's prediction upon participants' decisions." or "To confirm if using the AI predictor changes outcomes compared to no AI predictor used." Three out of the six participants who successfully completed either use chatGPT or other AI tools at least every week and can, therefore, be referred to as experts according to our definition in the instructions of the main experiments. Two of the participants regularly using AI tools decided to provide the tool to the participants in the main experiment, and one decided against it.