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 – December 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 the 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. Diffusing such individual pivotality for algorithmic choices increased the share of selfish decisions and weakened revealed prosocial preferences. This does not result from a change in the structure of incentives. Rather, our results show that Big Data offers an excuse for selfish behavior through lower responsibility for one's and others' fate.

Keywords:
Artificial Intelligence, Big Data, Pivotality, Ethics, 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]

December 20, 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 the 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. Diffusing such individual pivotality for algorithmic choices increased the share of selfish decisions and weakened revealed prosocial preferences. This does not result from a change in the structure of incentives. Rather, our results show that Big Data offers an excuse for selfish behavior through lower responsibility for one's and others' fate.

Keywords: Artificial Intelligence, Big Data, Pivotality, Ethics, 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 research support from the Leibniz Institute for Financial Research SAFE, the Goethe University Frankfurt, 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 3, 60323 Frankfurt, Germany. Center for Humans & Machines, Max Planck Institute for Human Development, Berlin, Germany. klockmann@econ.uni-frankfurt.de.

[b]Goethe University Frankfurt, Theodor-W.-Adorno-Platz 3, 60323 Frankfurt, Germany. Center for Humans & Machines, Max Planck Institute for Human Development, 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] At the same time, IBM's 2021 Global Artificial Intelligence Adoption Index indicates that three-quarters of companies are now using Artificial Intelligence (AI) and this proportion increases even more rapidly since the pandemic. Despite undisputed benefits of the broad availability of massive data and AI, such as higher accuracy and prediction quality of 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. In line with the "bystander effect" (Darley and Latané, 1968), people might use their low pivotality, *i.e.*, lacking crucial importance, for certain outcomes to rationalize their selfish behavior that possibly hurts others *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 AI 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. This question is all the more important since in more and more situations individuals' decisions are used by companies to train AI with potential future consequences on one's and others' outcomes. Typical examples are provided by predictive analytics applied to customer data by firms to suggest products to customers, or wealth management tools used by financial companies to approve loans or recommend products.[3]

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

[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] For example, companies are using algorithms that track consumers' search and choices and personalize digital ads and online product recommendations. These programs may be more or less aggressive toward the consumers. Financial companies use knowledge of financial advisers to generate robo-advice that provides clients with guidance. For example, the Personal Advisor Services of the investment company Vanguard combine investment advising by robots and human advisors (for other examples, see https://hbr.org/2018/01/artificial-intelligence-for-the-real-world). The content of advice, notably regarding the composition of portfolios, impacts how the surplus is shared between providers of financial products and investors. Finally, via the training of algorithms, chat bots can learn one's social preferences from past conversations and may thereby affect the utility of another party. For instance, the chat bot Tay was influenced by politically incorrect and racist messages on Twitter and started to release also inflammatory messages through this learning process, leading to its withdrawal (see https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism). These examples illustrate how social preferences may interfere with the training of AI.

difficult to track all humans involved in the history of one particular technological action and to attribute responsibility to one single person such as the programmer of the algorithm or its first user. 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?[4]

This paper analyzes the causal effect of changing individuals' pivotality for an algorithm's decisions on social behavior by means of an online experiment. We seek to answer to what extent the introduction of Big Data and adding training data sources affect the selfishness or prosociality of individuals' decisions. Increasing the number of sources of training is expected to reduce the prosociality of decisions for different reasons. First, the structure of incentives changes. Increasing the number of training sources for the AI's allocation choice reduces the expected future payoff from one's current decision via this AI's decision. In other words, it puts more weight on the player's current decisions and this may lead to more selfish choices to compensate for the loss of influence of ones' current decisions on one's future payoff. Second, it may offer excuses for selfish behavior to individuals who face a trade-off between self-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. Thus, as part of Big Data, individuals whose responsibility for their own fate and that of others is reduced may act more to their own selfish advantage. To disentangle between these reasons behind the expected reduction of prosocial choices when training data come from a higher number of players, we also studied the decisions of individuals who train an algorithm that impacts only a third party rather than themselves. These individuals are fully pivotal but they gain no personal future payoff from their current decisions. We can thus isolate the effect of pivotality for AI training on moral behavior by taking away personal monetary incentives in the AI's prediction. Here, we further ask whether the prosociality of training data people claim for others matches the standards they implement for decisions affecting their own payoff.

To answer our questions, we designed an experiment in which the ethics of AI was addressed through monetary allocation choices that could 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

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

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. The algorithm had to predict a hypothetical allocation choice of the dictator in a 31st period. This prediction determined the payoffs of the dictator and the receiver in this period and amounted to half of their total payoff. This made the training data meaningful in terms of incentives. As mentioned earlier, outside of the laboratory algorithms are typically using data from many sources, even if they provide personalized recommendations to individuals. This is precisely the advantage of the laboratory to let us represent a "what if" scenario, such as this Full Pivotality benchmark that does not intend to mimic reality but allows us to measure, through various treatments, the extent to which individuals adjust their behavior when responsibility diffuses, such as in the domain of AI training.

Our treatments varied the sources of the algorithm's training data. The variations allowed us to manipulate the pivotality of individuals' behavior for the AI's training and control for monetary incentives in the algorithm's prediction, diminishing or increasing the perceived individual responsibility of participants for arising inequality in the AI's allocation choices. In comparison with the baseline Full Pivotality treatment we have just described in the previous paragraph, in the No Pivotality treatment the algorithm made a decision in the 31st period based not only on the preferences of the player but also on the preferences of 99 other players, as expressed by their choices in the first 30 periods. These treatments vary both the degree of pivotality and the structure of incentives. We included the Full Pivotality-Others treatment in which a dictator was the only source of training of the AI, but the AI's decision impacted the payoffs of a pair of participants other than the one used for training. Thus, the weight of one's current decisions on one's future payoffs through the AI's choice was null and comparable to the No Pivotality treatment. However, there was full responsibility in the determination of future payoffs via training, like in the Full Pivotality treatment.Finally, the Shared Pivotality treatment represents an intermediate case between the Full Pivotality and the No Pivotality treatments. In this treatment the algorithm was trained by the 30 decisions of the individual and the 30 decisions of one other dictator, and its decision in the 31st period determined the payoffs of both pairs of participants.

Our findings revealed that pivotality in generating training data for the AI has a crucial influence on individuals' behavior. Dictators who accounted for a negligible part of the algorithm's training data and who had no whatsoever responsibility in determining others' payoffs (No Pivotality treatment), and those who shared responsibility for the algorithmic choices with another dictator (Shared Pivotality treatment), behaved significantly more selfishly than their counterparts who were fully pivotal (Full Pivotality and Full Pivotality-Others treatments). 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.

Importantly, we found that when dictators were fully pivotal for training an AI whose

predictions impacted the payoffs of a pair other than their own, they did not make significantly different payoff allocations than dictators who generated training data that solely influenced their own outcome. There was no discrepancy between an AI's prosociality individuals demanded when they were not monetarily affected by its choices and the principles they taught the AI to apply to themselves. This indicates that the higher selfishness observed when pivotality diffused was not motivated by the change in the structure of incentives (the fact that the current decisions had less impact on one's future payoffs), but by the change in the feeling of responsibility for others' payoffs.

Finally, we explored the dictators' beliefs about other dictators' behavior because anticipating others' selfishness might influence one's decisions when the AI is trained with external sources. We found that the selfishness of decisions increased when dictators believed that the AI was trained with the data of other selfish dictators in the context in which they were fully or partly pivotal for others, but not when they were not pivotal. This is interesting because in both extreme cases, the dictator's payoff in period 31 was (almost) fully determined by others' preferences. This suggests that the influence of beliefs about others' selfishness on one's choices is associated with the responsibility for determining others' payoffs.

Overall, our results show that when various agents train machines and accountability for ethical technological actions is mitigated, individuals behave more selfishly. Through its influence on training data, the dilution of responsibility has a major impact on the ethics of AI. One implication is a need for designing machines that embed explicit ethical guidelines, rather than using a simple aggregation of revealed preferences. This raises 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 permitted by the diffusion of pivotality.

Our study contributes to several strands in the literature. First, it 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 (*e.g.*, Bonnefon et al., 2016; Awad et al., 2020, in the context of moral dilemmas of autonomous vehicles). Compared to these investigations that elicit individuals' reported preferences, the novelty of our study is using revealed preferences through monetary choices that generate training data for an algorithm that makes consequential predictions. 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). 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 (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. In our experiment the AI learned from the current players' actions and there was no strategic interaction between the players and the AI. Importantly, as it is the case with many self-learning systems, individuals were aware of training an algorithm. Indeed, we are not investigating the impact of the awareness of training an algorithm, but the impact of pivotality when knowing that one's decisions are used to train an AI. Following the European Union's proposed artificial intelligence regulation, we made it clear to participants that their behavior created training data that would later be fed into an actual machine learning algorithm.[5] Due to these guidelines and the growing use of intelligent algorithms, we expect the investigated scenarios to occur increasingly in everyday life. It has further 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. 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. 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 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

[5]The proposed AI regulation of April 21, 2021 demands that "Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use." (Title IV, Article 52). Another illustration is provided by a report adopted in September 2021 by the European Parliament on the social rights of workers on digital platforms which notably requires the transparency of algorithms. In Spain, digital platforms are already obliged to inform workers' representatives of the rules on which the algorithms used by these platforms are based and which affect the conditions of access and retention in employment and remuneration. We acknowledge that in reality individuals are not always aware of how their data are used. For instance, the Harvard Business Review surveyed 900 people in five countries, of whom 97% were concerned that their data was being misused, even if they were unclear about what data exactly was being collected. (https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust). However, there are more and more instances in which it becomes difficult to ignore that there is an AI using one's data. For example, consumers can easily recognize that their decisions (*e.g.*, the products or services they search for and buy on the internet) influence the ads that are displayed on their computer screens following their purchases.

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. Individuals were shown to be in favor of moral or socially desirable behavior and policies, as long as they are not harmful to themselves – see the "Not in my backyard" literature in the context of nuclear waste repositories (Frey et al., 1996; Frey and Oberholzer-Gee, 1997) or the studies reported in Bonnefon et al. (2016) revealing a preference for utilitarian self-driving cars only when not being the car driver oneself. 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.

2 Design and Procedures

In this section, we first present the design of the experiment[6] and our treatments. Afterward, we describe our experimental procedures.

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 first 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 B10 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 (see Appendix B for an English translation of the instructions). Each participant was anonymously paired with another participant and was randomly assigned either the role of the dictator

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

("participant A") or of the receiver ("participant B"). Matching and roles were kept fixed throughout the experiment. All participants were informed upfront that the dictator's 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 and Y. 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 his or 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, whether the dictator was in an advantageous or disadvantageous relative position in the pair, and whether there was a conflict of interest between dictator and receiver or option X was Pareto dominant. The aim was to identify the participant's distributional preferences. 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 Appendix C lists all pairs of options. 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.[7]

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 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. For the decision of the machine learning tool, one of the six games represented by a dashed line in

[7]Previous work in the field of experimental economics discusses the question of whether to pay one or all rounds in multi-period experiments. In a theoretical analysis, Azrieli et al. (2018) prove that in rather general strategic environments with few assumptions on utility functions, paying for only one period is the best and only incentive-compatible mechanism. See also Cox et al. (2015) when risk is involved. For eliciting social preferences that inevitably requires multiple decisions per individual, Charness et al. (2016) do not make a strong recommendation for either paying one or several rounds, and refer to the variety of methods used in the literature. In our case, it was crucial to make period 31 salient and have participants pay strong attention to each of their decisions in stage 1. This is why we paid only one choice in stage 1.

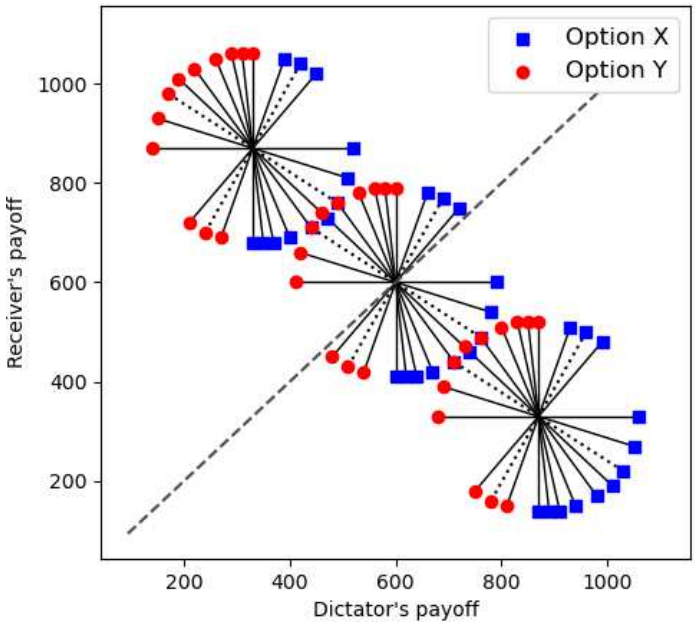

Figure 1: Dictator Games

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.

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, while the other three pairs of options are summarized in Table C3 in Appendix C. In the first decision, option Y was Pareto-dominant. 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 elicited participants' familiarity with artificial intelligence and machine learning, their confidence in these technologies, and assessed their understanding of the functionality of the random forest algorithm. We asked dictators about their satisfaction with the AI prediction in period 31, and how accurate the implemented decision reflected their preferences.

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, No Pivotality, Shared Pivotality, and Full Pivotality-Others. Figure D1 in Appendix D presents a simplified overview of these treatments.

Our baseline condition is the *Full Pivotality treatment*. It exactly follows the above described experimental design.[8] The random forest algorithm built exclusively on the 30 choices of the respective dictator for its predictions, and these data were only used for 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, and had full monetary incentives in what data s/he transferred as training to the algorithm. One can consider this treatment as a "what if" scenario in which an AI that usually relies on a broad range of data sources is trained exclusively with the behavior of one individual.

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.[9] Hence, the dictator was only responsible for 1% of the training data used by the random forest algorithm. Being not pivotal for the AI's resulting decision, the dictator also has hardly any monetary incentives in how to train the AI. This condition is consistent with the way training of intelligent systems often occurs. Though AI entirely relies on observed patterns in human behavior, no individual can solely be made accountable for the outcome.

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. 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

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

[9]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.

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. 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 contrast to the No Pivotality treatment, all dictators who contributed to the AI training were non-marginally monetarily affected in period 31 by their own preferences through the AI's prediction, since their decisions accounted for 50% instead of 1% of the training data for the AI. Therefore, two opposing mechanisms could simultaneously influence the dictator's decisions. On the one hand, compared to the Full Pivotality treatment, pivotality was reduced through a lower weight of each decision on one's future payoffs (which might increase selfishness, but to a lesser extent than in the No Pivotality treatment). 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 (which might decrease selfishness).

Finally, in the *Full Pivotality-Others treatment*, we used the 30 decisions of the dictator in the pair as training data for the AI that decided for another, third pair in the same session. Meanwhile, 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. Hence, the dictator in the Full Pivotality-Others treatment was fully pivotal for the AI's decision in another pair but had no monetary influence on the AI prediction in the own pair. Thus, the weight of the dictators' current decisions on their future payoff through the AI training was null, making this treatment close to the No Pivotality treatment in terms of its incentive structure. In terms of responsibility, however, decision making in this treatment is similar to the Full Pivotality treatment. If decisions in the Full Pivotality-Others treatment were closer to those in the No Pivotality treatment, it would suggest that the incentive structure (the weight of current decisions in future payoffs) drives behavior; if they were closer to those in the Full Pivotality treatment, it would suggest that responsibility for the AI's training itself matters more. This treatment was not intended to mimic real-world situations but to isolate the effect of taking away the monetary incentives for influencing the AI's prediction. Thus, with this treatment we can estimate the impact of removing or diffusing pivotality for the AI's training in the No Pivotality and Shared Pivotality treatments. The Full Pivotality-Others treatment also informs us on which ethical guidelines participants would want intelligent systems to follow when they are not monetarily affected by the algorithm's decision.

In the three latter treatments (No Pivotality, Shared Pivotality, and Full Pivotality-Others), the monetary outcome of dictators crucially depends on the behavior of previous participants. Therefore, we elicited beliefs about past decisions as described above and later

correlated them with observed behavior in section 4. Testing the extent to which beliefs influenced behavior across these treatments is also informative on the players' motivation. This is especially the case when comparing the Full Pivotality-Others and the No Pivotality treatments in which players' payoff in period 31 depends on others (almost) exclusively.

2.3 Procedures

Due to the 2020 coronavirus pandemic, we implemented an online experiment using oTree (Chen et al., 2016) after recruiting students from the regular subject pool of the Frankfurt Laboratory for Experimental Economic Research (FLEX) through ORSEE (Greiner, 2015).

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%. We chose a binomial distribution of decisions 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 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.[10] There were no drop-outs 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 C4 in Appendix C summarizes the main sociodemographic variables in each treatment. There were 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. Apart from measuring the frequency of choosing the selfish option X, we estimate social preference parameters. Bruhin et al. (2019) built on Fehr and Schmidt (1999) and Charness

[10] For the Shared Pivotality treatment, we pre-registered 60 dictators and eventually collected data from 61 dictators for procedural reasons related to our companion paper (Klockmann et al., 2021). For the analysis reported in this paper, we relied on the 34 participants who took part in the Shared Pivotality treatment during the same time period as the participants in the other conditions.

and Rabin (2002) to set up a model of social preferences for two players that is fitted to the data, using maximum likelihood. 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. Note that this utility function also captures efficiency concerns because the parameters are not restricted in terms of sign. In fact, if both $\alpha, \beta > 0$, the dictator always puts positive weight on the receiver's payoff, independent of her relative position. If even $\alpha = \beta = 0.5$, the dictator seeks to maximize total welfare in any case.

Regarding the estimations of α and β, we followed the theoretical framework only referring to distributional preferences in the pair. The reasoning is as follows. If we took into account that the decision of the dictator also affects the payoff in period 31, we would add a multiplication factor to each π_D and π_R. For instance, in the Full Pivotality treatment, each decision affects the current payoff and the payoff in period 31 equally. Hence, each $\pi_i, i = D, R$ would be multiplied by two in the utility function. This, however, would not affect the estimated parameters as the values for α and β that maximize the log likelihood would remain unaffected if we simply multiplied the objective function with any positive factor.[11] Still, the utility function does not take the possible externality on another pair into account. However, when just adding the payoffs of the other dictator and receiver, we would implicitly assume that the dictator weigh her own payoff the same as the payoff of the other dictator, which is most likely not true. Finally, we also do not explicitly incorporate pivotality in the utility function. We do so to answer our research questions on how varying pivotality alters not only choices, but also revealed distributional preferences of the dictator. Apart from that, in contrast to the well-established theories from the literature on social

[11] To be precise, we followed the econometric strategy by Bruhin et al. (2019) that builds on a random utility model. When choosing an allocation $X = (\pi_D^X, \pi_R^X)$, the dictator' utility is given by $\mathcal{U}(\pi_D^X, \pi_R^X; \alpha, \beta, \sigma) = u_D(\pi_D^X, \pi_R^X; \alpha, \beta) + \varepsilon_X$, where u_D is the deterministic utility from above and ε_X is a random component following a type 1 extreme value distribution with scale parameter $1/\sigma$. The likelihood of choosing allocation X over Y then reads (see Bruhin et al., 2019)

$$Pr(X|\pi_D^X, \pi_R^X, \pi_D^Y, \pi_R^Y; \alpha, \beta, \sigma) = \frac{\exp(\sigma u_D(\pi_D^X, \pi_R^X; \alpha, \beta))}{\exp(\sigma u_D(\pi_D^X, \pi_R^X; \alpha, \beta)) + \exp(\sigma u_D(\pi_D^Y, \pi_R^Y; \alpha, \beta))}.$$

Any positive multiplication factor in the deterministic utility u_D only inversely proportionally changes σ, but leaves the estimated values of α and β unchanged.

preferences, it also not clear how to incorporate pivotality in the utility function.

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 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 prosocial preferences can actively increase the fairness of payoffs by training the AI with egalitarian decisions.

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. We expect the dictator to be less reluctant in allocating ECU to himself or herself at the expense of the receiver in this treatment. In line with Bénabou et al. (2018), reduced pivotality may induce individuals to feel less responsible for their selfish choices than in the baseline condition and to downplay negative externalities. Adding 99 other dictators' decisions in the AI training data may lower individuals' perceived responsibility and create an excuse for making more selfish decisions. This may also be due to the modified incentive structure: in each period, the expected utility of one's current decision is decreased since this decision weights a hundred times less in the No Pivotality treatment compared to the Full Pivotality treatment for period 31. This may lead to more selfish choices to compensate for this loss of future expected utility from one's current decisions.

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 reduces the expected utility in period 31 from one's current decision and offers an excuse for selfish behavior. On the other hand, as the dictators' decisions now also affect another group and thereby create an externality, dictators might be willing to reduce inequality for these individuals, by making less selfish decisions than in the Full Pivotality treatment. Still, in our companion paper Klockmann et al. (2021) we show that informing dictators about an externality of their training on others did not affect their allocation decisions. Even internalizing the externality by introducing monetary interdependence did not alter behavior when dictators could only benefit repeatedly from training the AI with selfish choices. Therefore, we conjecture the first channel that concerns 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 Full Pivotality treatment, their decisions do not affect themselves in period 31. The structure of incentives is such that the expected monetary utility from one's current decision for period 31 is null in this treatment. Still, the dictator is fully pivotal for the decision of the AI. We conjecture that dictators may behave more selfishly than in the Full Pivotality treatment to compensate for this reduced expected utility of each current decision, and this may be reinforced the more players believe that other decision makers are selfish. If this conjecture is rejected (*i.e.*, players behave like in the Full Pivotality treatment), this would indicate that individuals in our study mainly react to their degree of responsibility when training the AI. One could argue that the Full Pivotality-Others treatment, as the Shared Pivotality treatment, further introduces an externality by training an AI that makes an allocation for another pair of players. Based on the results of our companion paper Klockmann et al. (2021), we expect no influence from this externality. We thus focus on the aspect of altered incentives while preserving the pivotality when comparing the Full Pivotality-Others and Full Pivotality treatments.[12]

Conjecture 3. *Compared to the Full Pivotality treatment, dictators in the Full Pivotality-Others 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 Full Pivotality-Others treatment than in the Full Pivotality treatment.*

4 Results

To identify how varying the degree of pivotality for training data generated for an artificially intelligent algorithm affects dictators' 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 1 and 2 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

[12] In Klockmann et al. (2021), the Baseline condition was similar to our Full Pivotality treatment. In an Externality treatment, the dictator's choices generated training data for the AI's allocation decision in their group and for another group that would participate in the same experiment in the future, with no impact on payoffs in the current pair. The Offspring treatment was similar to the Externality treatment, except that the earnings of the first pair were affected by the prediction made by the AI for the future pair. Both the dictator and the receiver in the past generation received half of the payoff of the corresponding player in the future generation (the subjects playing the Shared Pivotality treatment here), in addition to their other earnings. The results showed that the dictators' decisions did not react to the introduction of such externalities.

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.[13]

Table 1: Overview of the Frequency of Choices of the Selfish Option X across Treatments

Treatments	Nb Obs.	Option X	Option X [Restricted Sample]
Full Pivotality	34	70.29% (0.029)	66.01% (0.046)
No Pivotality	29	79.20% (0.024)	80.27% (0.035)
Shared Pivotality	34	77.25% (0.018)	80.07% (0.027)
Full Pivotality-Others	30	70.89% (0.028)	68.52% (0.043)
Treatment comparisons *(p-values)*			
Full Pivotality vs. No Pivotality		0.025	0.019
Full Pivotality vs. Shared Pivotality		0.045	0.010
Full Pivotality vs. Full Pivotality-Others		0.883	0.693

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.

We start with comparing the Full Pivotality treatment with the Full Pivotality-Others treatment. In both cases, dictators were fully responsible for training the algorithm but the weight of each current decision on the expected future payoff differed dramatically. Here, we found no significant difference in the share of selfish choices dictators made ($p = 0.883$ in the full sample and $p = 0.693$ in the restricted sample). Neither the estimated social preference parameter α did differ significantly ($p = 0.676$) between these two treatments, nor the estimated β parameter ($p = 0.850$).

The comparison reveals no effect of taking away the monetary incentives of the dictator for selfish training to influence the AI's prediction in the own group, while keeping the pivotality for the AI's decision constant.[14] We found no evidence of a discrepancy between how participants would expect an AI to function for themselves and what behavior they would implement for others. What possibly motivated them is a minimal sense of belonging with the participant in the role of dictator in the other group. Therefore, what seemed to actually matter more than the structure of incentives was the decision-maker's responsibility for training the AI, regardless of whether this training was used for one's or another pair's payoffs.[15]

[13] In addition, Figures D2 and D3 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.

[14] As noted earlier, our companion paper Klockmann et al. (2021) shows that there is no evidence of an effect on behavior of introducing an externality (the fact that players' choices determined another pair's payoffs).

[15] We acknowledge that the algorithm relies on different amounts of data across treatments, and that this

Table 2: Estimated Parameters of Social Preferences across Treatments

Treatments	Nb Obs.	Dictators	α	β
Full Pivotality	1020	34	0.082 (0.048)	0.394 (0.051)
No Pivotality	870	29	−0.050 (0.044)	0.253 (0.051)
Shared Pivotality	1020	34	−0.047 (0.048)	0.246 (0.037)
Full Pivotality-Others	900	30	0.054 (0.048)	0.382 (0.046)
Treatment comparisons *(p-values)*				
Full Pivotality vs. No Pivotality			0.040	0.050
Full Pivotality vs. Shared Pivotality			0.056	0.020
Full Pivotality vs. Full Pivotality-Others			0.676	0.850

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.

This analysis contradicts Conjecture 3, and it is summarized in Result 1.

Result 1. *Individuals who taught an AI algorithm that exclusively decided for others did not behave differently than those who generated training data that solely influenced their own outcome. The percentage of choices of the selfish option X and the estimated social preference parameters were not significantly different in the Full Pivotality-Others treatment than in the Full Pivotality treatment.*

Tables 1 and 2 reveal that reducing pivotality for training data, without introducing an externality of players' decisions on another pair of participants, 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). The positive value of α and β in the two Full Pivotality treatments differs from what is usually observed. The representative dictator in the Full Pivotality treatment values the payoff of the receiver

could have made the prediction less noisy in the No Pivotality treatment, though it turned out that this was actually not the case (see Table A1 in Appendix A). We have no information on whether dictators believed in different noise in the prediction of the algorithm in the No Pivotality treatment compared to the Full Pivotality treatment, which might have influenced their decisions. It would have been useful to measure ex ante the dictators' perceptions of the environment noisiness. Nevertheless, our ex post final questionnaire indicates that their satisfaction with the prediction in period 31 and their assessment of how accurately the AI's decision matched their true preferences were not significantly different across treatments. These ex post evaluations suggest that perceived noisiness in the AI training may not have been a serious source of concern across treatments.

positively ($\alpha > 0$ and $\beta > 0$ with $p = 0.084$ and $p < 0.01$, respectively), regardless of whether the other player is better or worse off. This cannot be explained by the fact that current decisions entailed another future payoff in period 31, since this did not apply in the Full Pivotality-Others treatment. The positive weighting of the receiver's payoff independent of the relative position could indicate higher concerns for social welfare when dictators are fully pivotal. Indeed, efficiency concerns were significantly strengthened when pivotality was increased. Table C6 in Appendix C shows that dictators in the Full Pivotality treatment, compared to those in No Pivotality, chose significantly less frequently the selfish option X when choosing the alternative option Y maximized the sum of payoffs ($p = 0.027$).

As mentioned in section 2.2, adding 99 sources of training data not only reduced the dictator's pivotality for the prediction of the AI, but also lowered the expected future payoff of the dictators' current decisions. Still, as summarized by Result 1 describing the comparison of Full Pivotality and Full Pivotality-Others, taking away the monetary incentives while maintaining the pivotality for the algorithm's decision does not affect the dictators' behavior, nor the revealed social preferences. Therefore, we conclude that the observed significant changes in behavior in the No Pivotality treatment resulted from vanished responsibility and the possibility of self-excuses rather than altered incentive structures.

This analysis supports Conjecture 1 and leads to Result 2.

Result 2. *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 C6 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, diffusing pivotality by letting dictators from two groups independently influence the AI's training that determined future payoffs in both groups resulted in less egalitarian payoff allocations and in the estimation of more selfish preferences. Participants in the Shared Pivotality treatment behaved very similarly to their counterparts in the No Pivotality treatment, although the impact of one's individual decisions on future payoffs was considerably more reduced in the latter treatment. Wald tests using estimation results from Table C8 in the appendix conclude to no significant differences between these two treatments ($p > 0.58$ for all three specifications).

This analysis supports Conjecture 2. It is summarized in Result 3.

Result 3. *Individuals made more selfish decisions when pivotality was diffused by letting only 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 and the estimated social preference parameters were significantly lower in the Shared Pivotality treatment than in the Full Pivotality treatment.*

Table C8 in the appendix reports regression results of the share of option X on treatment dummies and several demographic variables with the Full Pivotality treatment as the baseline and omitted category. Column (1) confirms the t-test conducted above. Columns (2) and (3) show that the observed treatment differences between Full Pivotality and No Pivotality, as well as between Full Pivotality and Shared Pivotality, remain significant at the 5 or 10%-level when controlling for gender and age, and at the 10%-level when adding an indicator for studying economics and the current semester as further control variables.

Finally, we report an exploratory analysis of the relationships between beliefs and behavior. Before informing participants about their payoff from the AI's prediction, we elicited the dictators' beliefs about the others' training data in all treatments, except Full Pivotality. There were no significant differences in average beliefs across treatments ($p > 0.10$ in all pairwise t-tests, see Table C9 in Appendix C). Dictators did not anticipate treatment effects on others' choices and there is no evidence that they manipulated their beliefs strategically depending on their degree of pivotality. Still, beliefs may have affected behavior differently across treatments. To explore this, Table 3 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. 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 elicited scenarios (see section 2.2).

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.

The picture is slightly different when estimating this effect for each treatment separately. We detect the strongest correlation between beliefs and actual behavior in the Full Pivotality-Others treatment. In this treatment, a dictator who believed that the dictator in the other group put least weight on the receiver's payoff (*i.e.,* Belief Option X = 3) picked option X herself more than 80% more frequently. This supports the notion that participants did not counterbalance their exogenous power position when training an intelligent system

Table 3: 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

for a third, independent, group if they expected the participant in the other group to behave in a similar way, favoring the dictator's payoff.

In contrast, beliefs had no significant impact on dictators' behavior in the No Pivotality treatment. This difference in the impact of beliefs across treatments is interesting because in both the No Pivotality and Full Pivotality-Others treatments, the dictator's payoff in period 31 was (almost) fully determined by others' preferences. The impact of beliefs did not depend on whether one's payoff was determined by other dictators in period 31. It seems that when they knew they were not pivotal, individuals did not need beliefs about others' selfishness to justify their own selfish choices; they made more selfish choices independently of what they believed other people would do in the same position. This was not the case when they were fully pivotal. The importance of beliefs about others' preferences on one's behavior is associated with the responsibility for determining others' payoffs. This is consistent with Falk et al. (2020) who provided evidence that to behave more selfishly while keeping a good image, individuals tend to rely on the narrative that their actions do not influence an outcome. In our case, individuals do not need narratives, they simply react to the change in pivotality.

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, but only when they were to some extent pivotal for others' payoffs.*

5 Discussion and Conclusion

Before making decisions, artificial intelligence has to be trained based on data. In various situations such as when judges employ decision support systems that predict recidivism

or are used for bail decisions, when financial companies use robots to advise investors on their portfolio composition, or when human resources systems develop the use of intelligent automation, the data partly stems from humans and their behavior. From an ethical perspective, this means that individuals are responsible for the data they generate for the training of AI, and that the more they behave in a polarized manner themselves, for example, 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 experiment, we studied to what extent varying the common knowledge pivotality of individuals in the training of an artificially intelligent algorithm affected 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. We provided evidence that this effect was driven by altered individual responsibility (that is, the power on the own or others' fate) and not by different incentive structures (that is, the expected future payoff of one's current decisions through the AI's training).

Overall, our findings suggest that humans' tendency to behave less morally when feeling less responsible is also reflected in AI training, except when pivotality, and related responsibility, diffuses due to Big Data and immoral choices become obfuscated. This interpretation is backed by the positive correlation we identified between the participants' beliefs about others' selfishness reflected in the training and the selfishness in their own choices, only when they were pivotal for others. When they were not pivotal, in contrast, dictators behaved selfishly independently on their beliefs about others' selfishness. 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 tends to translate into more selfish or less moral behavior. This is a challenging dilemma. In our companion paper Klockmann et al. (2021), we considered another aspect of AI training and 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.

In terms of policy implications, our findings suggest a need for attributing explicit and salient individual responsibility to those affecting algorithmic predictions. Especially when algorithms are trained with a multiplicity of data sources, the changes in behavior towards more selfish decisions bear the risk of biased decisions of algorithms. Our findings mimic the insights of previous studies (particularly Darley and Latané, 1968; Bénabou et al., 2018; Falk et al., 2020) and show that undesirable behavioral patterns occur in the context of

intelligent algorithms as well. More than that, it has a larger impact as these algorithms will use the training data for multiple future predictions and decisions that will ultimately also affect others in the society. Our findings showing how reduced pivotality increases the selfishness of individuals' decisions emphasize the risk that selfishness develops through AI when individuals anticipate that machines learn from the data of many hands. Like for elections where every vote counts although each voter may feel powerless at the individual level, it is certainly important that companies remind their employees that ethics matters for every decision that they make – even if these individual decisions will be combined with that of many other employees to train algorithms. However, this is certainly not sufficient to prevent the development of "selfish algorithms", which is why we also suggest to incorporate general ethical principles in the design of these programs. In fact, with our findings we reopen a larger debate in the field of AI ethics mentioned by Coeckelbergh (2020) about whether computer scientists should make the training datasets for intelligent systems more diverse and perhaps even create "idealized" data, as Eric Horvitz, a technical fellow at Microsoft, suggested. Or should datasets reflect those human biases that exist in the real world? Should developers build affirmative action into their algorithms or should they create "blind" algorithms that mirror human decision-making and (social) preferences, even if this potentially increases inequality? "Human in the loop" interventions should certainly be involved to validate models, check AI's decisions, and flag harmful consequences, as recommended by Chui et al. (2018).[16]

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.

[16]"Human analysis of the data used to train models may be able to identify issues such as bias and lack of representation. Fairness and security 'red teams' could carry out solution tests, and in some cases third parties could be brought in to test solutions by using an adversarial approach." (Chui et al. (2018), p.41).

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.

Azrieli, Y., C. P. Chambers, and P. J. Healy (2018): "Incentives in experiments: A theoretical analysis," *Journal of Political Economy*, 126, 1472–1503.

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., U. Gneezy, and B. Halladay (2016): "Experimental methods: Pay one or pay all," *Journal of Economic Behavior & Organization*, 131, 141–150.

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.

Chui, M., M. Harrysson, J. Manyika, R. Roberts, R. Chung, P. Nel, and A. van Heteren (2018): "Notes from the AI frontier: Applying artificial intelligence for social good," *Working Paper McKinsey Global Institute*.

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.

Cox, J. C., V. Sadiraj, and U. Schmidt (2015): "Paradoxes and mechanisms for choice under risk," *Experimental Economics*, 18, 215–250.

Darley, J. M. and B. Latané (1968): "Bystander intervention in emergencies: Diffusion of responsibility." *Journal of Personality and Social Psychology*, 10, 215–221.

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.

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.

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 84.0%, a precision of 87.7%, a recall of 91.4%, and a F_1 score of 0.895. 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. In Table A1, we report the precision, recall, and F1 score of the random forest algorithm in each treatment. We conclude that noisiness of the AI prediction was not different across treatments.

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

Table A1: Performance Metrics of the Random Forest Algorithm by Treatment

Treatment	Dictators	Training Obs.	Accuracy	Precision	Recall	F_1 score
Full Pivotality	34	24	84.31%	89.93%	87.41%	0.8865
No Pivotality	29	2994	86.21%	93.89%	88.49%	0.9111
Shared Pivotality	34	54	77.45%	89.47%	81.93%	0.8553
Full Pivotality-Others	30	24	81.11%	91.27%	83.33%	0.8712

Notes: The table provides an overview of different performance metrics of the random forest algorithm, separate for each treatment. For each dictator, we made an 80:20 split, *i.e.*, 24 observations for training and 6 observations for testing.The column Training Obs. reports the number of training observations used for each dictator. For the treatment No Pivotality, we additionally trained the algorithm with the 30 decisions from each of 99 randomly selected previous dictators. For the Shared Pivotality treatment, we additionally trained the algorithm with the 30 decisions from one previous dictator.

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.

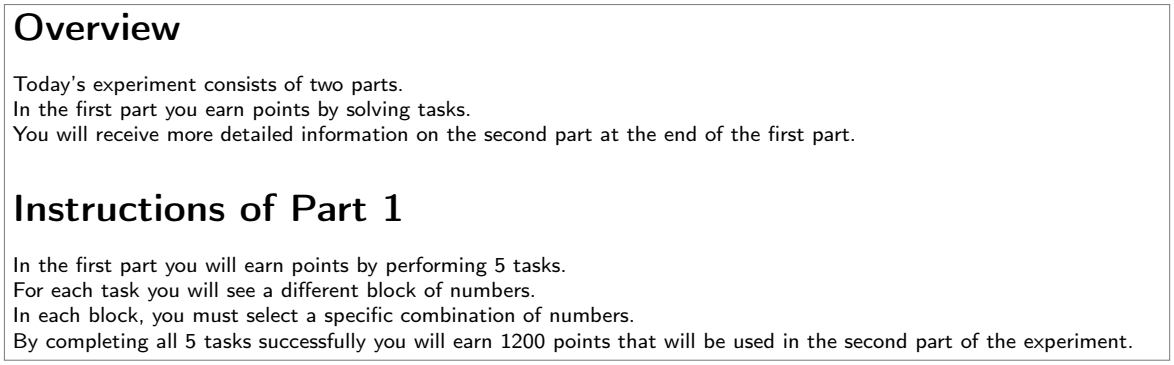

Figure B1: Instructions – Real Effort Tasks

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

Figure B2: 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 B3: 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 B4: 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 B4: 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 B4: 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 B4: 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 B5: 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 B5: 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 B5: 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 B6: 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 B7: 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 B8: 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 B9: 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.

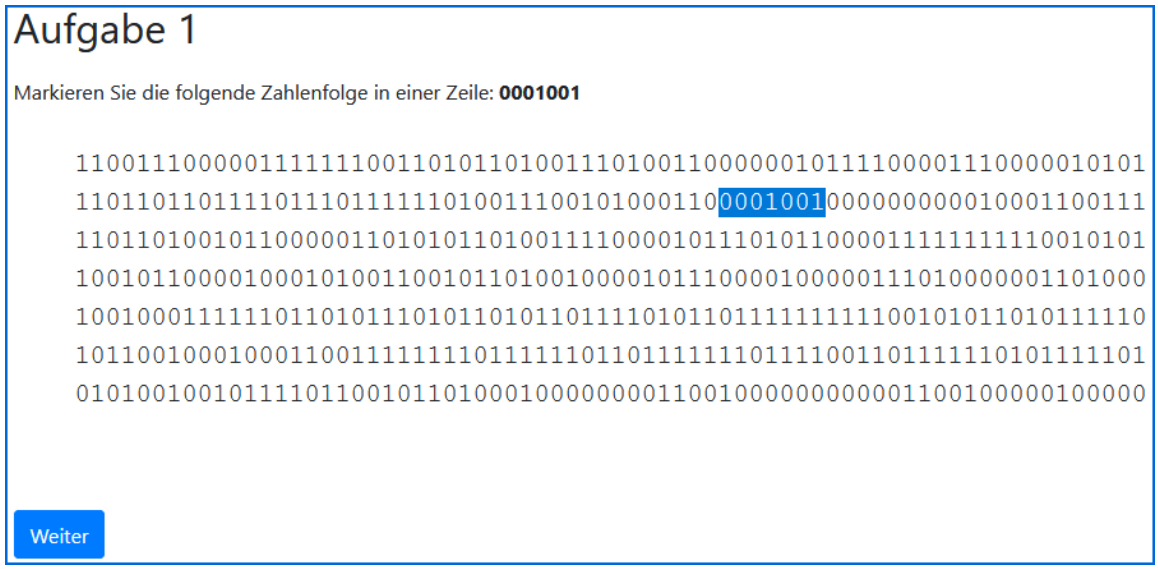

Figure B10: Real Effort Task

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

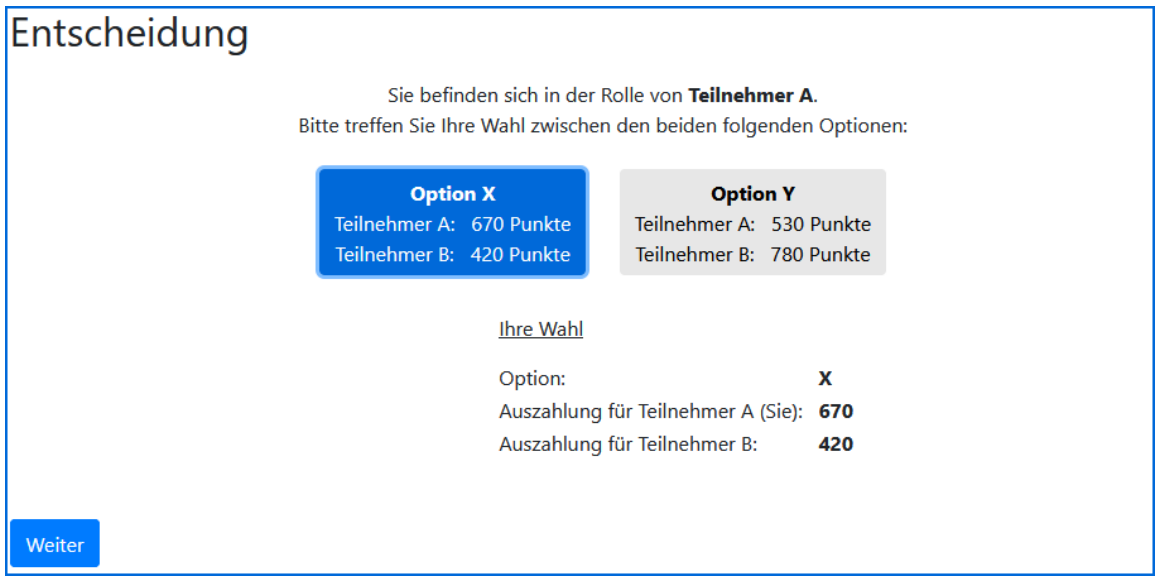

Figure B11: Decision Screen of the Dictator

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: Elicited Beliefs about Other Dictators' Behavior

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

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.

Table C4: 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. There are no significant differences across treatments at the threshold of 5%.

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

Treatments	Nb obs.	Option X [Advantageous]	Option X [Disadvantageous]	Option X [mixed]
Full Pivotality	34	57.94% (0.0417)	79.42% (0.0273)	73.53% (0.0355)
No Pivotality	29	68.62% (0.0426)	87.93% (0.0213)	81.03% (0.0240)
Shared Pivotality	34	67.65% (0.0305)	82.65% (0.0199)	81.47% (0.0180)
Full Pivotality-Others	30	57.67% (0.0392)	78.67% (0.0270)	76.33% (0.0297)
Treatment comparisons *(p-values)*				
Full Pivotality vs. No Pivotality		0.080	0.020	0.096
Full Pivotality vs. Shared Pivotality		0.065	0.343	0.050
Full Pivotality vs. Full Pivotality-Others		0.962	0.848	0.553

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

Treatments	Nb. obs.	Option X [X efficient]	Option X [Y efficient]
Full Pivotality	34	96.32% (0.0101)	48.63% (0.0502)
No Pivotality	29	96.26% (0.0098)	63.91% (0.0436)
Shared Pivotality	34	93.63% (0.0154)	62.16% (0.0336)
Full Pivotality-Others	30	95.56% (0.0143)	50.00% (0.0476)
Treatment comparisons *(p-values)*			
Full Pivotality vs. No Pivotality		0.967	0.027
Full Pivotality vs. Shared Pivotality		0.147	0.028
Full Pivotality vs. Full Pivotality-Others		0.656	0.844

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 C7: Relative Frequency of Choices of the Selfish Option X, by Treatment and Relative Inequality

Treatments	Nb obs.	Option X [X fairer]	Option X [Y fairer]	Option X [equal]
Full Pivotality	34	79.71% (0.0272)	57.65% (0.0420)	73.53% (0.0355)
No Pivotality	29	87.93% (0.0207)	68.62% (0.0440)	81.03% (0.0240)
Shared Pivotality	34	84.41% (0.0203)	65.88% (0.0319)	81.47% (0.0180)
Full Pivotality-Others	30	80.33% (0.0277)	56.00% (0.0411)	76.33% (0.0297)
Treatment comparisons *(p-values)*				
Full Pivotality vs. No Pivotality		0.022	0.077	0.096
Full Pivotality vs. Shared Pivotality		0.170	0.123	0.050
Full Pivotality vs. Full Pivotality-Others		0.873	0.782	0.553

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 C8: Regression of Relative Frequency of Choices of the Selfish Option X

	(1)	(2)	(3)
No Pivotality	0.089**	0.082**	0.065*
	(0.036)	(0.036)	(0.036)
Shared Pivotality	0.070**	0.067*	0.059*
	(0.034)	(0.035)	(0.034)
Full Pivotality-Others	0.006	-0.002	-0.013
	(0.035)	(0.035)	(0.035)
Female		0.033	0.051*
		(0.026)	(0.026)
Age		0.003	0.003
		(0.003)	(0.003)
Economics			0.061**
			(0.026)
Semester			0.004
			(0.004)
Constant	0.703***	0.602***	0.552***
	(0.024)	(0.073)	(0.075)
Number of observations	127	127	127
R^2	0.072	0.096	0.141

Notes: The table reports OLS estimates of the percentage of choices of the selfish option X in the 30 games on treatment dummies and demographics. The constant report the relative frequency of option X in the omitted treatment Full Pivotality. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C9: 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

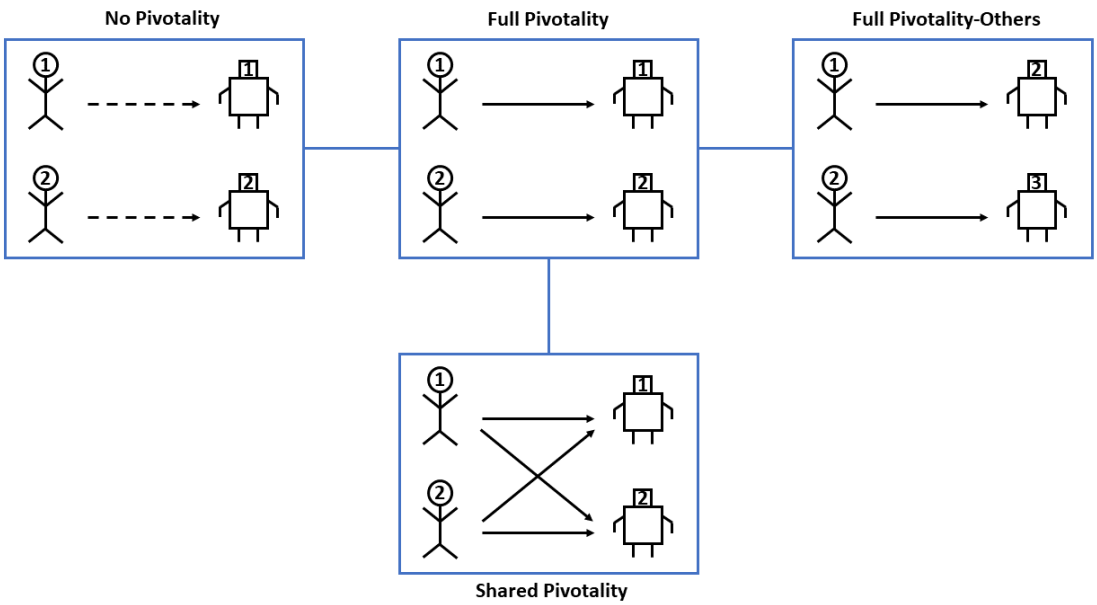

Figure D1: Illustration of Treatments

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.

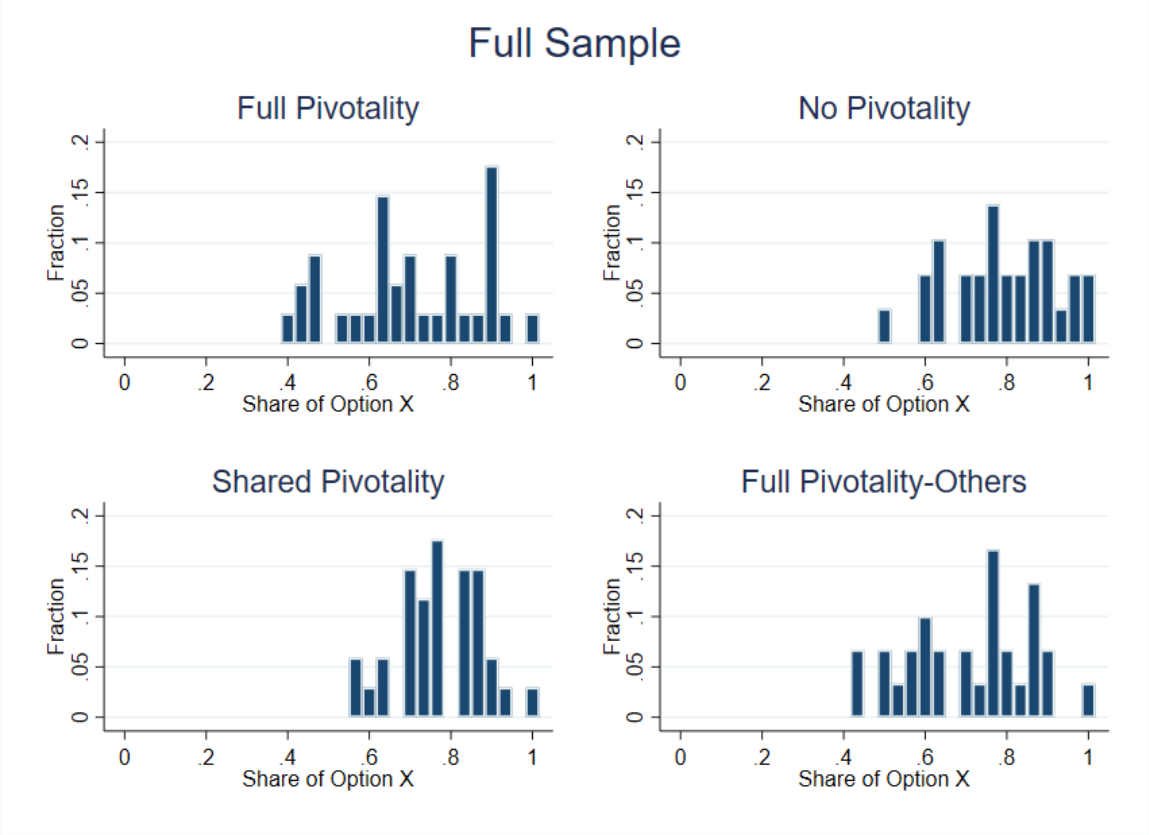

Figure D2: Distribution of the Shares of Selfish Choices by the Dictators, by Treatment

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.

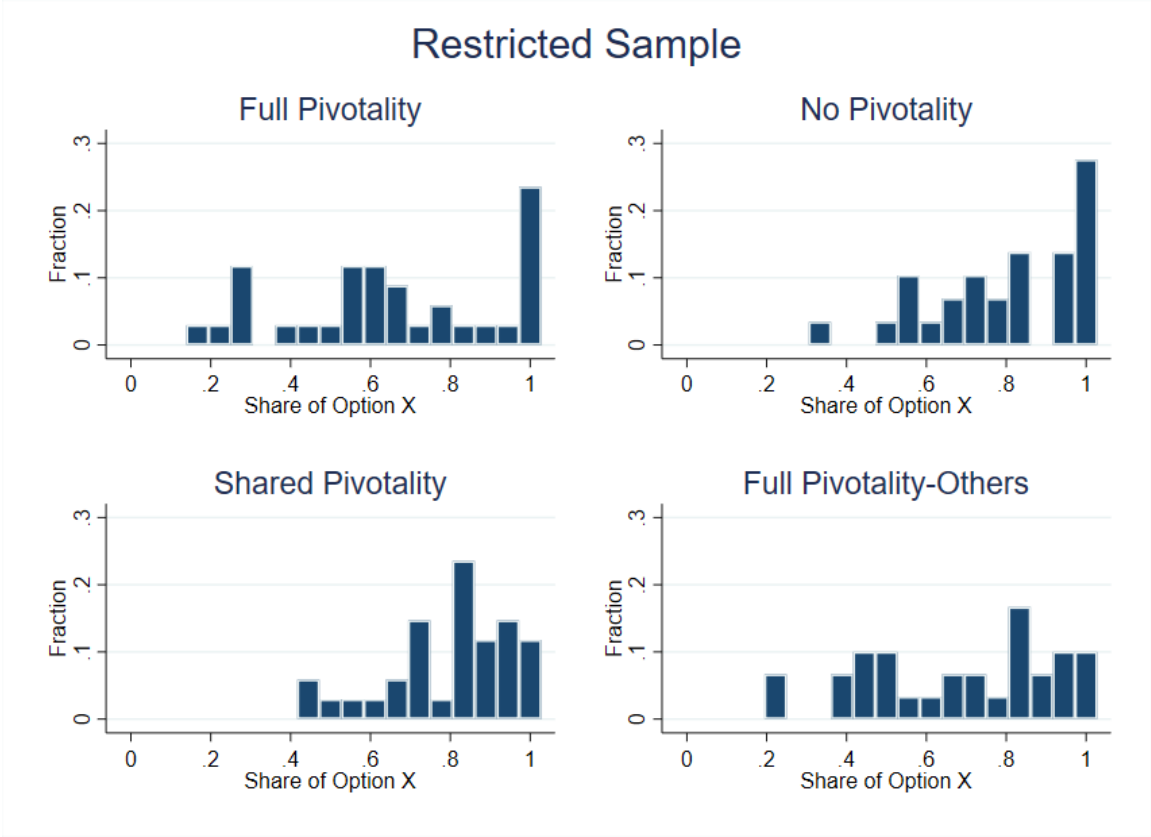

Figure D3: Distribution of the Shares of Selfish Choices by the Dictators, by Treatment (Restricted Sample)

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".