Artificial Intelligence, Ethics, and Diffused Pivotality

Victor Klockmann, Alicia von Schenk, Marie Claire Villeval

GATE LYON SAINT-ÉTIENNE
UMR 5824

93, chemin des Mouilles
69130 Ecully - France

Maison de l'Université, Bâtiment B
10, rue Tréfilerie
42023 Saint-Etienne cedex 02 - France

http://www.gate.cnrs.fr
gate@gate.cnrs.fr

WP 2111 – May 2021

Artificial Intelligence, Ethics, and Diffused Pivotality

Victor Klockmann, Alicia von Schenk, Marie Claire Villeval

Abstract:

With Big Data, decisions made by machine learning algorithms depend on training data generated by many individuals. In an experiment, we identify the effect of varying individual responsibility for moral choices of an artificially intelligent algorithm. Across treatments, we manipulated the sources of training data and thus the impact of each individual's decisions on the algorithm. Reducing or diffusing pivotality for algorithmic choices increased the share of selfish decisions. Once the generated training data exclusively affected others' payoffs, individuals opted for more egalitarian payoff allocations. These results suggest that Big Data offers a "moral wiggle room" for selfish behavior.

Keywords:
Artificial Intelligence, Pivotality, Ethics, Externalities, Experiment

JEL codes:
C49, C91, D10, D63, D64, O33

Artificial Intelligence, Ethics, and Diffused Pivotality

Victor Klockmann[a] Alicia von Schenk[b] Marie Claire Villeval[c]

May 18, 2021

Abstract

With Big Data, decisions made by machine learning algorithms depend on training data generated by many individuals. In an experiment, we identify the effect of varying individual responsibility for moral choices of an artificially intelligent algorithm. Across treatments, we manipulated the sources of training data and thus the impact of each individual's decisions on the algorithm. Reducing or diffusing pivotality for algorithmic choices increased the share of selfish decisions. Once the generated training data exclusively affected others' payoffs, individuals opted for more egalitarian payoff allocations. These results suggest that Big Data offers a "moral wiggle room" for selfish behavior.

Keywords: Artificial Intelligence, Pivotality, Ethics, Externalities, Experiment

JEL Codes: C49, C91, D10, D63, D64, O33

We are grateful to Ferdinand von Siemens, Matthias Blonski, Michael Kosfeld and seminar participants at the Goethe University Frankfurt and GATE for useful comments. Financial support from the Goethe University Frankfurt, Leibniz Institute for Financial Research SAFE, and the LABEX CORTEX (ANR-11-LABX-0042) of Universite de Lyon, within the program Investissements Avenir (ANR-11-IDEX-007) operated by the French National Research Agency (ANR) is gratefully acknowledged.

[a]Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany. Center for Humans & Machines, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. klockmann@econ.uni-frankfurt.de.

[b]Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany. Center for Humans & Machines, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. vonSchenk@econ.uni-frankfurt.de.

[c]Univ Lyon, CNRS, GATE UMR 5824, 93 Chemin des Mouilles, F-69130, Ecully, France. IZA, Bonn, Germany. villeval@gate.cnrs.fr.

1 Introduction

The amount of data created every day has been increasing steadily in the 21st century. The World Economic Forum estimates that by 2025, 463 exabytes or 463,000,000 terabytes of data, an equivalent of 212,765,957 DVDs, will be generated on a daily basis.[1] In the same spirit, by 2018, 90% of all data in the world have been generated between 2017 and 2018, and the tendency is rising.[2] The broad availability of massive data opens up new possibilities to program and train algorithms that form decisions based on a huge number of individual observations or data points. Despite undisputed benefits, such as higher accuracy and prediction quality due to extensive data sources for improving the algorithms, this study asks whether human behavior that serves as a source of training for artificial intelligence adapts to the fact of making up a negligible share of all observations. Already more than 50 years ago, Darley and Latane (1968) described the "bystander effect". This effect characterizes the tendency of people to underestimate their pivotality for certain outcomes, act in a self-serving way while possibly hurting others, and rationalize their selfish behavior *ex post* in the sense of "if I don't do it, someone else will". More recently, Bénabou et al. (2018) explained how individuals build narratives of not being pivotal to maintain a positive self-image while acting in a morally questionable manner. We can therefore legitimately wonder whether, in the debate on artificial intelligence and ethics, the absence of pivotality of individuals for the created training data contributes to the emergence of algorithms that make less ethical or prosocial decisions.

A particular noteworthy problem in the context of artificial intelligence (AI) and new technologies is the concept of "many hands". AI systems have histories and a multiplicity of individuals determines the outcome of an AI's prediction and decision. Hereby, it is very difficult to track all humans involved in the history of one particular technological action and to attribute responsibility to one single person. Neither the programmer of the algorithm, nor the first user, nor any individual generating training data might be solely responsible for a given outcome, but all of them constitute interconnected elements that contribute to the AI's choices. Coeckelbergh (2020) describes obstacles when attributing moral responsibility in human-machine interaction in his philosophical perspective on artificial intelligence ethics. On the one hand, the system itself cannot be made responsible for possible undesired moral consequences since it lacks consciousness, moral reasoning, free will, or emotions. On the other hand, it is difficult to make the engineers designing the AI fully accountable for these consequences. One could claim that an agent must be fully aware of her actions and her actions' consequences. But if more than one person trains an AI system, how do individuals take the consequences of their behavior for the algorithm's choices into consideration?[3]

[1] www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

[2] www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[3] This responsibility is all the more important as once artificial intelligent algorithms are involved in the process of problem-solving, individuals may tend to fully rely on the AI for making a decision. This is a

This paper analyzes the causal effect of changing individuals' pivotality for an algorithm's decisions on social behavior. The first question we seek to answer is to what extent the introduction of Big Data affects the selfishness or prosociality of individuals' decisions. In particular, we compare a situation in which an individual is the only source of training data for an algorithm with situations in which there are one or 99 additional sources of training. Increasing the number of the sources of training may create a "moral wiggle room" for the individuals who face a trade-off between selfish interest and social preferences. This is because the relative contribution of the individual's behavior to the training data diminishes, and so does the feeling of individual responsibility. We thus ask if individuals make use of such moral wiggle room offered by Big Data at their own advantage. The second aspect of our research question is whether individuals behave differently when being fully pivotal for an algorithm that takes decisions that will impact a third party rather than themselves. Does the prosociality of training data they claim for others match the standards they implement for decisions affecting their own payoff?

By addressing these questions, we contribute to several strands in the literature. First, our study complements the literature on artificial intelligence, ethics and algorithmic biases (Anderson and Anderson, 2007; Bostrom and Yudkowsky, 2014; Bonnefon et al., 2016; Awad et al., 2018; Lambrecht and Tucker, 2019; Rahwan et al., 2019). This literature has insisted on the importance of involving public morality in designing machine ethics that cannot be derived from the technical rules of robotics. In this perspective, researchers have collected large-scale data on how citizens prefer that AI solves moral dilemmas in the context of autonomous vehicles, (*e.g.*, Bonnefon et al., 2016; Awad et al., 2020). Compared to these investigations that elicit individuals' reported preferences, the novelty of our study is using revealed preferences through monetary choices to analyze to what extent individual responsibility in generating training data for an algorithm that makes consequential predictions influences behavior. These studies also do not explore whether building the morality of AI on the views of a large number of citizens comes along with decreased individuals' responsibility when their preferences are aggregated with those of many others to train the AI. Moreover, instead of considering ethics in terms of moral dilemmas, we approach ethics in terms of social preferences. This approach responds to the call for broad scientific research agendas to better understand machine behavior, including their implications in terms of fairness and accountability (Rahwan et al., 2019).

Second, our focus on training constitutes a novelty relative to the experimental literature on human-machine interactions (see, for a recent survey of this literature, Chugunova and Sele, 2020). In this literature, robots were mainly introduced to isolate the role of social preferences when studying strategic decision making (*e.g.*, Houser and Kurzban, 2002; Ferraro et al., 2003; Yamakawa et al., 2016), to create a social preferences vacuum chamber

frequently raised critic in the domain of predictive justice, for example, where conformity might replace the judge's personal conviction.

(Benndorf et al., 2020), to eliminate the opponent's strategic behavior in a competition (Houy et al., 2020), or to isolate the role of social incentives in teams (Corgnet et al., 2019). In these studies, robots made decisions based on predetermined rules or choices, while in our experiment the AI learned from the current players' actions and individuals were aware of training an algorithm. We also differ from studies on auctions investigating how bidding behavior is affected by interacting with a computer that adapts to traders' behavior (Van den Bos et al., 2008; Ivanov et al., 2010; Teubner et al., 2015). In our case, there was no strategic interaction between the players and the AI. It has also been found that interacting with a machine harms the ethical behavior of individuals in a cheating task because of image concerns (Cohn et al., 2018). In contrast, our algorithm was not passive, it learned and made a prediction that impacted the players' payoffs. Through its focus on pivotality, this study complements our companion paper (Klockmann et al., 2021) in which we explored how the intergenerational transmission of training data for an AI influenced individuals' behavior. The current study differs from the other one in that, instead of looking at how individuals' decisions affect the well-being of future generations, it examines how they are affected when individuals know that their AI training data are aggregated with those of other decision makers. By focusing on different channels of responsibility, both studies contribute to explaining how individual responsibility shapes the development of moral algorithms.

Third, we contribute to the literature on the role of pivotality in decision making. For example, Bénabou et al. (2018) showed theoretically how individuals use narratives in a moral context to downplay the externalities of their actions, or to pretend not being decisive in final outcomes. Falk et al. (2020) provided experimental evidence that individuals tend to rely on the narrative that their actions do not influence an outcome, when it is available. In group settings, when one member's choice is sufficient to trigger an immoral action, individuals are more willing to act in self-serving ways, compared with individual settings. In a study on collective decision making processes, Bartling et al. (2015) reported that individuals vote strategically to avoid being pivotal for an unpopular voting result. By designing treatments in which individuals were either not pivotal, fully pivotal, or shared responsibility equally in their contribution to the training data of an algorithm, our study follows up on this idea in the context of AI.

Finally, our study adds to the literature that examines the potential discrepancy between the moral guidelines individuals claim and the principles they want to be applied to themselves. The "Not in my backyard" literature has shown, notably in the context of nuclear waste repositories, that individuals are in favor of moral or socially desirable behavior and policies, as long as they are not harmful to themselves (Frey et al., 1996; Frey and Oberholzer-Gee, 1997). In the domain of AI, the studies reported in Bonnefon et al. (2016) revealed that individuals preferred utilitarian self-driving cars – that is, autonomous vehicles that would sacrifice their few passengers to save a larger number of pedestrians – only if they

were not themselves the car driver. Once they were asked what type of vehicle they would buy, individuals preferred to ride in an autonomous vehicle that protects its passengers at all costs. We contribute to this reflection by asking whether such discrepancy also exists in the context of training AI algorithms that make moral decisions that exclusively affect others or that also influence ones' outcomes.

To answer our questions, we designed a laboratory experiment in which the ethics of AI was addressed through monetary allocation choices that can be either selfish or altruistic. After completing real effort tasks to generate a predetermined endowment, participants were randomly paired with another participant to play a series of 30 binary dictator games inspired by Bruhin et al. (2019). Like in Klockmann et al. (2021), the repeated allocation decisions of the dictator in each pair were used to train a standard machine learning algorithm. This algorithm had to predict a hypothetical allocation choice of the dictator in a 31^{st} period. This prediction determined the payoffs of the dictator and the receiver in this period. The participants' earnings in the experiment consisted in equal parts of their payoff in one randomly selected period among the first 30 ones and of the payoff determined by the algorithm's prediction in period 31. This made the training data meaningful in terms of incentives.

Four treatments varied the sources of the algorithm's training data. These variations allowed us to manipulate the pivotality of individuals' behavior in the training of the AI, creating or removing the moral wiggle room for players to justify selfish actions. In the *Full Pivotality treatment*, which represents our baseline condition, the dictators' decisions were the unique source of training of the algorithm. In contrast, in the *No Pivotality treatment*, the dictator's 30 decisions were pooled with the 30 choices of 99 other dictators in the experiment. That is, the dictator was responsible for only 1% of the training data. The comparison between the Full Pivotality and the No Pivotality treatments allowed us to test whether individuals behaved less selfishly and their choices exhibited more social preferences when they were fully pivotal for the AI training.

In the *Full Pivotality-Others treatment*, the dictators' decisions did not train an algorithm in their own group, but dictators were fully pivotal for the AI training in another pair of players in the same session. Likewise, the algorithm in the own group was trained with data generated by another dictator. Finally, the *Shared Pivotality treatment* built a link between the Full Pivotality and the Full Pivotality-Others treatments. In this treatment, the dictators from two pairs trained a single algorithm that decided about payoff allocations in period 31 in both pairs. Compared to the two previous treatments (Full and No Pivotality), these treatments introduced a responsibility of the individuals for the payoffs of another pair of participants. In contrast to the Full Pivotality treatment, pivotality in the Shared Pivotality treatment was diffused since both dictators shared responsibility for the outcome in both pairs. Comparing these two treatments enabled to test whether reducing pivotality increased the selfishness of the training data. The comparison between the Shared Pivotality

and the Full Pivotality-Others treatments informed on whether being exclusively responsible for the outcome of another pair of players eliminated the wiggle room arising from reduced pivotality.

Our findings revealed that pivotality in generating training data for the AI has a crucial influence on individuals' behavior. Dictators in the No Pivotality treatment, who accounted for a negligible part of the algorithm's training data, and those in the Shared Pivotality treatment in which responsibility for the algorithmic choices was shared among two dictators, behaved significantly more selfishly than their counterparts in the Full Pivotality baseline. The main effect of diffused pivotality was observed when moving from a share of 100% to 50% of contribution to the training data. The likelihood of choosing the option that increased the dictator's payoff at the detriment of the receiver significantly increased and the estimated social preferences parameters significantly decreased when pivotality diffused. Moreover, when dictators were fully pivotal for training an AI whose predictions impacted the payoffs of a pair other than their own, they made more egalitarian payoff allocations than dictators who generated training data that also partially influenced their own outcome. An additional exploratory analysis exploring the dictators' beliefs about other dictators' behavior revealed that the selfishness of their decisions increased when dictators believed that the AI was trained with the data of other selfish dictators.

Overall, our results show that when machines are trained by a multiplicity of agents, individuals may use the moral wiggle room to behave more selfishly that is offered by diffused pivotality, attenuating their responsibility for unethical technological actions. One implication of these findings is a need for designing machines that embed explicit ethical guidelines and basic fairness principles, rather than using a simple aggregation of revealed preferences. At the same time, these findings raise a challenging question for democracies on how to involve citizens in the design of artificially intelligent machines that will decide for them in the future, without the moral weakening and increased social distancing permitted by the diffusion of pivotality.

The remainder of this paper is organized as follows. The next section describes our experimental design and procedures. We formulate behavioral predictions in section 3. Section 4 presents the experimental results. Section 5 offers concluding remarks.

2 Design and Procedures

In this section, we first present the design of the experiment[4] and our treatments. Afterward, we describe our recruitment methods and experimental procedures.

[4]This subsection that presents the general features of our game is close to the one in our companion paper (Klockmann et al., 2021)

2.1 Design

The experiment comprised three parts. In part 1, participants completed a set of real effort tasks. Part 2 comprised two stages. In the fist stage participants played several mini-dictator games. The second stage comprised the prediction of the machine learning algorithm based on observed participants' choices in the role of dictators in the first stage. Meanwhile, we elicited dictators' beliefs about other dictators' behavior. In part 3, we elicited sociodemographic and other individual information.

Part 1 In part 1, each participant completed five tedious real effort tasks that comprised finding a sequence of binary numbers within a matrix of zeros and ones (Figure B.10 in Appendix B). Participants were informed upfront that completing these tasks would earn them 1200 points that would be used in the second part of the experiment. The objective was to generate a feeling of entitlement without introducing any earnings inequality in this part. On average, participants spent approximately 90 seconds per task.

Part 2 In the first stage of part 2, participants played 30 mini-dictator games, one per period (see Appendix B for an English translation of the instructions). Each participant was anonymously paired with another participant and matching remained fixed for this part. One of the two pair members was randomly assigned the role of the dictator ("participant A" in the instructions) and the other one the role of the receiver ("participant B"). Roles were kept fixed throughout the experiment. The dictator's task in each period was to decide between two options on how to split points between herself and the passive receiver. All participants were informed upfront that these decisions would later serve as training data for a machine learning algorithm that would make a prediction in a later stage that could affect their earnings.

In each period, the dictator could choose one of two possible unique payoff allocations, $X = (\Pi_X^1, \Pi_X^2)$ or $Y = (\Pi_Y^1, \Pi_Y^2)$, for which the sum of all four payoffs was kept constant in each period. The dictator's amount was always weakly higher with option X (the "selfish option") and weakly smaller with option Y (the "altruistic option"). Because both pair members had to complete the same tasks to generate the group endowment, opting for the selfish option X when the receiver would be strictly better off with option Y would indicate that the dictator takes advantage of her exogenously assigned power of decision.

Across the 30 games, we systematically varied the payoffs in the two options to manipulate the inequality between pair members, the sum of payoffs in each option, and whether the dictator was in an advantageous or disadvantageous relative position in the pair. The order of the 30 games was random, but fixed for all the participants. The calibration of payoffs was inspired by Bruhin et al. (2019). Figure 1 illustrates the dictator games and represents each game by a solid line that connects option X and option Y. Table C1 in the

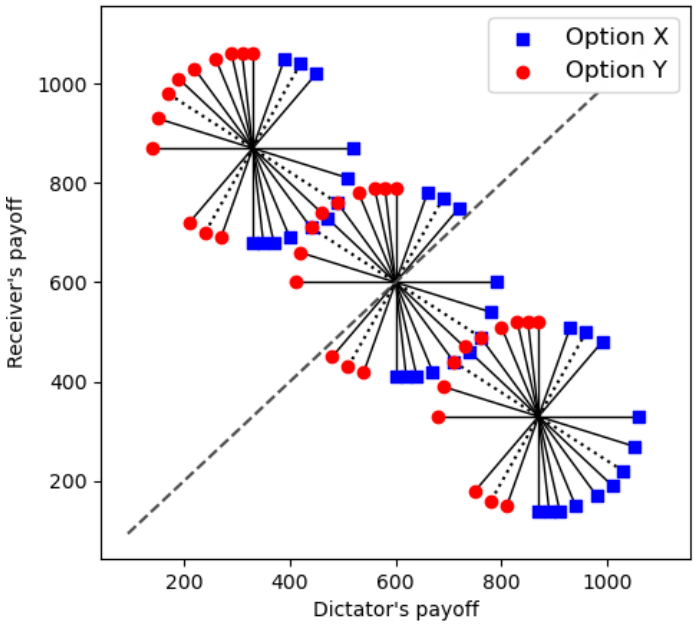

Figure 1: Dictator Games

Notes: Each circle represents 12 binary dictator games. Each game is represented by a line that connects option X (in blue) and option Y (in red). The dotted lines correspond to games that were not presented to the participants. Rather, one of them was picked at random for the AI's prediction and decision. The slope of each line represents the cost for the dictator to increase the receiver's payoff. In the top-left circle, both options in each game represent disadvantageous inequality for the dictator. In the bottom-right circle, both options in each game represent advantageous inequality for the dictator. In the middle circle, the relative position depends on the chosen option.

Appendix C lists all pairs of options.

We categorized the decisions along two dimensions. First, there were games in which option X Pareto-dominated option Y, games in which the receiver or the dictator was indifferent between both choices, and games in which the dictator got a strictly higher payoff with option X while the receiver was monetarily better off with option Y. Second, either the receiver or the dictator was better off than the other pair member in both alternatives, or the dictator's decision determined whose payoff was higher. The aim of these variations was to identify the participant's distributional preferences.[5] At the end of this stage, one of the 30 decisions was picked at random and this determined the payoff of both the dictator and receiver in this stage.

In the second stage, there was a 31st pair of options X and Y. But instead of the dictator choosing one of the two options, there was a random forest algorithm used as a standard supervised classification method making the choice (see Appendix A for details). Participants received detailed information on the concept of machine learning and classification in an

[5]This design created a large variety of situations while still using a simple game. Alternatively, we could have used strategic games, such as trust games or ultimatum bargaining games, which would have allowed us to also examine reciprocity behavior. But for the purpose of the current study, a simpler game, such as the dictator game, was sufficient to test our conjectures.

Scenario	Option X Selfish	Option Y Altruistic	Preference type
1	(870, 140)	(870, 520)	Efficiency/Altruism
2	(1050, 270)	(690, 390)	Fairness
3	(670, 420)	(530, 780)	Selfishness

Table 1: Elicited Beliefs about Other Dictators' Behavior

Notes: All dictators except those in the Full Pivotality treatment were asked to assess whether an AI trained with the other dictators' choices would select option X or option Y in the three listed decision scenarios. The other participants were those whose data also trained the AI that determined the payoff in the current pair, as explained in subsection 2.2. The fourth pair of options corresponded to the randomly chosen binary dictator game for the AI's prediction that was paid out to the current pair (*i.e.*, one of the menus of options shown in Table C2 in Appendix C.

information box included in the instructions (see Appendix B). The exact functionality of the algorithm was not crucial for the research question and, as it was kept constant across conditions, it did not affect the treatment differences. The focus rather is the training of AI as an a priori neutral technology with behavioral data.

As explained to the participants before they made their first stage decisions, in the baseline condition that we call the Full Pivotality treatment (see below) the algorithm used the 30 decisions of the dictator as training data to make an out-of-sample prediction of how the dictator would have decided in period 31 when facing a new pair of options. The machine learning tool did not build models or estimate parameters; it simply predicted from patterns in the data. We used the payoffs and the sum and difference of points allocated to the players in the chosen and rejected options as features for classification. For the decision of the AI, one of the six games represented by a dashed line in Figure 1 was chosen at random. Table C2 in Appendix C lists all six possible out-of-sample decisions of the AI.

After the dictators made their decisions but before they received feedback on the AI's choice in period 31, we elicited the dictators' beliefs about other past dictators' training and decisions of the AI, except in the Full Pivotality treatment for a reason that will become clear in subsection 2.2. Participants had to assess whether the random forest algorithm would choose option X or option Y in four decision scenarios. The alternatives in the fourth scenario corresponded to the actual menu of options the AI faced in the 31st period, as shown in Table C2 in Appendix C. The other three pairs of options are summarized in Table 1. In the first decision, option Y was Pareto-dominant. While selfish dictators should be indifferent between the two options since both paid the same to the decision-maker, preferences for efficiency or altruism should motivate the choice of option Y. In the second decision, option Y increased the receiver's payoff and decreased inequality but at a cost for the dictator. In the third decision, switching from option X to option Y put the dictator in a disadvantaged position, but reduced the absolute difference in payoffs between the two pair members. Belief elicitation was incentivized: each accurate answer paid 100 points to the participant.

Part 3 In the last part we collected sociodemographic information such as gender, age, field and length of study, average weekly spending, and school graduation grade (Abitur). We also collected information about the participants' familiarity with artificial intelligence and machine learning and their confidence in these technologies. We asked questions about the individuals' satisfaction with the AI prediction in period 31, and how accurate the implemented decision reflected the dictator's preferences. Finally, one question assessed the participants' understanding of the functionality of the random forest algorithm.

2.2 Treatments

The experiment comprises four between-subjects treatments that vary the relative impact of the dictator's decisions in a pair on the AI's training and future prediction. These treatments are called Full Pivotality, Shared Pivotality, Full Pivotality-Others, and No Pivotality. Figure 2 presents a simplified overview of these treatments.

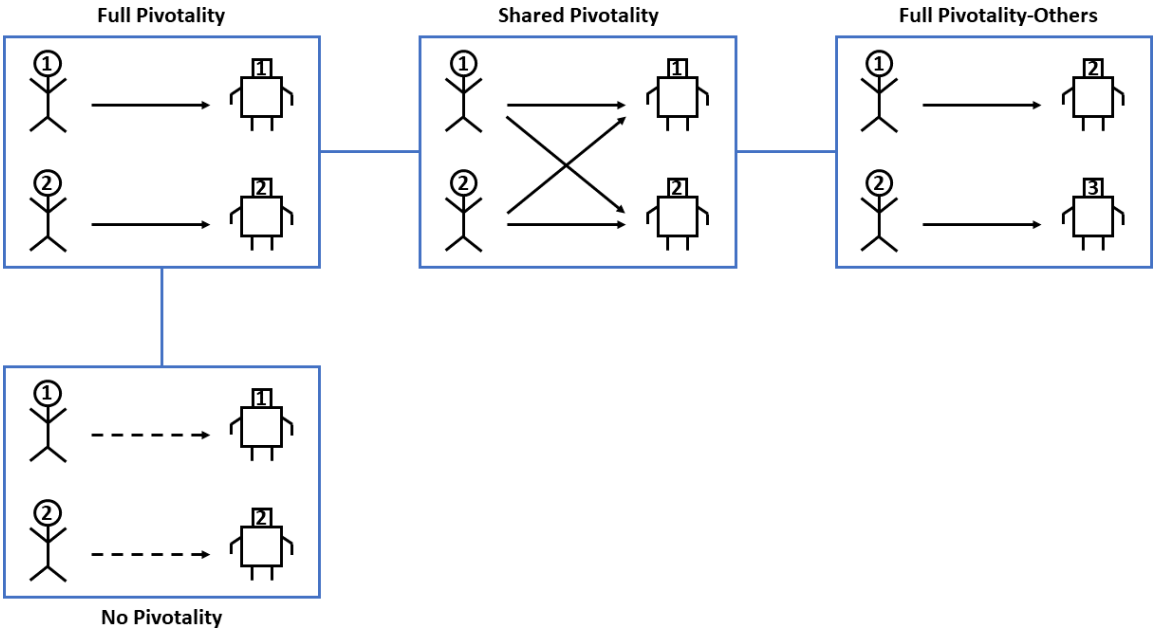

Figure 2: Illustration of Treatments

Notes: The figure shows a simplified overview of the treatment variations with only two dictators and two pairs. The stickmen on the left of each block represent dictators with their pair number. The robots on the right represent an AI algorithm that determines the payoffs of the respective pair. A solid arrow stands for pivotal influence on the training data; the dashed arrow in the No Pivotality treatment depicts the individual's negligible impact on AI's training due to pooling with data from 99 other dictators.

Our baseline condition is the *Full Pivotality treatment*. It exactly follows the above described experimental design.[6] The random forest algorithm built exclusively on the 30 choices of the respective dictator for its predictions, and these data were only used for

[6]This baseline treatment, and only this one, is common to this paper and our companion paper (Klockmann et al., 2021)

determining payoffs in the 31st period. There was no connection between pairs in this treatment. Therefore, each dictator in a pair was fully responsible for the AI's training and its subsequent payoff consequences for the pair. Figure 2 depicts this feature by means of solid arrows that connect each dictator with a robot that decides about the payoff in the dictator's pair.

In the *No Pivotality treatment*, the training and the allocation decision of the AI in a pair relied not only on the dictator's choices in this pair, but also on the training data generated by other dictators in the past. The 30 decisions reflecting the preferences of the current dictator were pooled with the 30 choices from each of 99 randomly selected previous participants in the same experiment.[7] Hence, the dictator was only responsible for 1% of the training data used by the random forest algorithm and was thus not (or hardly) pivotal for its resulting decision, as illustrated by the dashed arrows in Figure 2.

We also compared the dictators' behavior in the Full Pivotality treatment with their decisions in the *Shared Pivotality treatment*. This treatment adds to the baseline an interdependence between two pairs in terms of AI training and payoff allocation in period 31. In this variation, the algorithm was additionally trained with the data generated by a dictator from another pair in a past session of the experiment. We call this pair the "predecessor pair". In addition to and independent of the dictator's own payoff in the current session, the predecessor pair received an additional payoff amounting to half of the earnings generated by the option chosen by the AI. In short, participants in the Shared Pivotality treatment inherited training data from another pair's dictator and passed on their own training data to the members of this other pair. We pooled the data generated by the dictator with data from past sessions and not from the current session to avoid strategically different behavior. This was made common knowledge. In such settings, both dictators, past and present, equally shared the responsibility for the payoffs determined by the AI's prediction in the predecessor pair and in the current pair, such that pivotality was diffused. They both affected the decision of the algorithm in both groups. In Figure 2, this is represented by solid arrows pointing from each dictator to each of two robots. In contrast to the No Pivotality treatment variation, all dictators who contributed to the AI training were also monetarily affected by its prediction. Therefore, two opposing mechanisms could simultaneously influence the dictator's decisions. On the one hand, pivotality was reduced (which might increase selfishness) and, on the other hand, the dictator's choices now (partially) determined the degree of inequality in both one's own pair and the other pair.

Finally, in the *Full Pivotality-Others treatment*, the AI whose prediction determined the payoffs of a pair in period 31 was exclusively trained with the 30 decisions of a dictator from another pair in the session. Similarly, the 30 decisions of the dictator in the pair were used as training data for the AI that decided for another, third, pair in the same session. Hence,

[7]These participants were informed that their decisions would affect future participants' earnings through an AI training. The impact of this information on behavior is analyzed in our companion paper.

the dictator in the Full Pivotality-Others treatment was fully pivotal for the AI's decision in another pair, but had no influence on the AI prediction in his or her own pair. We depict this in Figure 2 by solid arrows that point from a dictator to the robot in another group.

2.3 Procedures

Due to the 2020 coronavirus pandemic and the enforced social distancing rules, we have not been able to invite students to the laboratory for in-person sessions. Therefore, we implemented an online experiment after recruiting students from the regular subject pool of the Frankfurt Laboratory for Experimental Economic Research (FLEX) through ORSEE (Greiner, 2015). All the subjects from this pool were informed about the new format and less than 1% decided to opt out from online experiments. We programmed the experiment in oTree (Chen et al., 2016) and hosted it on a Heroku server to allow participants to take part from outside the laboratory. The registered participants received the details for joining the virtual Zoom meeting via email and a unique ID to ensure anonymity in the experimental session. On the day of the experiment, they first came into a waiting room so that we could check that they registered for the respective session and rule out multiple participation of the same person. If there was an odd number of participants, one person was selected at random, who then received 4 Euros for connecting on time and could leave. All participants were muted and had to keep their video turned off. After a short introduction from the experimenters, the oTree link to the program was distributed. As soon as everyone was in the oTree room, the experiment started. Questions were asked and answered through the chat with the experimenters. To prevent the dictators from simply clicking through the 30 decisions, we forced them to wait for three seconds after their choice of option before they could validate their decision.

We determined a target of 30 dictators and 30 receivers in each treatment, based on a prior statistical power analysis to detect a medium-size effect with a significance level of 5% and a power of 80%. For the underlying distribution of decisions, we chose a binomial distribution where we varied the probability of selecting the selfish option X. A total of 308 participants (154 dictators or independent observations) were recruited for the experiment between July and September 2020. We ran two pilot sessions in June 2020 to check for technical functionality and calibrate the decision space of the dictator; the data from these pilot sessions were not used in the analysis. We preregistered the project, the sample size, and our hypotheses on AsPredicted (#44010) in July 2020. In the experiment, there were 34 dictators in the Full Pivotality treatment, 29 in the No Pivotality treatment, 34 in the Shared Pivotality treatment, and 30 in the Full Pivotality-Others treatment.[8] There were

[8]For the Shared Pivotality treatment, we pre-registered 60 dictators and eventually collected data from 61 dictators. This is because the AI was trained with the data generated both by a dictator in this treatment and by a dictator who participated in a past session of a treatment reported in our companion paper (Klockmann et al., 2021). Thus, for each pair in Shared Pivotality we needed one pair from a corresponding previous session that served as data source. For the analysis reported in this paper, we relied on the 34 participants

no drop-outs in any treatment and we did not exclude any observation. The average age of the participants was 24.7 years. About 60% were female. Their predominant field of study was economics, and they were on average in their 6^{th} to 7^{th} semester. Table C3 in Appendix C summarizes the main sociodemographic variables in each treatment. There is no significant differences across treatments at the threshold of 5%.

On average, participants earned 14.25 Euro (S.D. 3.60), including the bonus from the belief elicitation, and received their payoff by using PayPal. The conversion rate was 1 Euro per 100 points. Each session lasted approximately 45 minutes.

3 Behavioral Predictions

A model with standard preferences would not predict any differences in the dictators' decisions: dictators are expected to choose the selfish option in all games, regardless of the treatment. However, the presence of social preferences may lead to different predictions. We employ two measures to assess the dictators' distributional preferences. The first measure is the frequency of choosing the selfish option X in the whole sample. The second measure is given by the estimation of social preference parameters. Bruhin et al. (2019) built on Fehr and Schmidt (1999) and Charness and Rabin (2002) to set up a model of social preferences for two players that is fitted to the data, using maximum likelihood. In this model, the dictator's utility function is given by

$$u_D(\pi_D, \pi_R) = (1 - \alpha s - \beta r)\pi_D + (\alpha s + \beta r)\pi_R.$$

where π_D denotes the payoff of the dictator and π_R the payoff of the receiver. The indicator functions $s = \mathbb{1}\{\pi_R - \pi_D > 0\}$ and $r = \mathbb{1}\{\pi_D - \pi_R > 0\}$ equal one in case of disadvantageous and advantageous inequity for the dictator, respectively. As in Bruhin et al. (2019), the sign of the parameters α and β describe the preference type of the dictator. $\alpha < 0$ indicates that the dictator is envious of the receiver's payoff whenever receiving a lower amount, and it captures behindness aversion. $\beta > 0$ indicates that the dictator seeks to increase the other's payoff whenever receiving a larger amount, which reveals aheadness aversion. The absolute values of α and β measure how envious or empathetic the individual is. Moreover, $\alpha, \beta < 0$ indicates the presence of spiteful preferences, that is, the dictator dislikes the receiver's payoff, regardless of whether the receiver earns more or less than the dictator. $\alpha = \beta = 0$ reveals purely selfish preferences. Finally, $\alpha, \beta > 0$ indicates altruistic preferences, that is, the dictator derives utility from the receiver's payoff.

We now introduce our conjectures. When the implemented allocation resulting from the AI's prediction in period 31 only relies on the 30 choices of the dictator, decision makers

who took part in the Shared Pivotality treatment during the same time period as the participants in the other conditions.

have to cope with a trade-off between selfishness and social preferences. Being fully pivotal for the algorithmic prediction directly implies full pivotality for the resulting inequality. A dictator exhibiting social preferences can actively increase the fairness of payoffs by training the AI with egalitarian decisions. Choices are thus expected to be less selfish than under standard preferences. When the allocation implemented in period 31 relies not only on the dictator's 30 decisions, but also on training data generated by 99 other dictators, this results in reduced pivotality of the dictator for the AI's actual choice. In line with Bénabou et al. (2018), this reduced pivotality may induce individuals to feel less responsible for their selfish choices than in the baseline condition and to downplay negative externalities. We conjecture that adding 99 other dictators' decisions in the AI training data creates a moral wiggle room for making more selfish decisions. Because his or her behavior has a smaller impact on the fairness of the AI decision when the dictator only provides 1% of the AI training data, we expect him or her to be less reluctant in allocating ECU to himself or herself at the expense of the receiver.

Conjecture 1. *Compared to the Full Pivotality treatment, dictators in the No Pivotality treatment choose the selfish option more frequently in decisions in which the receiver would get a higher payoff with the alternative. The estimated social preference parameters, α and β, are lower in No Pivotality treatment than in the Full Pivotality treatment.*

The comparison between the Full Pivotality and Shared Pivotality treatments requires to account for two possible opposing mechanisms. On the one hand, similar to the No Pivotality treatment, being responsible for a smaller share of the AI training data in the Shared Pivotality treatment offers moral wiggle room for selfishness. On the other hand, as the dictator's decisions now also affect another group, she might be willing to reduce inequality for these individuals, by making less selfish decisions than in the Full Pivotality treatment. We conjecture the first mechanism, which is directly connected to the AI's allocation choice in the own group, to be dominant.

Conjecture 2. *Compared to the Full Pivotality treatment, dictators in the Shared Pivotality treatment choose the selfish option more frequently in decisions in which the receiver would get a higher payoff with the alternative. The estimated social preference parameters, α and β, are lower in the Shared Pivotality treatment than in the Full Pivotality treatment.*

In the Full Pivotality-Others treatment, dictators train the AI exclusively for another pair of players. In contrast with the Shared Pivotality treatment, their decisions affect only another pair's future payoffs in period 31. Since dictators are fully responsible for the outcome of another pair, the moral wiggle room for selfish actions provided by reduced pivotality disappears. And, if they have social preferences, dictators may be concerned with reducing inequality in the other pair.

Conjecture 3. *Compared to the Shared Pivotality treatment, dictators in the Full Pivotality-Others treatment choose the selfish option less frequently in decisions in which the receiver*

would get a higher payoff with the alternative. The estimated social preference parameters, α and β, are larger in the Full Pivotality-Others treatment than in the Shared Pivotality treatment.

4 Results

To identify how varying the degree of pivotality for training data generated for an artificially intelligent algorithm affects dictators' ethical behavior, our analysis focuses on two measures of moral behavior: the proportion of the selfish option X chosen by the dictator and the social preference parameters of a representative agent for each treatment, following Bruhin et al. (2019). Tables 2 and 3 report pairwise tests that compare the differences of these measures across treatments. Our analysis was conducted both on the full sample of observations and on a restricted sample. This restricted sample refers to the set of decisions characterized by conflicting interests, that is, those in which the dictator obtains a higher payoff with option X while the receiver is monetarily better off with option Y.[9]

Table 2: Overview of the Frequency of Choices of the Selfish Option X across Treatments

Treatments	Nb Obs.	Option X	*p–values*	Option X [Restricted Sample]	*p–values*
No Pivotality	29	79.20% (0.024)	0.025	80.27% (0.035)	0.019
Full Pivotality	34	70.29% (0.029)	0.045	66.01% (0.046)	0.010
Shared Pivotality	34	77.25% (0.018)	0.052	80.07% (0.027)	0.023
Full Pivotality-Others	30	70.89% (0.028)		68.52% (0.043)	

Notes: The table reports the relative frequency of the choice of the selfish option X in each treatment, with standard errors of means in parentheses. Each dictator in periods 1-30 represents one independent observation. Column "Option X [Restricted Sample]" includes only the decisions in games characterized by conflicting interests, that is, in which the dictator obtains a strictly higher payoff with option X and the receiver gets a strictly higher payoff with option Y. p-values refer to two-sided t-tests for differences in means.

Tables 2 and 3 reveal that almost any variation of pivotality for training data played a significant role in dictators' behavior. In the No Pivotality treatment, participants were significantly more likely to choose the selfish option X than in the Full Pivotality treatment, both in the full and the restricted samples ($p = 0.025$ and $p = 0.019$, respectively). The estimated social preferences parameters, α and β, were significantly lower in the No Pivotality than in the Full Pivotality treatment ($p = 0.040$ and $p = 0.050$, respectively). Furthermore,

[9] In addition, Figures D.1 and D.2 in Appendix D display the distribution of the shares of selfish option X chosen by the dictators, by treatment, in the full sample and the restricted sample, respectively.

Table 3: Estimated Parameters of Social Preferences across Treatments

Treatments	Nb Obs.	Dictators	α	*p–values*	β	*p–values*
No Pivotality	870	29	−0.050 (0.044)	*0.040*	0.253*** (0.051)	*0.050*
Full Pivotality	1020	34	0.082* (0.048)	*0.056*	0.394*** (0.051)	*0.020*
Shared Pivotality	1020	34	−0.047 (0.048)	*0.135*	0.246*** (0.037)	*0.022*
Full Pivotality-Others	900	30	0.054 (0.048)		0.382*** (0.046)	

Notes: The table reports the estimates of the α and β parameters of advantageous and disadvantageous inequality aversion, respectively, for a representative agent in the treatments, with robust standard errors clustered at the individual level in parentheses. One observation corresponds to one dictator in one period. The number of observations shows how many data were used to estimate inequity aversion in each treatment. p-values refer to z-tests for differences in estimates. * $p < 0.10$, *** $p < 0.01$.

efficiency concerns were significantly weakened when pivotality was decreased. Table C5 in Appendix C shows that dictators chose significantly more frequently the selfish option X even when choosing the alternative option Y would have maximized the sum of payoffs ($p = 0.027$).

This analysis supports Conjecture 1 and leads to Result 1.

Result 1. *Reducing pivotality for the AI's training induced dictators to behave more selfishly. The percentage of choices of the selfish option X was significantly higher in the No Pivotality than in the Full Pivotality treatment, regardless of whether option Y was efficient or not. The estimated social preference parameters were significantly lower in the No Pivotality treatment.*

A similar picture emerged when comparing the Full Pivotality with the Shared Pivotality treatments. The percentage of the selfish option X increased significantly in the latter treatment ($p = 0.045$ in the full sample and $p = 0.010$ in the restricted sample). The estimated social preference parameters, α and β, were significantly lower in the Shared Pivotality treatment ($p = 0.056$ and $p = 0.020$, respectively). Furthermore, Table C5 in Appendix C shows that dictators cared less about efficiency in this treatment. Indeed, they opted for the selfish alternative significantly more frequently even when choosing option Y would have increased welfare ($p = 0.028$). To sum up, participants in the Shared Pivotality treatment behaved very similarly to their counterparts in No Pivotality treatment. Diffusing pivotality by letting dictators from two groups independently influence the AI's training that determine future payoffs in both groups resulted in less egalitarian payoff allocations and in the estimation of more selfish preferences.

This analysis supports Conjecture 2. It is summarized in Result 2.

Result 2. *Individuals made more selfish decisions when pivotality was diffused by letting two dictators train the AI, compared with a setting in which they were fully responsible for*

the algorithmic outcome. The percentage of the selfish option X was significantly higher in the Shared Pivotality treatment than in the Full Pivotality treatment. The estimated social preference parameters were significantly lower in the Shared Pivotality treatment.

When comparing the Shared Pivotality treatment with the Full Pivotality-Others treatment, in which dictators trained the AI exclusively for another group, we found that dictators made significantly less selfish choices in the latter treatment, particularly when decisions presented a conflict of interest ($p = 0.052$ in the full sample and $p = 0.023$ in the restricted sample). The estimated disadvantageous inequality aversion parameter α did not differ significantly ($p = 0.135$) between these two treatments, but the estimated advantageous inequality aversion β parameter did, with a higher value of β in Full Pivotality-Others ($p = 0.022$). Along the same lines, dictators selected the selfish option less frequently in the Full Pivotality-Others treatment than in the Shared Pivotality treatment when they were in an advantageous position relative to the receiver regardless of their choice ($p = 0.046$, Table C4 in Appendix C), and when option Y reduced inequality ($p = 0.059$, Table C6 also in C). This suggests that when training an AI for a third party, individuals tended to counterbalance the power position that was exogenously given to them when assigned the role of a dictator, because they knew they were fully responsible for determining others' payoffs. Furthermore, dictators were not affected monetarily in period 31 by their own choices in periods 1 to 30, which may have led them to set aside their self-interest in favor of a fairness norm in this treatment.

This analysis supports Conjecture 3 and it is summarized in Result 3.

Result 3. *Individuals who taught an AI algorithm that exclusively decided for others behaved less selfishly than those who generated training data that also partially influenced their own outcome. The percentage of the selfish option X was significantly lower and the advantageous inequality aversion parameter was significantly higher in the Full Pivotality-Others treatment than in the Shared Pivotality treatment. This difference was most pronounced when the dictator was in an advantageous position regardless of the chosen option, and when option Y reduced inequality.*

Finally, we report an exploratory analysis of the relationships between beliefs and behavior. In the No Pivotality, Shared Pivotality, and Full Pivotality-Others treatments, the prediction of the AI for one's own group depended on the behavior of other dictators. Before informing participants about their payoff from the AI's prediction, we elicited the dictators' beliefs about the training data. Dictators had to guess which option would be selected by the AI in four different pairs of alternatives, as described in section 2.2. Out of the four questions, participants made three correct guesses on average, without any significant differences across treatments ($p > 0.39$ in all pairwise t-tests). Moreover, as reported in Table C7 in Appendix C, the average belief about the number of times option X was selected in these scenarios did not significantly differ across treatments ($p > 0.10$ in all pairwise t-tests). In

addition, Table 4 reports the results of an Ordinary Least Square regression analysis that tested for the relationship between the dictators' beliefs about the AI's training and their own behavior.[10] The dependent variable is the relative frequency of the dictator's choices of the selfish option X in the 30 games and the independent variable is the dictator's belief about the number of selfish choices by the AI in the three scenarios.

Table 4: Relationship between the Dictators' Beliefs and their Behavior, by Treatment

	Aggregate	No Pivotality	Shared Pivotality	Full Pivotality-Others
Belief Option X	0.088***	0.053	0.076***	0.135***
	(0.019)	(0.031)	(0.027)	(0.038)
Constant	0.604***	0.697***	0.632***	0.494***
	(0.035)	(0.060)	(0.053)	(0.066)
Number of observations	93	29	34	30
R^2	0.198	0.097	0.196	0.305

Notes: The table reports OLS estimates of the percentage of choices of the selfish option X in the 30 games on the belief on how frequently option X was selected by the AI in the three scenarios (variable "Belief Option X"). We excluded the fourth pair of alternatives corresponding to the actual menu of options the AI faced in period 31 because this varied across groups. Thus, the belief variable takes value 0, 1, 2 or 3. Dictators were asked to guess the decision of the AI that was trained only with the choices of those dictators whose generated data affected the algorithm's decision in their own group. The Aggregate column pools the data from all treatments except Full Pivotality. Standard errors are in parentheses. *** $p < 0.01$.

When pooling all three treatments, increasing a dictator's belief by one came along with a significant increase in the percentage of option X being chosen by this dictator by about 9 percentage points (*pp*). On average, these estimates showed that, compared to a dictator who believed that AI would always predict the altruistic option Y, a dictator who believed that the AI would always predict the selfish option X picked option X herself about 50% more frequently. We also estimated this effect for each treatment separately. While there was no statistically significant relationship between beliefs and behavior in the No Pivotality treatment, dictators in the Shared Pivotality and the Full Pivotality-Others treatments chose, respectively, 7.6*pp* and 13.5*pp* more frequently the selfish option X when expecting self-serving behavior by the other dictators ($p < 0.01$ in both cases).

We interpret this positive, significant correlation between dictators' behavior and beliefs as follows. When their outcome was not solely determined by their own behavior, dictators seemed to take advantage of the moral wiggle room offered by others' training of the AI and by not being pivotal for the algorithm' decision. Of course, this positive relationship could also be driven by a false consensus effect (although we paid participants for accurate beliefs). However, while behavior varied across treatments, beliefs did not. This suggests that rather than biasing their beliefs *ex post* according to their actual behavior because of

[10] A Tobit model was not necessary since less than 10% of the observations were either left or right censored. On aggregate, only 4 out of 93 dictators in these three treatments always chose option X (4.3%), and nobody always chose option Y.

a false consensus effect, dictators actively adjusted their beliefs to justify their self-serving training of the algorithm itself.

We conclude this exploratory analysis with the following, last, result.

Result 4. *Individuals who believed that the AI was trained by other selfish dictators behaved more selfishly themselves when generating training data.*

5 Discussion and Conclusion

Before making decisions, artificial intelligence has to be trained based on data. In various scenarios such as when judges employ decision support systems that predict recidivism or are used for bail decisions, this data stems from humans and their behavior. In the ethical domain, this means that individuals are responsible for the data they generate for the training of AI, and that the more they behave, for example, in a polarized manner themselves, the more likely the future decisions made by algorithms will also exhibit this polarization. However, the feeling of individual responsibility in this training may be more or less diffuse, depending notably on the pivotality of an individual's decisions in generating training data for the AI. In our laboratory experiment, using mini-dictator games, we studied to what extent varying the common knowledge pivotality of individuals in the training of an artificially intelligent algorithm affected the prosociality in human behavior. In some treatments, the individual's decisions were the exclusive source of training of the AI, while in others these decisions represented only half, 1% or none of the AI training data. Our results demonstrated that reduced pivotality increased selfishness in how humans trained the AI.

When the dictators' decisions were pooled with those of many others (in the No Pivotality treatment), individuals took advantage of the moral wiggle room offered by Big Data. They opted significantly more frequently for selfish choices and the estimation of social preference parameters showed that less weight was put by individuals on others' payoff. We observed similar behavior when dictators from two groups independently trained the algorithm that decided upon payoff allocations in both groups. Dictators behaved more selfishly in the Shared Pivotality treatment compared to the setting in which they were fully accountable for the training of the algorithm (the Full Pivotality treatment). Here again, we interpret this behavior as the expression of a moral wiggle room for the individuals when their behavior was merged with that of one or 99 other individuals to train the AI. This interpretation is backed by the positive correlation we identified between the participants' beliefs about others' selfishness reflected in the training data and the selfishness in their own choices. In other words, anticipating that others might make selfish choices gave an excuse to people for behaving similarly, as if they were complying with an empirical social norm in a society. In contrast, we observed more egalitarian allocation choices when individuals taught an AI deciding exclusively for a third party (in the Full Pivotality-Others

treatment), compared to the case in which they trained an algorithm also taking decisions for themselves (in the Shared Pivotality treatment). In that setting, individuals were again fully accountable for what was going to happen to another pair of participants, and this contributed to reduce selfish inclinations.

Overall, our results suggest that distortions in human behavior permitted by the exploitation of moral wiggle room are also reflected in AI training when pivotality, and related responsibility, diffuses due to Big Data. Training an intelligent system in groups or with many observations probably provides an excuse for acting less prosocially, simply because the individual may feel only partly responsible for the algorithmic prediction or outcome. If this behavioral pattern replicates for many people, this could result in much more selfish algorithmic recommendations or decisions in a society. On the one hand, the AI is trained in a more "democratic" manner, but, on the other hand, this does not necessarily translate into more prosocial or more moral behavior. This is a challenging dilemma. In our companion paper (Klockmann et al., 2021) we considered another aspect of AI training and we concluded that when individuals' decisions that train an AI had an impact on future generations, decision makers behaved less selfishly only when they could be harmed themselves by the AI's future decisions. Taken together, these results highlight the importance of making the externalities of their own actions more salient to individuals when they interact with intelligent machines.

Of course, our study has limitations. In our design, the AI training data come from decisions that revealed individuals' selfishness and prosociality with relatively limited monetary consequences. It would be interesting to explore with a similar manipulation of pivotality other decisions that involve moral and ethical dilemmas with larger, or also non-monetary consequences, or that affect many more individuals. It might be also interesting to explore more complex settings, using strategic games such as trust or bargaining games, to analyze how individuals would take into account how varying pivotality may influence their decisions. In terms of policy implications, our findings suggest a need for attributing explicit and salient individual responsibility to those affecting algorithmic predictions. Future research on human-machine interaction could tackle the question of the best way to increase the awareness of individuals about the externalities of their individual behavior through Big Data. Another policy recommendation from our results is to extend machine learning algorithms with classical programming that explicitly sets guidelines regarding morality or fairness. This would require a consensus on developing artificial intelligence that implements these principles.

References

Anderson, M. and S. L. Anderson (2007): "Machine ethics: Creating an ethical intelligent agent," *AI Magazine*, 28, 15.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan (2018): "The moral machine experiment," *Nature*, 563, 59–64.

Awad, E., S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon (2020): "Universals and variations in moral decisions made in 42 countries by 70,000 participants," *Proceedings of the National Academy of Sciences of the United States of America*, 117, 2332–2337.

Bartling, B., U. Fischbacher, and S. Schudy (2015): "Pivotality and responsibility attribution in sequential voting," *Journal of Public Economics*, 128, 133–139.

Bénabou, R., A. Falk, and J. Tirole (2018): "Narratives, imperatives, and moral reasoning," Working Paper 24798, National Bureau of Economic Research.

Benndorf, V., T. Grosse Brinkhaus, and F. von Siemens (2020): "Ultimatum Game Behavior in a Social-Preferences Vacuum Chamber," Mimeo, Goethe University Frankfurt.

Bonnefon, J.-F., A. Shariff, and I. Rahwan (2016): "The social dilemma of autonomous vehicles," *Science*, 352, 1573–1576.

Bostrom, N. and E. Yudkowsky (2014): "The ethics of artificial intelligence," in *The Cambridge handbook of artificial intelligence*, ed. by K. Frankish and W. M. Ramsey, Cambridge: Cambridge University Press, 316–334.

Breiman, L. (2001): "Random forests," *Machine Learning*, 45, 5–32.

Bruhin, A., E. Fehr, and D. Schunk (2019): "The many faces of human sociality: Uncovering the distribution and stability of social preferences," *Journal of the European Economic Association*, 17, 1025–1069.

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

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 (2020): "We and It: An Interdisciplinary Review of the Experimental Evidence on Human-Machine Interaction," Research Paper 20-15, Max Planck Institute for Innovation & Competition.

Coeckelbergh, M. (2020): *AI Ethics*, MIT Press.

Cohn, A., T. Gesche, and M. Maréchal (2018): "Honesty in the digital age," CESIfo Working Paper 6996.

Corgnet, B., R. Hernán-Gonzalez, and R. Mateo (2019): "Rac(g)e Against the Machine? Social Incentives When Humans Meet Robots," Working paper, University of Lyon.

Darley, J. M. and B. Latane (1968): "Bystander intervention in emergencies: Diffusion of responsibility." *Journal of Personality and Social Psychology*, 8, 377.

Falk, A., T. Neuber, and N. Szech (2020): "Diffusion of being pivotal and immoral outcomes," *Review of Economic Studies*, 87, 2205–2229.

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Ferraro, P. J., D. Rondeau, and G. L. Poe (2003): "Detecting other-regarding behavior with virtual players," *Journal of Economic Behavior & Organization*, 51, 99–109.

Frey, B. S. and F. Oberholzer-Gee (1997): "The cost of price incentives: An empirical analysis of motivation crowding-out," *American Economic Review*, 87, 746–755.

Frey, B. S., F. Oberholzer-Gee, and R. Eichenberger (1996): "The old lady visits your backyard: A tale of morals and markets," *Journal of Political Economy*, 104, 1297–1313.

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

Houser, D. and R. Kurzban (2002): "Revisiting kindness and confusion in public goods experiments," *American Economic Review*, 92, 1062–1069.

Houy, N., J.-P. Nicolai, and M. C. Villeval (2020): "Always doing your best? Effort and performance in dynamic settings," *Theory and Decision*, 89, 249–286.

Ivanov, A., D. Levin, and M. Niederle (2010): "Can relaxation of beliefs rationalize the winner's curse? An experimental study," *Econometrica*, 78, 1435–1452.

Klockmann, V., A. von Schenk, and M. C. Villeval (2021): "Artificial Intelligence, Ethics, and Intergenerational Responsibility," Mimeo, university of frankfurt and gate.

Lambrecht, A. and C. Tucker (2019): "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Management Science*, 65, 2966–2981.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011): "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Rahwan, I., M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, and M. Wellman (2019): "Machine behaviour," *Nature*, 568, 477–486.

Teubner, T., M. Adam, and R. Riordan (2015): "The impact of computerized agents on immediate emotions, overall arousal and bidding behavior in electronic auctions," *Journal of the Association for Information Systems*, 16, 838–879.

Van den Bos, W., J. Li, T. Lau, E. Maskin, J. D. Cohen, P. R. Montague, and S. M. McClure (2008): "The value of victory: Social origins of the winner's curse in common value auctions," *Judgment and Decision Making*, 3, 483–492.

Yamakawa, T., Y. Okano, and T. Saijo (2016): "Detecting motives for cooperation in public goods experiments," *Experimental Economics*, 19, 500–512.

A Appendix Random Forest Algorithm

For predicting the out-of-sample choice of the dictator in the 31st period, and as explained in Klockmann et al. (2021), we relied on a random forest algorithm as a standard classification method (Breiman, 2001). A Random Forest consists of several uncorrelated Decision Trees as building blocks. The goal of using a Decision Tree is to create a training model that can be used to predict the class of a target variable. It learns simple decision rules inferred from prior data (training data). In our experiment, the target variable was the decision of the dictator in period 31 and the training data corresponded to the dictator's previous decisions in periods 1 to 30. The algorithm took eight features as input variables to predict the binary outcome option X or option Y in period 31. Apart from the four payoffs for both players from both options, we further added the sum and difference between payoffs for each option as features.

All decision trees have grown under two types of randomization during the learning process. First, at each node, a random subset of features was selected to be considered when looking for the best split of observations. Hereby, we relied on the usual heuristics and allowed up to $\sqrt{8} \approx 3$ features. Second, only a random subset of observations was used to build each tree (bootstrapping). For each dictator, a forest consisted of ten different classification trees. To make the final decision on whether option X or option Y was the dictator's hypothetical 31st choice, each tree in the forest made a decision and the option with the most votes determined the final classification.

Due to the Python foundation of oTree, we made use of the random forest implementation of the scikit-learn package (Pedregosa et al., 2011). We further set a fixed random state or seed to ensure reproducibility of results. To assess the accuracy of the algorithm ex post, we split the decision data of each dictator in a training and test data set with 24 (80%) and 6 (20%) observations, respectively. For each individual, we thus trained a random forest with 24 randomly selected allocation choices and let it predict the six remaining ones. For all 127 dictators, this yielded an average predictive accuracy of about 84%. Note that this number should be rather taken as a lower bound on the actual accuracy of the algorithm in the experiment that actually used all 30, 60, or 3000 decisions of the dictator(s) for training to make the out-of-sample prediction.

The questionnaire in the final stage included several questions about the participants' attitudes toward AI in general and toward the machine learning algorithm in our experiment in particular. On a scale from 1 to 5, we asked dictators to rate their familiarity with and confidence in this technology (averages of 2.6 and 3.8, respectively), their satisfaction with the prediction in period 31 (average of 4.0), and their assessment of how accurately the AI's decision matched their true preferences (average of 4.2). There were no significant differences across treatments in any of these variables. Satisfaction with and assessed accuracy of the algorithm were not only very high, but also strongly correlated (Spearman rank correlation: 0.483, $p < 0.001$).

B Instructions

The experiment was conducted online with student subjects from Goethe University Frankfurt in German language. This section provides the instructions translated into English and the screenshots.

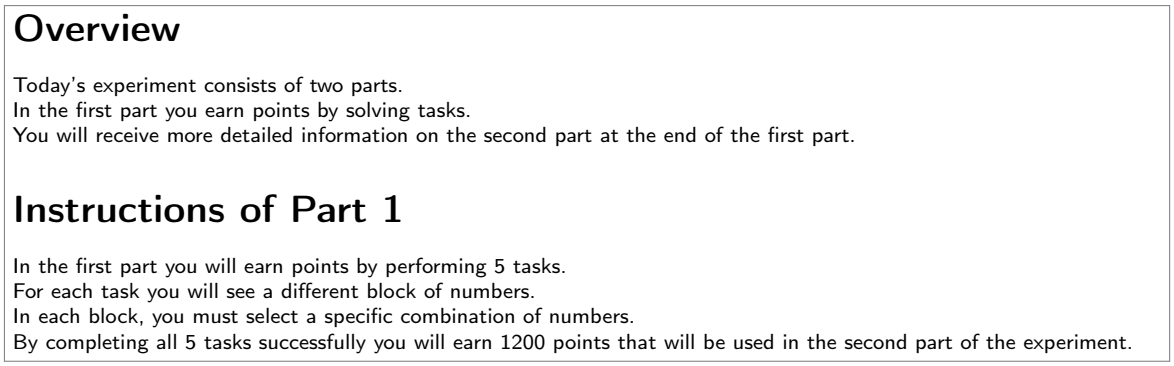

Figure B.1: Instructions – Real Effort Tasks

Note: This screen was displayed in all treatment pairs before participants performed the real effort tasks.

Figure B.2: Instructions – End of Real Effort Tasks

Note: This screen was displayed in all treatment pairs after participants completed the real effort tasks.

Instructions of Part 2

Instructions

The following instructions are shown to all participants. Please read carefully.
Afterwards, you need to answer a set of control questions to ensure your understanding before you can continue.

Overview

This part consists of 30 independent periods and a period 31 which differs from the previous 30 periods, as explained below.
At the beginning of the part, you will be randomly assigned a role, either participant A or participant B. You will keep this role throughout this part.
At the beginning of the part, you are going to be randomly matched with another participant to form a pair.
The pair of participant A and participant B will remain the same throughout the rest of the experiment.

Decisions in Periods 1 to 30

In each of these 30 periods, participant A has to choose between two options: option X and option Y.
Each option represents the share of a number of points between participant A and participant B.
The points that are distributed correspond to your earnings and the earnings of the other participant in your pair in the first part of the experiment.
In each option, the first number corresponds to the payoff of participant A, the second amount corresponds to the payoff of participant B.
In the entire experiment, 100 points correspond to one euro.
To validate his or her choice, participant A has to click on the option he or she prefers and then, validate by pressing the OK button.
It is very important to look carefully at the two amounts of each option before choosing the preferred option.
Note that participant B has no decision to make in this part.

Figure B.3: Instructions – Main Part and Decisions

Note: This screen was displayed in all treatment pairs.

Instructions of Part 2

Period 31

You will receive also a payoff for period 31 that will be added to your payoff in one of the previous periods.
Thus, your total payoff is determined by one of the 30 decisions made in periods 1 to 30, and by the unique decision made in period 31.

Your payoff in period 31 is determined as follows.
The previous 30 decisions of participant A are used to train an artificially intelligent Random Forest algorithm (see Info Box).
It is a machine learning algorithm that observes and learns from participant A's behavior.

[Full Pivotality]
Based on the 30 decisions of the participant A in your pair today, the algorithm makes a prediction.
The algorithm gives the full weight of 100% to participant A in your pair in forming its prediction.

Building on this source of training data, the algorithm chooses between option X and option Y in period 31.
Note that the two options X and Y between which the algorithm decides are randomly chosen.
They are of the same type as in the 30 previous decisions made by participant A.

In fact, participant A in your pair does not make a decision in period 31: it is the algorithm that makes the decision based on its prediction what participant A would prefer, given this participant A's choices of options in the first 30 periods.

The option chosen by the algorithm in this prediction determines your payoff in period 31 and the payoff of the other participant in your pair.
Example: If the algorithm predicts that participant A would prefer option X with payoffs (K, V): A in the pair receives K points and B receives V points.

Figure B.4: Instructions – Prediction of the Algorithm in Period 31

[No Pivotality]
The algorithm is additionally trained with data generated by 99 randomly selected participants A from past sessions of the experiment.
These 99 other participants participated in the same experiment as you in the exactly same conditions as you in periods 1 to 30.

Based on the 30 decisions of participant A in your pair today and the respective 30 decisions by 99 other participants A from your predecessors, the algorithm makes a prediction.
The algorithm gives the same weight of 1% to each participant A in forming its prediction.

Building on the 100 sources of training data, the algorithm chooses between option X and option Y in period 31.
Note that the two options X and Y between which the algorithm decides are randomly chosen.
They are of the same type as in the 30 previous decisions made by participant A.

In fact, participant A in your pair does not make a decision in period 31: it is the algorithm that makes the decision based on its prediction on what the 100 participants A, including the participant A in your pair, would prefer given the choice of options.

The option chosen in this prediction determines your payoff and the payoff of the other participant in your pair in period 31.
Example: If the algorithm predicts that the 100 participants A would prefer option X with payoffs (K, V): A in the pair receives K points and B receives V points.

[Shared Pivotality]
The algorithm is additionally trained with data generated by a participant A from another pair in a past session of the experiment.
This other participant participated in the same experiment as you in the exactly same conditions as you in periods 1 to 30.
We call this pair your predecessor pair.

Based on the 30 decisions of participant A in your pair today and the 30 decisions by the other participant A your predecessor pair, the algorithm makes a prediction.
The algorithm gives the same weight of 50% to each of the two participants A in forming its prediction.
This means that the 30 decisions of participant A from your pair represent one half of the training data for the algorithm.
The other half of the training data consists of the 30 decisions of participant A from your predecessor pair.

Figure B.4: Instructions – Prediction of the Algorithm in Period 31 (cont'd)

Building on these two sources of training data, the algorithm chooses between option X and option Y in period 31.
Note that the two options X and Y between which the algorithm decides are randomly chosen.
They are of the same type as in the 30 previous decisions made by participant A.

In fact, participant A in your pair does not make a decision in period 31: it is the algorithm that makes the decision based on its prediction on what the two participants A would prefer, given these two participants A' choices of options in the first 30 periods.

The option chosen by the algorithm in this prediction determines your payoff and the payoff of the other participant in your pair in period 31.
Additional to and fully independent of your own payoff, your predecessor pair receives a payoff amounting to 50% of the option chosen by the algorithm.
Example: If the algorithm predicts that the two participants A would prefer option X with payoffs (K, V): A in the pair receives K points and B receives V points.

[Full Pivotality-Others]
Based on the 30 decisions of the participant A the algorithm makes a prediction.
The algorithm that makes a prediction for your pair is trained with data from a participant A's 30 decisions from another pair in your session.
Similarly, the 30 decisions of the participant A in your pair are used as training data for the algorithm that decides for another pair in your session in period 31.
Thus, the 30 decisions of participant A in your pair have monetary consequences in period 31, not for your pair but for both participants (A and B) of another pair of participants in your session.
The algorithm gives the full weight of 100% to the participant A in your pair in forming its prediction for the other pair.

Building on this source of training data, the algorithm chooses between option X and option Y in period 31.
Note that the two options X and Y between which the algorithm decides in period 31 are randomly chosen.
They are of the same type as in the 30 previous decisions made by participant A.

In fact, participant A in your pair does not make a decision in period 31: it is the algorithm that makes the decision based on its prediction what participant A from the other pair would prefer given the choice of options.

The option chosen by the algorithm in this prediction determines your payoff and the payoff of the other participant in your pair in period 31.
Example: If the algorithm predicts that participant A from the other pair would prefer option X with payoffs (K, V): A receives K points and B receives V points.

Figure B.4: Instructions – Prediction of the Algorithm in Period 31 (cont'd)

Info box: Random Forest Algorithm

A Random Forest is a classification method.
Classification is a two-step process in machine learning: there is a learning step and a prediction step.
In the learning step, the model is developed based on given training data.
In the prediction step, the model is used to predict the response for given data.
A Random Forest consists of several uncorrelated Decision Trees as building blocks.
The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of a target variable.
It learns simple decision rules inferred from prior data (training data).

In this experiment, the target variable is the decision of participant A in period 31.
The training data correspond to previous decisions in periods 1 to 30.

All decision trees have grown under a certain type of randomization during the learning process.
For a classification, each tree in that forest makes a decision and the class with the most votes decides the final classification.

Figure B.4: Instructions – Prediction of the Algorithm in Period 31 (cont'd)

Notes: This screen was displayed in all treatment pairs. The content of the paragraphs which dynamically varied across treatments is marked accordingly.

Control Questions

Please answer the following control questions.
You must answer all questions correctly before you can continue with the experiment.

Question 1

Will your pair of participant A and participant B remain the same throughout the whole experiment?

- ◉ Yes
- ○ No

Question 2

Which kind of decisions will be made and who will make them?

- ◉ Participant A decides upon the distribution of the earnings of both participants (A and B) from solving tasks in the first part of the experiment.
- ○ Participant B decides upon the distribution of the earnings of both participants (A and B) from solving tasks in the first part of the experiment.
- ○ Both participants jointly decide upon the distribution of the earnings of both participants (A and B) from solving tasks in the first part of the experiment.
- ○ Participant A proposes a distribution of the earnings of both participants (A and B) from solving tasks in the first part of the experiment. Participant B can accept or reject this proposal.

Question 3

Does participant B make any decision with regard to the distribution of the endowment earned by both participants within your pair?

- ○ Yes
- ◉ No, participant A decides upon the allocation of the endowment of both participants.

Question 4

How will your payoff from the first 30 periods be determined?

- ○ There is no payoff from the first 30 periods.
- ○ All decisions of participant A in all periods will be paid out.
- ◉ One decision of participant A in one randomly selected period will be paid out.

Figure B.5: Control Questions

Question 5

How will your payoff from period 31 be determined?

- ○ There is no payoff from period 31.
- ○ Participant A makes another decision that is paid out.
- ◉ Participant A makes no decision, but there is an artificially intelligent algorithm that makes a prediction for period 31 based on learned behavior, which determines the payoffs in the pair.

Question 6

Where does the artificially intelligent algorithm in the experiment get its training data from?

- ○ Exclusively from the 30 decisions of participant A in your pair. [Full Pivotality]
- ○ Exclusively from the 30 decisions of participant A in another pair in your session. [Full Pivotality-Other]
- ○ From the 30 decisions of participant A in your pair and the 30 decisions of participant A in another pair. [Shared Pivotality]
- ○ From the 30 decisions of participant A in your pair and the 30 decisions of participant A in 99 other pairs. [No Pivotality]

Question 7

For the algorithm of which pair do the 30 decisions of participant A in your pair generate training data?

- ○ Exclusively for your pair. [Full Pivotality, No Pivotality, Shared Pivotality]
- ○ Exclusively for another pair in your session. [Full Pivotality-Other]

Question 8

What is the composition of your final payoff? (Multiple selections possible!)

- ◉ One decision of participant A in the first 30 periods is implemented for your pair.
- ◉ The decision of the artificially intelligent algorithm in period 31 is implemented for your pair.

Figure B.5: Control Questions (cont'd)

Question 9

Can roles within your pair be switched for payoff?

- ◉ No, I always keep my role.
- ○ Yes, it might be that with 50% probability I get the payoff of the other participant in my pair for the decision in the randomly selected period between 1 and 30.
- ○ Yes, it might be that with 50% probability I get the payoff of the other participant in my pair for the decision by the artificially intelligent algorithm in period 31.

Figure B.5: Control Questions (cont'd)

Notes: The selected answers are the correct ones for all treatment pairs. Some answers to the control questions vary across treatments and the correct ones are marked accordingly.

Results *[Example]*

Randomly selected round: Period 5
Options in this round: (640, 410) and (560, 790)
Decision in this round: **Option X**
Your payoff: **640 points**

Figure B.6: Results of Part 2 (Example)

Notes: This screen was shown to all subjects after the dictator has made the 30 decisions. The numbers and option choice are for illustrative purposes only.

Your Beliefs

Now, before the artificially intelligent algorithm makes a decision in period 31, we would like you to state your beliefs.
Reminder:
[No Pivotality]
The algorithm is also trained with decisions made by 99 randomly selected participants A from previous sessions of the experiment.
Now suppose for a moment that the algorithm had only this training data and your decisions were irrelevant.
How do you think the algorithm would decide based on only this training data given the following four choices?

[Shared Pivotality]
The algorithm is also trained with choices made by participant A from another pair.
Now suppose for a moment that the algorithm had only this training data and your choices were irrelevant.
How do you think the algorithm would decide based on only this training data given the following four choices?

[Full Pivotality-Other]
The algorithm is trained using only choices made by participant A from another pair in your session.
Based on this training data, how do you think the algorithm would decide given the following four choices?

You can earn 100 points for each correct answer.

Figure B.7: Belief Elicitation

Notes: This screen was shown to all the participants before learning the decision of the AI. The content of the paragraphs which dynamically varied across treatments is marked accordingly.

Period 31: Prediction *[Example]*

The artificially intelligent algorithm decided between (760, 490) and (440, 710) in this period 31.

[Full Pivotality]
Based on the previous decisions of participant A in your pair, the prediction and decision of the algorithm was **option X**.

[No Pivotality]
Based on the previous decisions of participant A in your pair and the decisions in previous pairs, the prediction and decision of the algorithm was **option X**.

[Shared Pivotality]
Based on the previous decisions of participant A in your pair and the decisions in a previous pair, the prediction and decision of the algorithm was **option X**.

[Full Pivotality-Other]
Based on the previous decisions in another pair, the prediction and decision of the algorithm was **option X**.

Your payout from period 31 is therefore 760 points.

[No Pivotality, Shared Pivotality, Full Pivotality-Other]
Overall, 2 of your beliefs were correct.
Therefore, you will receive a bonus of 200 points for the payoff of period 31.

Figure B.8: Results of Part 3 (Example)

Notes: This screen was shown to all the participants after the algorithm had made its prediction. The numbers and option choice are for illustrative purposes only. The content of the paragraphs which dynamically varied across treatments is marked accordingly.

Final Results *[Example]*

Your payoff from the randomly selected period 5 is 640 points.
Your payoff from period 31 is 760 points.
In total, you will thus receive a payoff of 1400 points.
This is equivalent to **14 euros**.

Figure B.9: Final Results (Example)

Notes: This screen was shown to all the participants at the end of the experiment before the final questionnaire. The numbers and option choice are for illustrative purposes only.

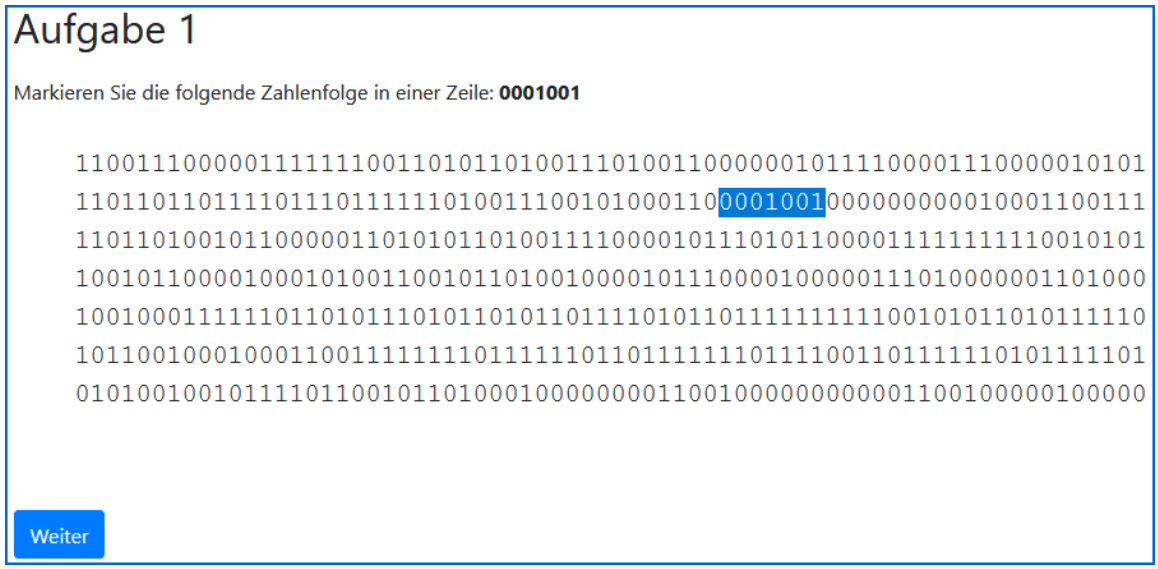

Figure B.10: Real Effort Task

Notes: Exemplary real effort task from the first part of the experiment. The correct solution needed to be marked as shown in the screenshot.

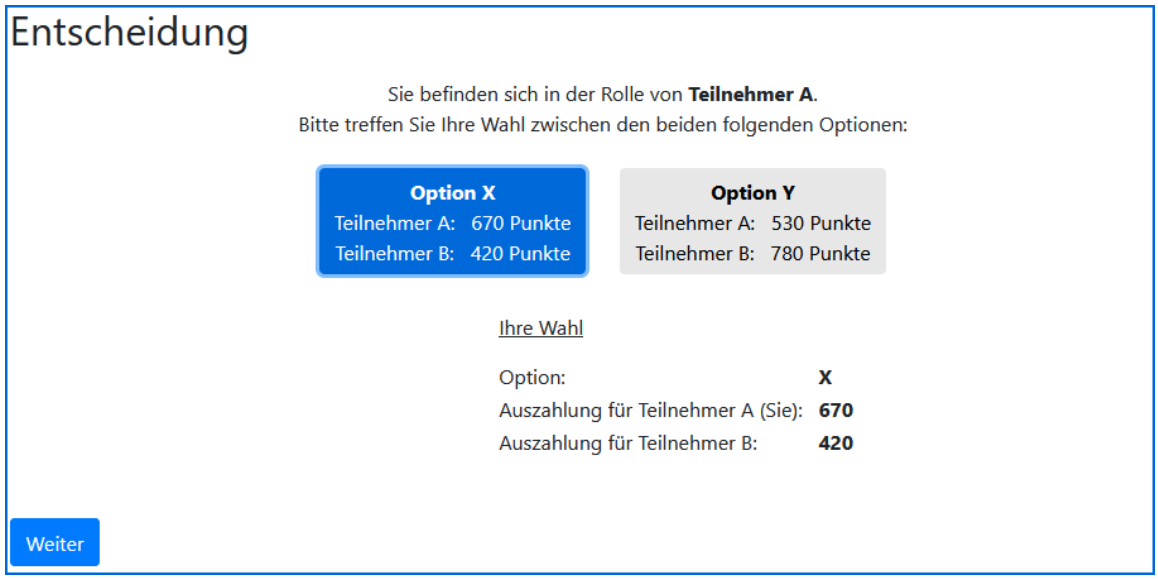

Figure B.11: Decision Screen of the Dictator

Notes: Exemplary decision screen of the dictator. The "Next" button appeared only 5 seconds after selecting an option to avoid rush decisions.

C Appendix Tables

Table C1: Decision Space of the Dictator Games

Game	Option X (Selfish)	Option Y (Altruistic)	Category 1 (Slope)	Category 2 (Dictator's Position)	Category 3 (Highest Efficiency)	Category 4 (Lowest Inequality)
1*	(890, 140)	(850, 520)	Selfish	Advantageous	Y	Y
2*	(910, 140)	(830, 520)	Selfish	Advantageous	Y	Y
3*	(940, 150)	(800, 510)	Selfish	Advantageous	Y	Y
4*	(980, 170)	(760, 490)	Selfish	Advantageous	Y	Y
5*	(1010, 190)	(730, 470)	Selfish	Advantageous	None	Y
6*	(1050, 270)	(690, 390)	Selfish	Advantageous	X	Y
7	(1060, 330)	(680, 330)	Receiver indiff.	Advantageous	X	Y
8	(990, 480)	(750, 180)	X Pareto	Advantageous	X	X
9	(930, 510)	(810, 150)	X Pareto	Advantageous	X	X
10	(870, 140)	(870, 520)	Dictator indiff.	Advantageous	Y	Y
11*	(620, 410)	(580, 790)	Selfish	Mixed	Y	None
12*	(640, 410)	(560, 790)	Selfish	Mixed	Y	None
13*	(670, 420)	(530, 780)	Selfish	Mixed	Y	None
14*	(710, 440)	(490, 760)	Selfish	Mixed	Y	None
15*	(740, 460)	(460, 740)	Selfish	Mixed	None	None
16*	(780, 540)	(420, 660)	Selfish	Mixed	X	None
17	(790, 600)	(410, 600)	Receiver indiff.	Mixed	X	None
18	(720, 750)	(480, 450)	X Pareto-dom.	Mixed	X	None
19	(660, 780)	(540, 420)	X Pareto-dom.	Mixed	X	None
20	(600, 410)	(600, 790)	Dictator indiff.	Mixed	Y	None
21*	(350, 680)	(310, 1060)	Selfish	Disadvantageous	Y	X
22*	(370, 680)	(290, 1060)	Selfish	Disadvantageous	Y	X
23*	(400, 690)	(260, 1050)	Selfish	Disadvantageous	Y	X
24*	(440, 710)	(220, 1030)	Selfish	Disadvantageous	Y	X
25*	(470, 730)	(190, 1010)	Selfish	Disadvantageous	None	X
26*	(510, 810)	(150, 930)	Selfish	Disadvantageous	X	X
27	(520, 870)	(140, 870)	Receiver indiff.	Disadvantageous	X	X
28	(450, 1020)	(210, 720)	X Pareto-dom.	Disadvantageous	X	Y
29	(390, 1050)	(270, 690)	X Pareto-dom.	Disadvantageous	X	Y
30	(330, 680)	(330, 1060)	Dictator indiff.	Disadvantageous	Y	X

Notes: The first entry of option X and option Y is the dictator's payoff, the second one is the receiver's payoff. In category 1, selfish decisions are characterized by conflicting interests, that is, the dictator strictly prefers option X and the receiver strictly prefers option Y. Category 2 describes the relative position of the dictator. Category 3 states which option maximizes the sum of payoffs. Category 4 states which option minimizes the absolute difference of payoffs. Stars in column 1 refer to the sub-set of games characterized by conflicting interests, that is, games in which the dictator strictly prefers option X while the receiver strictly prefers option Y; these games correspond to what is characterized in the analysis as the "restricted sample".

Table C2: Possible Out-of-Sample Decisions of the AI

Prediction	Option X (Selfish)	Option Y (Altruistic)	Category 1 (Slope)	Category 2 (Dictator's Position)	Category 3 (Highest Efficiency)	Category 4 (Lowest Inequality)
1	(1030, 220)	(710, 440)	Selfish	Advantageous	X	Y
2	(960, 500)	(780, 160)	X Pareto	Advantageous	X	X
3	(760, 490)	(440, 710)	Selfish	Mixed	X	None
4	(690, 770)	(510, 430)	X Pareto	Mixed	X	None
5	(490, 760)	(170, 980)	Selfish	Disadvantageous	X	X
6	(420, 1040)	(240, 700)	X Pareto	Disadvantageous	X	Y

Notes: One of the decision scenarios was randomly picked for the AI's prediction. The first entry of option X and option Y is the dictator's payoff, the second one is the receiver's payoff. In category 1, selfish decisions were characterized by conflicting interests, that is, the dictator strictly preferred option X and the receiver strictly preferred option Y. Category 2 describes the relative position of the dictator. Category 3 states which option maximized the sum of payoffs. Category 4 states which option minimized the absolute difference of payoffs.

Table C3: Summary Statistics, by Treatment

Treatments	Full Pivotality	No Pivotality	Shared Pivotality	Full Pivotality-Others
% Females	60.29	51.72	63.24	61.67
Mean age in years	24.28	25.33	24.03	25.20
	(0.60)	(0.53)	(0.40)	(0.49)
% Studies in Economics	33.82	48.28	50.00*	41.67
Mean nb Semesters	7.01	6.81	6.66	6.95
	(0.42)	(0.43)	(0.41)	(0.48)
Mean grade	1.96	2.05	1.83	1.91
	(0.08)	(0.08)	(0.06)	(0.07)
Mean expenses	1.41	1.69*	1.53	1.58
	(0.07)	(0.09)	(0.07)	(0.09)
N	68	58	68	60

Notes: The table displays summary statistics on the participants' sociodemographic characteristics, by treatment. Standard errors of means are in parentheses. Grade refers to the German Abitur grade and ranges from 1.0 (best) to 6.0 (worst). Expenses are on a weekly basis and coded by 1 (less than 100 Euros), 2 (between 101 and 200 Euros), and 3 (more than 200 Euros). The tests reported are based on comparisons with the Full Pivotality treatment. These tests are Fisher's exact tests, except for age, grade and semester, for which we used t-tests. * $p < 0.10$.

Table C4: Relative Frequency of Choices of the Selfish Option X, by Treatment and Relative Position of the Dictator

Treatments	Nb obs.	Option X [Advantageous]	*p–values*	Option X [Disadvantageous]	*p–values*
No Pivotality	29	68.62% (0.0426)		87.93% (0.0213)	
			0.080		*0.020*
Full Pivotality	34	57.94% (0.0417)		79.42% (0.0273)	
			0.065		*0.343*
Shared Pivotality	34	67.65% (0.0305)		82.65% (0.0199)	
			0.046		*0.233*
Full Pivotality-Others	30	57.67% (0.0392)		78.67% (0.0270)	

Treatments	Nb obs.	Option X [mixed]	*p–values*
No Pivotality	29	81.03% (0.0240)	
			0.096
Full Pivotality	34	73.53% (0.0355)	
			0.050
Shared Pivotality	34	81.47% (0.0180)	
			0.134
Full Pivotality-Others	30	76.33% (0.0297)	

Notes: This table reports the relative frequency of the choice of option X, by treatment and according to the relative position of the dictator in the game (advantageous, disadvantageous, or mixed), with standard errors of means in parentheses. One observation corresponds to one dictator. *p*-values refer to two-sided t-tests for differences in means.

Table C5: Relative Frequency of Choices of the Selfish Option X, by Treatment and Efficiency

Treatments	Nb. obs.	Option X [X efficient]	*p–values*	Option X [Y efficient]	*p–values*
No Pivotality	29	96.26% (0.0098)		63.91% (0.0436)	
			0.967		*0.027*
Full Pivotality	34	96.32% (0.0101)		48.63% (0.0502)	
			0.147		*0.028*
Shared Pivotality	34	93.63% (0.0154)		62.16% (0.0336)	
			0.365		*0.038*
Full Pivotality-Others	30	95.56% (0.0143)		50.00% (0.0476)	

Notes: This table reports the relative frequency of the choice of option X, by treatment and according to the efficiency of the option in the game, with standard errors of means in parentheses. Efficiency refers to the sum of payoffs. One observation corresponds to one dictator. *p*-values refer to two-sided t-tests for differences in means.

Table C6: Relative Frequency of Choices of the Selfish Option X, by Treatment and Relative Inequality

Treatments	Nb obs.	Option X [X fairer]	*p–values*	Option X [Y fairer]	*p–values*
No Pivotality	29	87.93% (0.0207)	*0.022*	68.62% (0.0440)	*0.077*
Full Pivotality	34	79.71% (0.0272)	*0.170*	57.65% (0.0420)	*0.123*
Shared Pivotality	34	84.41% (0.0203)	*0.233*	65.88% (0.0319)	*0.059*
Full Pivotality-Others	30	80.33% (0.0277)		56.00% (0.0411)	

Treatments	Nb obs.	Option X [equal]	*p–values*
No Pivotality	29	81.03% (0.0240)	*0.096*
Full Pivotality	34	73.53% (0.0355)	*0.050*
Shared Pivotality	34	81.47% (0.0180)	*0.134*
Full Pivotality-Others	30	76.33% (0.0297)	

Notes: This table reports the relative frequency of the choice of option X, by treatment and according to whether the option is fairer than the other option or not, with standard errors of means in parentheses. Fairness refers to the absolute difference in payoffs. One observation corresponds to one dictator. p-values refer to two-sided t-tests for differences in means.

Table C7: Distribution of Beliefs about Other's Behavior across Treatments

Treatments	Belief about the choice of option X			
	0	1	2	3
No Pivotality	6.90%	20.69%	58.62%	13.79%
Shared Pivotality	0.00%	26.47%	61.76%	11.76%
Full Pivotality-Others	6.67%	26.67%	66.67%	0.00%

Notes: The table reports participants' beliefs about how frequently option X was selected in the three scenarios presented to the dictators during the belief elicitation task. We excluded the fourth pair of alternatives corresponding to the actual menu of options the AI faced in the 31st period because it varied randomly across pairs. Participants were asked to guess the decision of the AI that was trained with the choices of the 99 other dictators in the No Pivotality treatment, with the choices of the other dictator in Shared Pivotality whose decisions we added to the own training data, and with the choices of the dictator who served as the only source of training data in Full Pivotality-Others.

D Appendix Figures

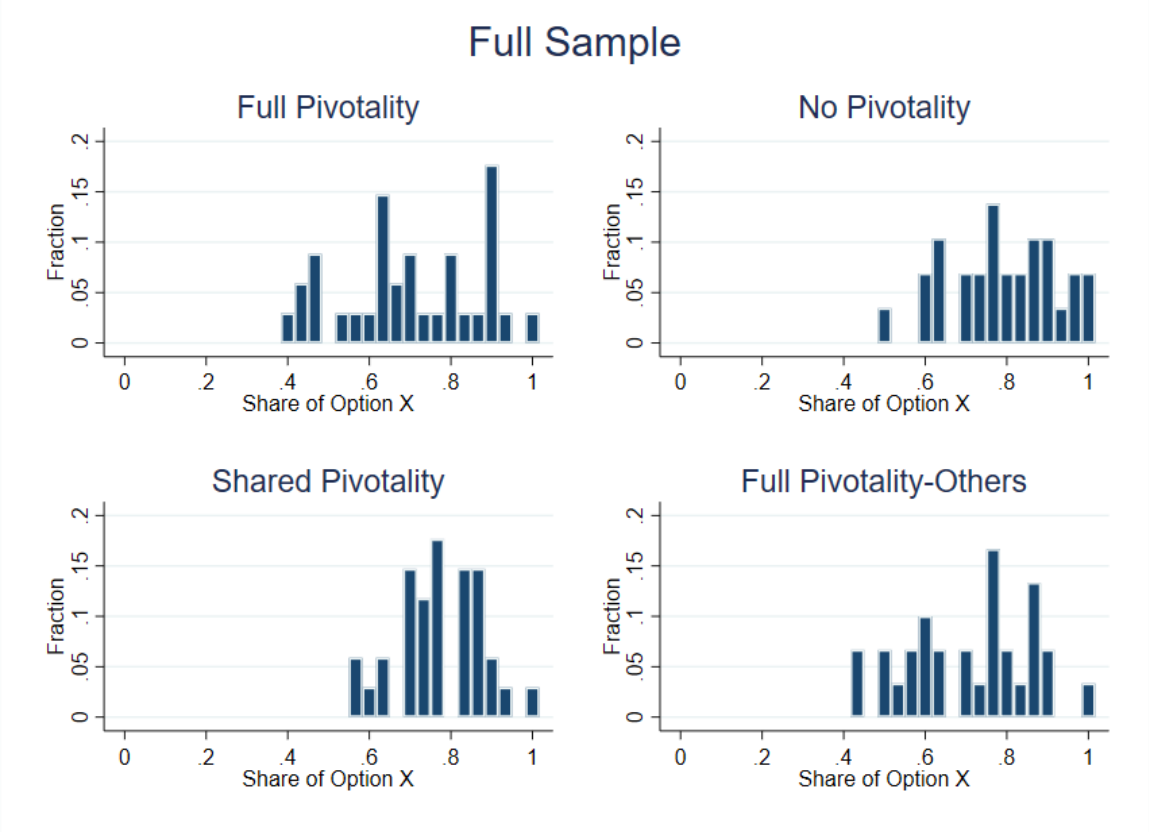

Figure D.1: Distribution of the Shares of Selfish Choices by the Dictators, by Treatment

Notes: The figure displays the distribution of the shares of choices of the selfish option X by the dictators in the 30 periods of the game, by treatment.

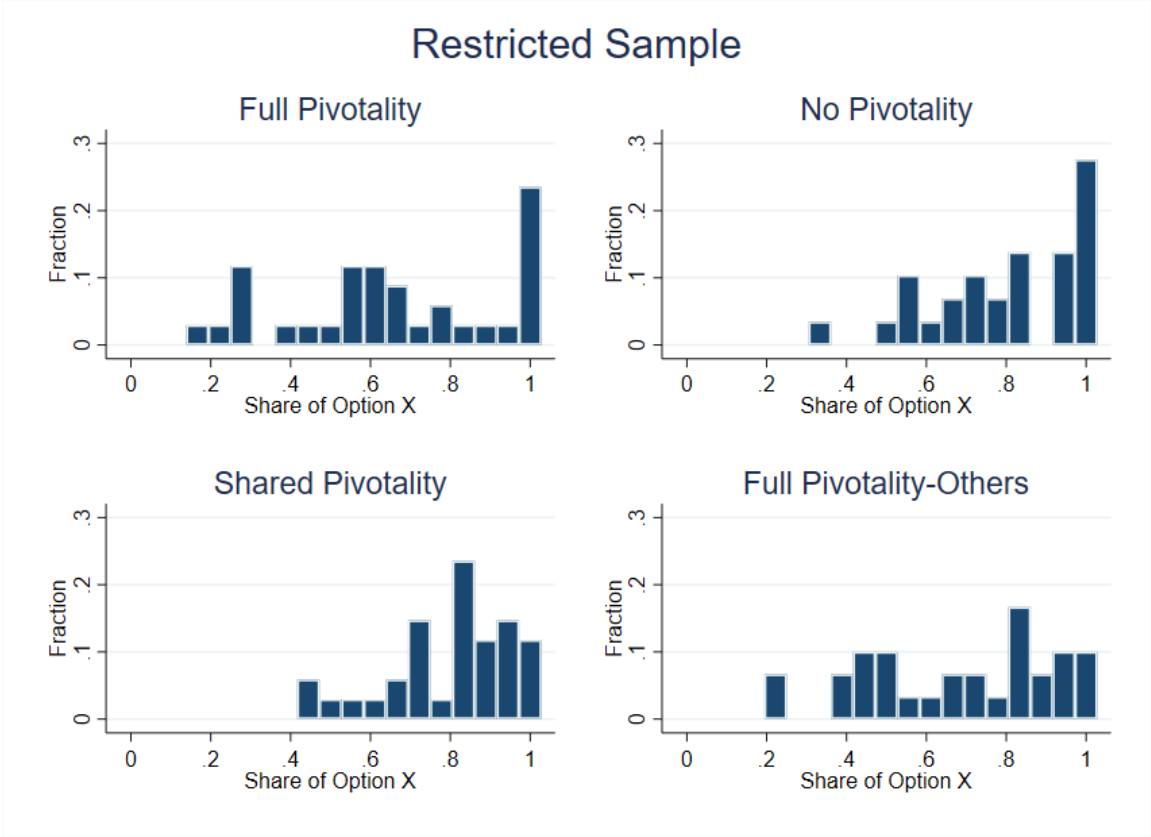

Figure D.2: Distribution of the Shares of Selfish Choices by the Dictators, by Treatment (Restricted Sample)

Notes: The figure displays the distribution of the shares of choices of the selfish option X by the dictators in the subset of games characterized by conflicting interests (that is, games in which the dictator gets strictly higher payoff with option X while the receiver gets strictly higher payoff with option Y), by treatment. These games correspond to what is characterized in the analysis as the "restricted sample".