ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Klockmann, Victor; von Schenk, Alicia; Villeval, Marie-Claire

Working Paper

Artificial intelligence, ethics, and intergenerational responsibility

SAFE Working Paper, No. 335

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Klockmann, Victor; von Schenk, Alicia; Villeval, Marie-Claire (2022) : Artificial intelligence, ethics, and intergenerational responsibility, SAFE Working Paper, No. 335, Leibniz Institute for Financial Research SAFE, Frankfurt a. M., https://doi.org/10.2139/ssrn.4002578

This Version is available at:
https://hdl.handle.net/10419/249178

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Victor Klockmann | Alicia von Schenk | Marie Claire Villeval

Artificial Intelligence, Ethics, and Intergenerational Responsibility

SAFE Working Paper No. 335 | January 2022

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de
Electronic copy available at: https://ssrn.com/abstract=4002578

Artificial Intelligence, Ethics, and Intergenerational Responsibility

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

December 21, 2021

Abstract

In more and more situations, artificially intelligent algorithms have to model humans' (social) preferences on whose behalf they increasingly make decisions. They can learn these preferences through the repeated observation of human behavior in social encounters. In such a context, do individuals adjust the selfishness or prosociality of their behavior when it is common knowledge that their actions produce various externalities through the training of an algorithm? In an online experiment, we let participants' choices in dictator games train an algorithm. Thereby, they create an externality on future decision making of an intelligent system that affects future participants. We show that individuals who are aware of the consequences of their training on the payoffs of a future generation behave more prosocially, but only when they bear the risk of being harmed themselves by future algorithmic choices. In that case, the externality of artificially intelligence training induces a significantly higher share of egalitarian decisions in the present.

Keywords: Artificial Intelligence, Morality, Prosociality, Generations, Externalities

JEL Codes: C49, C91, D10, D62, D63, O33

We are grateful to Ferdinand von Siemens, Matthias Blonski, Guido Friebel 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 4, 60323 Frankfurt, Germany. Center for Humans & Machines, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. klockmann@econ.uni-frankfurt.de.

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

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

Electronic copy available at: https://ssrn.com/abstract=4002578

1 Introduction

Technological progress moves artificial intelligence (AI) from performing narrow, well-defined tasks in closed environments to becoming an actor in the real world that actively decides for humans. This increasing autonomy induces a host of ethical questions related to moral machines and teaching robots to distinguish right from wrong. In extreme cases, AI systems may experience moral dilemmas with life-or-death consequences, such as in the context of self-driving cars (Bonnefon et al., 2016; Shariff et al., 2017; Awad et al., 2018), kidney exchanges (Dickerson and Sandholm, 2015) and personalized medicine. But in everyday life, algorithms are making more and more decisions that have a moral component and affect a large number of people. Consider, for instance, computers that approve credit card transactions (Bhattacharyya et al., 2011; Adewumi and Akinyelu, 2017), AI tools managing time schedules at the workplace, predictive policing, for which AI forecasts where crimes are likely to occur (Ensign et al., 2018), or algorithms that advise decision making in courts by predicting who will re-offend (Brennan et al., 2009; Flores et al., 2016), and that recommend bail decisions (Kleinberg et al., 2018; Cowgill, 2018). If an artificially intelligent system shall make decisions on behalf of an individual or a society, it needs to learn the preferences of that individual or society – this comprises, in particular, social preferences.

AI might exhibit biases or inefficiencies in some of its predictions. If this occurs, the fault will lie – apart from their programming – mostly in how the algorithms were taught and trained. Indeed, many AI systems learn from human choices of today to augment decision making of tomorrow. Biased predictions could thus result from contamination of the training data, for example, through feedback loops or discriminatory choices that then substantially affect decision making of the algorithm (Rahwan et al., 2019; Cowgill and Tucker, Forthcoming).[1] As long as our society exhibits biases or unethical behavior, these biases should be reflected in the data that are collected and the choices self-learning algorithms make for us and for others, today and in the future. Explicitly identifying and counteracting those biases, *e.g.* via affirmative actions regulating the algorithm's outcome, is a multilayered challenge and often hard to resolve. Hence, focusing on the transmission of social preferences through the training of an AI, our first research question is: How does human behavior that expresses selfish versus prosocial preferences react to the use of today's decisions for the training of an AI that will make decisions in the future?

The development of AI raises the question of attribution of responsibility. In his perspective on AI ethics, Coeckelbergh (2020) describes obstacles when attributing moral responsi-

[1]For example, algorithmic systems allocating police officers to parts of the city based on crime rate data may lead to feedback loops. With the police relying on the algorithm without patrolling also in other areas, they induce the system to allocate officers in the future to the same neighborhoods over and over again. In the domain of predictive justice, a frequent critique is that using algorithms may reduce the feeling of responsibility of judges who may be less willing to take the risk of deviating from the recommendation of the algorithm based on other judges' decisions. Conformity might then lead to decision errors in some circumstances.

Electronic copy available at: https://ssrn.com/abstract=4002578

bility in human-machine interaction. The guiding question is who is responsible for potential errors of biased algorithms or for immoral predictions. Though the AI's actions have moral consequences, it lacks consciousness, moral reasoning, free will, or emotions. Therefore, the system cannot be responsible for its decisions. Nevertheless, taking an anthropocentric ethical position and attributing responsibility to the humans designing or training the AI is difficult, as well. In his *Nicomachean Ethics*, Aristotle defines two conditions that must be satisfied for moral responsibility. First, each action must have its origin in the agent. From a normative perspective, this implies that having the possibility to decide means responsibility for one's choices. Second, the agent must be aware of her actions and her actions' consequences. Especially the latter requirement is difficult to satisfy in the case of humans training an AI. Human agents might not be fully conscious of the externalities generated through the seemingly neutral new technology. Certain characteristics specific to AI such as scalability or self-reinforcement through feedback loops aggravate this concern. Hence, our second research question is: Is it possible to strengthen individuals' sense of responsibility when training algorithms by informing them about the direct impact of their actions on training and by emphasizing the consequences for the well-being of future generations?

This study contributes to the recent literature on AI and ethics by highlighting the training aspect of AI and its ability to operate across individuals and generations. Observational data, when they exist, do not allow us to vary exogenously the externalities generated by the training of AI. Therefore, we designed an online experiment in which we manipulated the concern of players for a future generation of individuals. We tested whether an externality of today's training of an AI algorithm that would make decisions in the future influenced today's individuals' social preferences and decisions. Importantly, as is also required by the proposed artificial intelligence regulation by the European Union, participants were fully aware that their behavior trained an actual machine learning algorithm.[2] They also knew that they trained a sophisticated algorithm that did not simply play a mixed strategy. Rather, the algorithm aimed to predict their (social) preferences with the greatest possible accuracy by exploiting features of the decision space and randomization techniques. On the basis of this prediction, it then made an allocation decision on behalf of the individual.

The Baseline condition of our experiment comprised a set of 30 mini-dictator games inspired by Bruhin et al. (2019).[3] Participants were paired, and one pair member repeatedly chose between two options that determined the payoff allocation in the pair. By manipulating the sum of payoffs, payoff inequality, and whether the dictator was in an advantageous or

[2]The European Union's proposed AI regulation, released on April 21, 2021, contains the following paragraph "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)

[3]This particular type of game is of course much simpler than the situations faced by doctors, mediators and judges invoked at the beginning of the article, but it captures the main characteristics of the decision making process we are looking for: The game is non strategic and it entails an externality on others' well-being, making decision makers' social preferences potentially relevant.

Electronic copy available at: https://ssrn.com/abstract=4002578

disadvantageous position with both options, we measured revealed social preferences (*i.e.*, selfishness, advantageous and disadvantageous inequity aversion, spite, efficiency concerns). After participants made their 30 decisions, an intelligent (random forest) algorithm used these decisions as training data to predict a hypothetical allocation choice of the dictator in a 31st period, with monetary consequences for the two pair members. Using dictator games is simpler than using strategic games, as we did not need to account for the individuals' beliefs on how partners were affected by the perspective of training an algorithm.

The treatments varied: (i) whether the algorithm trained by the dictator's choices made an allocation decision only for the current pair of players or also for a future pair of participants and how this affected the current pair's payoff; and (ii) whether it was possible that the payoffs of the dictator and receiver from the AI's allocation choice were swapped.

The aim of the first variation was to increase the size of the externality. The Externality treatment had the same features as the Baseline condition, except that participants were informed that the dictator's choices generated training data also for another group that would participate in the same experiment under the same basic design in the future. We thereby followed up on Aristotle's epistemic condition for moral responsibility and aimed at raising awareness for the consequences of training on the well-being of a future generation. The Offspring treatment was similar to the Externality treatment, except that we added a monetary interdependence between the current pair (the "parents") and the future pair (the "offspring"). Current generation players received an additional payoff at a later date that depended on the future generation's payoff. This interdependence of payoffs is similar to what was done in previous intergenerational strategic games experiments (Schotter and Sopher, 2003, 2006, 2007) and can create solidarity across generations.

The aim of the second variation was to test whether, in the spirit of the thought experiment developed by John Rawls, letting participants decide in the so-called original position behind a "veil of ignorance" regarding the relative position they and others will take in the AI's decision would induce fairer allocation choices when they create training data. Previous studies (e.g. Huang et al., 2019) showed that putting decision makers into the original positions induces more utilitarian and socially beneficial choices. The Switch treatment mimicked the Baseline except that the payoffs of the dictator and the receiver, as determined by AI in period 31, could be switched with 50% probability. Finally, in the Offspring Switch variation, we combined the intergenerational dependence of payoffs and the uncertainty about the own position. With 50% probability, the additional, future payoffs of the current generation dictator and receiver were swapped.

Our results show that dictators did not consider the externality of generating training data for AI when they were simply informed about the existence of an externality for a future generation of players in the Externality treatment. The same behavior was observed in the Offspring treatment, that is, when dictators could benefit from their offspring's payoffs, provided relative positions across generations were stable and certain.

Electronic copy available at: https://ssrn.com/abstract=4002578

Behavior changed dramatically when relative positions for the algorithmic prediction became uncertain, to a degree that cannot be explained by expected payoff maximization. In Offspring Switch, we found evidence for intergenerational responsibility in terms of participants showing higher preferences for fairness when teaching the AI. They took (i) the efficient option more frequently even if this choice decreased their own immediate payoff, (ii) the selfish option less frequently even if their own payoff was lower than the one of the receiver in both options, and they chose (iii) the altruistic option more often if it reduced inequity and was fairer compared to the alternative. One possible interpretation is that future uncertainty forces individuals to take more distance with immediate selfish interests and help them take perspective by envisioning the situation more broadly from the start.

Our findings primarily complement the booming literature on AI, ethics, and algorithmic biases (see, for example, Anderson and Anderson (2007); Bostrom and Yudkowsky (2014); Bonnefon et al. (2016); Awad et al. (2018); Lambrecht and Tucker (2019); Rahwan et al. (2019); Awad et al. (2020)). This literature has started to explore the ethical principles that should guide AI in the presence of moral dilemmas. It has highlighted the cultural and individual diversity of moral preferences and the resulting complexity of combining these preferences to define acceptable guidelines for ethical AI. We contribute to this literature by considering a setting in which machines need to learn social preferences because the decisions to be made have an impact on several humans and do not depend on the answer to a true/false question. By focusing on general social preferences that play a role in many settings, we try to inform more broadly about situations in which an AI needs to learn such social preferences from repeated observation of human behavior. The system applies the learned preferences to generalize decision-making in new situations or in different types of interactions. We could think of cases where algorithms are used directly to allocate surplus or risk between parties (for example, algorithms that make recommendations to consumers or propose insurance plans), but there are many more cases where AI needs to incorporate social preferences as a parameter in its representation of the world (for example, to determine the acceptability of a policy in the domain of housing, environment, transportation, etc.). Hereby, our findings draw attention to the risks of teaching AI systems that are derived from legitimate concerns about algorithms imitating human behavior. Most humans are not ideal models of ethical agents, although they might have been taught egalitarian principles. In terms of policy implications, our experiment thus suggests a need for designing machines as agents with explicit ethical rules or guidelines, instead of simply aggregating the individual data of human decision makers. We suggest that this need is particularly vivid when machines operate in less mobile societies in which future status is certain and where social hierarchies are characterized by high inertia.

Further, our study adds to the literature that has used laboratory experiments to investigate human machine interactions. We refer to Chugunova and Sele (2020) for a comprehensive overview of this topic. The experimental research thus far has concentrated mainly

Electronic copy available at: https://ssrn.com/abstract=4002578

on settings where human subjects play against computers in strategic situations. Starting with Houser and Kurzban (2002), several papers have aimed at muting social preferences by replacing human players with computers in standard economic games, such as the public goods game.[4] In contrast to our experimental design, these studies either implemented predefined decision rules for the computer players or left participants uninformed of their impact on machine behavior. Closer to our approach of training that allows machines to learn from observed behavior are the studies in the context of auctions. Van den Bos et al. (2008) and Ivanov et al. (2010) have allowed computers to mimic the bidding strategy of individuals to study the winner's curse, and Teubner et al. (2015) used this method to investigate how playing against a computer affects individuals' bidding behavior and emotions. Hence, similar to the design of our study, the players' decisions determined machine behavior that ultimately affected their payoff. Finally, Corgnet et al. (2019) showed that social incentives are reduced when replacing human team members with robots. In general, one novelty of our paper is to introduce the AI not as an opponent in a strategic situation (in contrast also to Houy et al., 2020) but as a third-party observer that evaluates and learns from others' behavior and uses it for out-of-sample prediction and decision making.

Finally, operating through the data that underlies self-learning and intelligent algorithms, the paper broadly relates to studies on the intergenerational transmission of preferences and economic outcomes (Bisin and Verdier, 2001; Sacerdote, 2002; Björklund et al., 2006). As in Schotter and Sopher (2003, 2006, 2007) and Chaudhuri et al. (2006), we used a sequence of nonoverlapping generations of individuals playing a stage game for a finite number of periods. Additionally, we considered intergenerational income mobility as an additional factor influencing preferences and decision making (Bénabou and Ok, 2001; Alesina et al., 2018). By introducing uncertainty on future status in some treatments, our study also relates to the experimental literature on role uncertainty (*e.g.*, Engelmann and Strobel, 2004; Iriberri and Rey-Biel, 2011) and on the impact of playing both roles in games (*e.g.*, Gueth et al., 1982; Andreoni and Miller, 2002; Charness and Rabin, 2002; Burks et al., 2003). Although some studies showed that in such settings behavior is less selfish and more efficiency-oriented than when roles are certain and fixed (Iriberri and Rey-Biel, 2011), others have found that trust and reciprocity are reduced but only in the absence of uncertainty (Burks et al., 2003). Our results are in line with those in the literature and suggest that role uncertainty when generating training data for an AI can induce dictators to make less selfish decisions. When there was payoff interdependence with the future generation and dictators' future was more uncertain, they put higher weight on the receivers' payoffs.

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

[4]This subtraction method was further used by Ferraro et al. (2003) and Yamakawa et al. (2016) to disentangle motives for individuals' contributions to a public good, and by Benndorf et al. (2020) in the context of ultimatum bargaining games.

Electronic copy available at: https://ssrn.com/abstract=4002578

2 Design and Procedures

In this section, we first present the design of the experiment and our treatments.[5] Next, we describe our recruitment methods and experimental procedures.

2.1 Design

The experiment consisted of three parts. In part 1, participants performed real effort tasks. Part 2 comprised two stages. In the first stage, participants played several rounds of a mini-dictator game. The second stage comprised the prediction of a machine learning algorithm, based on the observed participants' behavior in the role of the dictator in stage 1, and its implementation. In the last part, we elicited sociodemographics and other information.

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

Part 2 The first stage of part 2 comprised 30 periods of a mini-dictator game.[6] Each participant was anonymously paired with another participant and matching was fixed for the whole part. One of the two pair members was randomly assigned the role of the dictator ("participant A" in the instructions); the other pair member played the role of the receiver ("participant B"). Roles remained fixed throughout the part. The dictator's task in each period was to select one of two options on how to split points between herself and the receiver she was matched to. All participants were informed upfront that these decisions would later serve as the only source of training data for a machine learning algorithm that would make a prediction and a decision in the second stage of this part and that this decision of the algorithm would affect their earnings. They were also informed that points earned in this part were convertible into Euros at the rate 100 points = 1 Euro.

In each period, the dictator could choose one of two possible payoff allocations, $X = (\Pi^1_X, \Pi^2_X)$ or $Y = (\Pi^1_Y, \Pi^2_Y)$, for which the sum of all four payoffs was kept constant in each period. The dictator's amount was always weakly higher with option X (the "selfish option") than with option Y. Because both participants had to complete the same set of tasks in the first part to generate the endowments, opting for the selfish option X when the receiver

[5] The general features of our game are similar as in our companion paper (Klockmann et al., 2021) but treatments differ.

[6] See Appendix B for an English translation of the instructions shown to the participants and of the control questionnaire.

Electronic copy available at: https://ssrn.com/abstract=4002578

would receive strictly higher payoff with option Y would indicate that the dictator takes advantage of her exogenously assigned power of decision.

Across the 30 binary dictator games, we systematically varied the payoffs in the two options to manipulate the inequality between pair members, the sum of payoffs in each option, and whether the dictator was in an advantageous or disadvantageous position relative to the receiver with both options. The calibration of payoffs was inspired by Bruhin et al. (2019). As in Bruhin et al. (2019), Figure 1 illustrates the payoff space and represents each game by a solid line that connects option X and option Y. Table C.1 in Appendix C lists all pairs of options.

We categorize the decisions along two dimensions. First, there were games in which option X Pareto-dominated option Y ($\Pi^1_X > \Pi^1_Y$ and $\Pi^2_X > \Pi^2_Y$), games in which the receiver or the dictator was monetarily indifferent between both choices ($\Pi^1_X = \Pi^1_Y$ or $\Pi^2_X = \Pi^2_Y$), and games in which the dictator receives strictly higher payoff with option X while the receiver receives strictly higher payoff with option Y ($\Pi^1_X > \Pi^1_Y$ and $\Pi^2_Y > \Pi^2_X$). Second, either the receiver or the dictator earns more money than the other player in both alternatives ($\Pi^1_X > \Pi^2_X$ and $\Pi^1_Y > \Pi^2_Y$, or $\Pi^2_X > \Pi^1_X$ and $\Pi^2_Y > \Pi^1_Y$), or the dictator's decision determined whose payoff was higher ($\Pi^1_X > \Pi^2_X$ and $\Pi^2_Y > \Pi^1_Y$, or $\Pi^2_X > \Pi^1_X$ and $\Pi^1_Y > \Pi^2_Y$). The aim of these variations was to identify the individual's distributional preferences. The order of the 30 decisions was random but was fixed for all the participants.

At the end of this stage, one of the 30 decisions was picked at random. This step determined the payoffs of both the dictator and receiver in this stage.

In stage 2, there was another 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). Notably, the exact functionality of the algorithm is not crucial for the research question and, as kept constant across conditions, does not affect the treatment differences. The focus is the training of AI as an a priori neutral technology with behavioral data. As explained to the participants before they made their first decision in stage 1, the algorithm used the dictator's 30 decisions as training data to make an out-of-sample prediction of how this dictator would have decided in period 31 when facing a new game.[7] The machine learning tool did not build models or estimate parameters; it simply predicted from patterns in the data. We used the payoffs and the sum and difference of points allocated to the players in the chosen and rejected options as features for classification. For the prediction of the AI, one of the six games represented by a dashed line in Figure 1 was chosen at random. Table C.2 in Appendix C lists all six

[7]Note that in many settings, algorithms are trained based on the behavior of many individuals. For our purpose of focusing on intergenerational responsibility, we let the algorithm learn from one individual only. We extensively study the question how behavior changes when a multiplicity of individuals provide training data as compared to a single one in our companion paper Klockmann et al. (2021).

Electronic copy available at: https://ssrn.com/abstract=4002578

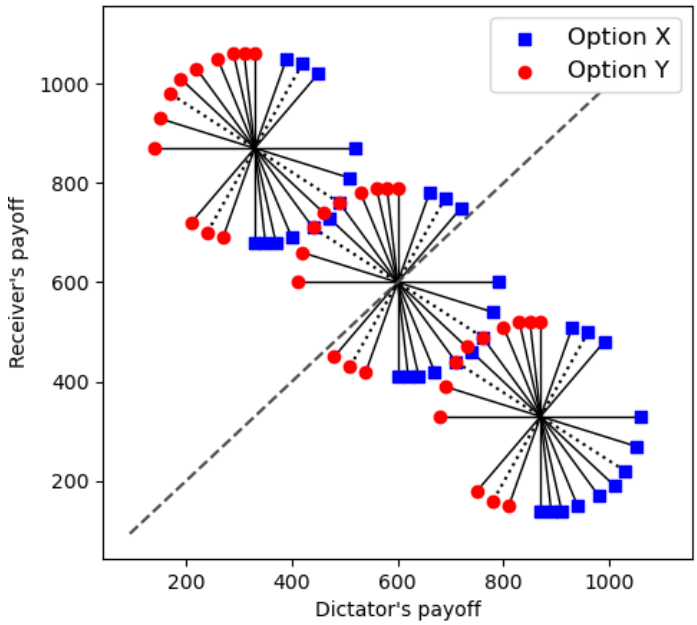

Figure 1: Dictator Games

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 in stage 2. 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.

possible out-of-sample decisions of the AI.

The prediction made by the AI in period 31 was implemented for the payment of the pair members. Therefore, this decision weighted as much as the 30 prior decisions in the determination of the participants' earnings in this part. This design choice was made to motivate the participants to pay attention to each of their decisions because not only each one could directly count for payment in stage 1 but was also used with certainty as training material for the AI, which affected payoffs in stage 2. If we had paid instead each of the 30 periods, the weight of the AI decision in period 31 would have been too small to identify the training effect, and a wealth effect through the length of training could have biased this identification.

Part 3 In the last part, we elicited sociodemographic information (gender, age, field and length of study, average weekly spending, and school graduation grade – Abitur). We further asked about familiarity with AI and machine learning, confidence in these technologies, satisfaction with the prediction made by the algorithm in period 31, and how accurate this prediction of the AI reflected the dictator's preferences. In addition, one question assessed the participants' understanding of the functionality of the random forest algorithm.

Electronic copy available at: https://ssrn.com/abstract=4002578

2.2 Treatments

As summarized in Table 1, the experiment includes a baseline condition and four between-subjects treatments.

Table 1: Overview of the Treatments

Treatments	Externality on Future	Payoff Dependency	Possible Role Switch
Baseline	No	No	No
Externality	Yes	No	No
Offspring	Yes	Yes	No
Offspring Switch	Yes	Yes	Yes
Switch	No	No	Yes

Our *Baseline condition* follows the aforementioned experimental design. The random forest algorithm made an allocation that was paid out exclusively to the current pair. There was no externality for any other player.

In the *Externality treatment*, participants were told that the dictator's choices generated training data not only for the AI's allocation decision in their group but also for another group that would participate in the same experiment with a similar basic design in the future, as described in sub-section 2.1.[8] Except for the mere information on this externality via the AI, there was no impact on payoffs in the current pair. The monetary incentives of the dictator therefore did not change between the Baseline and the Externality treatment.

The *Offspring treatment* added one payoff dimension to the Externality treatment. In this variation, the algorithm again learned from the dictator's choices to make a prediction not only for the dictator in the current pair but also for another dictator in a future pair. By contrast with the Externality treatment, the earnings of the first pair members were affected by the implementation of the prediction made by the AI for the future pair. The first generation received an additional payoff equal to half of the payoff allocated by the AI in the 31st period in the future generation. The dictator in the first generation received half of the payoff of the dictator in the future generation, and the receiver received half of the payoff of the receiver in the future generation. The objective of this monetary interdependence between the first pair (the "parents") and the future pair (the "offspring") was to make the externality through AI training more salient. Thereby, dictators were expected to internalize the monetary externality they exerted on the future through their generated training data. This payoff was added to the earnings made in one of the first 30 periods and in the 31st period, but was wired to participants at a later point in time. Compared to the Externality

[8]Precisely, participants were informed that in period 31 in their successor pair, the algorithm would decide based on the 30 decisions of the dictator in their pair and on the 30 decisions of the future dictator in their successor pair. The training data generated by their pair and by their successor pair were given the same weight in the prediction of the algorithm. The data of the participants in future pairs are not used in the current paper but in our companion paper (Klockmann et al., 2021).

Electronic copy available at: https://ssrn.com/abstract=4002578

treatment, the Offspring treatment preserves the impact on the future generation, but alters the monetary consequence of this impact for the parent generation.

The *Offspring Switch treatment* combined the externality of the decisions of the dictators from the current generation on the future generation's payoffs through the AI training with the risk of a switch in the payoff matching with the future generation. With 50% chance, payoff matching for the future, additional payoff was switched: If this occurred, dictators (receivers, respectively) in the current generation received half of the payoff of the receiver (dictator, respectively) from the AI's allocation in the future generation. Otherwise, payoff matching was not affected. The objective of this treatment was to increase dictators' awareness of the monetary consequences of their decisions for the receiver through AI training. Participants were not informed on whether the switch occurred or not during the session, but they could infer it when receiving their additional payment.

Finally, we used the *Switch treatment* as a control treatment aiming to isolate the pure effect of the possibility of swapped payoffs between the dictator and the receiver for the AI's allocation decision in period 31. This treatment mimicked the Baseline condition, except for the payoff stemming from the AI's prediction. With 50% probability, payoffs were switched between the dictator and the receiver: If this occurred, the dictator would receive the payoff assigned by the AI algorithm to the receiver in period 31, and the receiver would receive the payoff assigned to the dictator in period 31. Thus, after training the algorithm, the dictators had a self-interest to weigh more the receiver's payoff in their evaluation of the options. Nevertheless, in both treatments Switch and Offspring Switch with possible payoff swapping, the largest share of the dictator's total expected monetary outcome stems from the payoff initially allocated to the dictator (75% in Switch and 90% in Offspring Switch).

2.3 Procedures

Due to the 2020 coronavirus pandemic and the enforced social distancing rules to prevent its spread, we could not invite students to the laboratory to participate in an onsite experiment. Alternatively, we implemented a "virtual" laboratory experiment in the framework of the Frankfurt Laboratory for Experimental Economic Research (FLEX). Using ORSEE software (Greiner, 2015), we recruited students from Goethe University Frankfurt who provided their consent for participation in online experiments. We informed them about the new format and asked them to register for experimental sessions in the same manner used for a usual lab experiment.[9] Volunteers were informed that the experiment would be embedded in a virtual Zoom meeting, in which they would communicate with the experimenters, and that they would receive individual log-in information by email in advance.

To guarantee anonymity, we assigned each registered participant a unique ID in the

[9]Less than 1% of all students registered in the participants pool of the Frankfurt Laboratory for Experimental Economic Research opted out of online experiments. Therefore, there is no reason to fear a selection bias due to the novel format of the experiment.

Electronic copy available at: https://ssrn.com/abstract=4002578

form of a random alphanumeric string. The second purpose of these IDs was to ensure that only participants who registered for a given session participated in the meeting, and prevent individuals from participating multiple times. The third purpose was to identify participants from the Offspring and Offspring Switch treatments so that they could be assigned their payoff from the future generation. We programmed the experiment in oTree (Chen et al., 2016) and hosted it on a Heroku server.[10]

Once participants logged in, we assessed whether they had registered for the respective session. Participants were muted, and their video was disabled. After a short welcome from the experimenters, the oTree link to the experiment was posted to everyone on the chat function. The experiment started as soon as everyone was present in the oTree room. If there was an odd number of participants, one person was selected at random, who then received 4 Euros for arriving on time and who was dismissed. Participants could use the chat function to ask questions to the experimenters. To prevent the dictators from simply clicking through the 30 decisions to finish as soon as possible, we forced them to wait for 3 seconds after each choice of option before they could validate their decision.

Based on statistical power analyses with a significance level of 5%, an intended power of 80%, and a medium-size effect of about $7pp$, we determined a target sample size of 30 dictators per treatment. For the underlying distribution of decisions, we chose a binomial distribution where we varied the probability of selecting the selfish option X.[11] A total of 322 participants (161 dictators or independent observations, excluding the offspring subjects) were actually recruited between July and August 2020. Thirty-four dictators participated in the Baseline, 34 in the Externality treatment, 30 in the Offspring treatment, 31 in the Offspring Switch treatment, and 32 in the Switch treatment. There were no dropouts in any treatment, and we did not have to exclude any observation for technical or any other reasons. The participants' average age was 24.6 years, 52.5% were female, their predominant field of study was economics, and they were on average in their 7th to 8th semester. Table C.3 in Appendix C summarizes the main sociodemographic variables in each treatment. The only significant differences are that participants in the Externality treatment were on average 1.9 years older and there were fewer female participants than in the Baseline condition. If we only consider the dictators, there are no significant differences across treatments.

On average, participants earned 13.17 Euro (S.D. 3.21) and received their payoff by using PayPal. Participants in the Offspring and Offspring Switch treatments received a second payoff through PayPal once their successor pair participated in the experiment. Each session lasted approximately 45 minutes.

[10] This allowed us to benefit from a large flexibility in terms of monitoring the participants' progress, and it provided participants the possibility to login again and continue playing in the unlikely event of a disconnection.

[11] We preregistered the project, the sample size, and our hypotheses on AsPredicted (#44011) in July 2020 before starting the data collection. Before, we ran two pilot sessions in June 2020 to check for technical functionality and calibrated the decision space of the dictators.

Electronic copy available at: https://ssrn.com/abstract=4002578

3 Behavioral Predictions

To investigate how treatment manipulations influenced how individuals trained the AI algorithm, we developed behavioral conjectures based on the comparisons of the dictators' revealed distributional preferences across treatments. With standard preferences and individuals exclusively motivated by the maximization of their payoffs, we expected no difference across treatments Baseline, Externality, and Offspring. Since the expected payoff of dictators in the Switch and Offspring Switch partially depended on the monetary amount allocated to the receiver, we expected that a fully rational individual would choose option Y in some cases (see below for details).

To identify the distributional preferences revealed by the dictator's decisions, we considered the frequency of choosing a certain option and estimated parameters of social preferences. We refer to Bruhin et al. (2019), who built on Fehr and Schmidt (1999) and Charness and Rabin (2002), to set up a two-player model of social preferences that was fitted to the data, using maximum likelihood. Denoting the payoffs of the dictator by π_D and that of the receiver by π_R, the dictator's utility function is given by

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

Hereby, $s = \mathbb{1}\{\pi_R - \pi_D > 0\}$ and $r = \mathbb{1}\{\pi_D - \pi_R > 0\}$ are indicators of disadvantageous and advantageous inequality for the dictator, respectively. Following Bruhin et al. (2019), the sign of the parameters α and β describe the preference type of the dictator. $\alpha < 0$ reveals "behindness averse" decisions motivated by envy of the receiver's payoff whenever receiving a lower amount. Similarly, $\beta > 0$ reveals "aheadness averse" decisions motivated by a willingness to increase the other's payoff whenever receiving a larger amount.[12] Depending on the absolute value of α and β, the choices reveal more envious or more empathetic preferences. Furthermore, the parameters may have the same sign. First, if $\alpha, \beta < 0$, choices are spiteful, independent of whether the receiver earns more or less than the dictator. Second, if $\alpha = \beta = 0$, choices are purely selfish and the dictator does not put any weight on the other's payoff. Finally, if $\alpha, \beta > 0$, choices are altruistic because the dictator always derives utility from the receiver receiving a payoff.

When estimating the social preference parameters of a representative dictator, we closely adhered to above described method by Bruhin et al. (2019) for the following reasons. First, though dictators in the Offspring treatment can benefit repeatedly from selfish training of the algorithm, we need not incorporate this in the utility functions. The estimated parameters α and β only describe the relative weighting of payoffs rather than absolute weights. The

[12] For the sake of simplicity, we assume that the parameters of the model can capture the same preferences when the dictator is oneself and when the dictator is another individual (*i.e.*, when one's decisions train an algorithm that will decide for another couple of individuals). We simply allow the values of these parameters to vary cross conditions.

Electronic copy available at: https://ssrn.com/abstract=4002578

additional future payoff in the Offspring treatment proportionately increases stakes for both dictator and receiver and does not affect our parameter estimation.[13] Second, we do not explicitly model the externality of training data on the future that dictators create in the Externality, Offspring, and Offspring Switch treatments and thus do not directly incorporate the payoffs of the future pair(s) in the utility functions. If we just added these future payoffs of the other dictator and receiver, we would implicitly assume that the dictator assigns equal weights to her own payoff and the payoff of the other (future) dictator, which is most likely not true. Our aim is to measure how changes in the externality of AI training and its consequences affect revealed social preferences as defined in the dictator's utility function in (1). Finally, the utility function in (1) can already capture the possibility of switched payoffs between dictator and receiver in the Offspring Switch and Switch treatments. The dictator's expected payoff in these treatments is a weighted sum of π_D and π_R rather than only π_D. This can directly be represented by positive weights α and β beyond any social preferences. We refer to this in more detail in what follows.

We now formulate behavioral conjectures regarding the dictators' choices and the differences in the parameters capturing social preferences across treatments. In the first set of treatments (Baseline, Externality, and Offspring), the experimental design manipulated the relationship between dictators and another pair of participants in a future generation, and highlighted the transfer of data and choice preferences via the AI's training.

We first test whether an externality of today's behavior on future predictions and payoffs through the AI affected today's decisions, by comparing the Baseline and the Externality treatment. The salience of such an externality could influence its internalization by the dictators. For example, if an individual was inequity averse, she might be even more willing to reduce inequality in a future group that would be affected by her training of the algorithm. We thus conjectured that the information on an externality of AI training for the future, if anything, would make selfish choices that increase the dictator's payoff at the expense of the receiver's payoff less likely.

Conjecture 1. *Compared with the Baseline, dictators in the Externality treatment choose the selfish option weakly less frequently in decisions in which the receiver would be monetarily better off with the alternative. The estimated social preferences parameters, α and β, are weakly larger in the Externality treatment than in the Baseline.*

[13] In technical terms, 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 (1) 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)

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

The additional future payoff in the Offspring treatment would simply correspond to multiplying the objective function – *i.e.*, the deterministic utility u_D – with a positive factor. This only inversely proportionally changes σ, but leaves the estimated values of α and β unchanged.

Electronic copy available at: https://ssrn.com/abstract=4002578

We expected that introducing a monetary dependence between generations would induce stronger self-interest in the dictators in the Offspring treatment than in the Externality treatment. This is because the dictators could now benefit twice from the AI's predictions favoring them since a second payment was based on how many points the algorithm allocated to the dictator in the offspring pair. Therefore, dictators might be willing to opt more frequently for the selfish option in the Offspring than in the Externality treatment.

Conjecture 2. *Compared with the Externality treatment, dictators in the Offspring treatment choose the selfish option more frequently in decisions in which the receiver would be monetarily better off with the alternative. The estimated social preferences parameters, α and β, are smaller in the Offspring than in the Externality treatment.*

In the second set of treatments (Switch and Offspring Switch), we introduced the possibility of switching payoffs in period 31 or in the future generation between the dictator and the receiver on the training data created by the dictator. In the Switch treatment, the dictator in periods 1-30 may receive in period 31 the receiver's payoff based on the AI prediction. Her expected payoff then directly depends on the amount allocated to the receiver. Assuming that in each round the dictator takes only the current set of options into account, her expected payoff would amount to $\pi_D + \frac{1}{2}(\pi_D + \pi_R)$. An individual maximizing this expression would choose option X in 70% of all rounds (all except games 1,2,10,11,12,20,21,22,30 in Table C.1). Apart from that, her estimated social preference parameters would, by design, raise to $\alpha = \beta = 0.25$ when normalizing the expected payoff to get the relative weights as defined in (1). Thus, this treatment is expected to raise dictators' weighting of self-interest in the receiver's monetary outcome from the very beginning of the game.

Conjecture 3. *Compared with the Baseline, dictators in the Switch treatment choose the selfish option less frequently in decisions in which the receiver would be monetarily better off with the alternative. The estimated social preferences parameters, α and β, become closer to 0.5, as dictators weight their payoff and that of the receiver more equally in the Switch treatment than in the Baseline condition.*

Finally, the Offspring Switch treatment combines the externality of AI on a future generation with the possible switching of payoffs between the dictator and the receiver. Thus, this treatment can be directly compared with both the Switch and Offspring treatments. Without discounting the future, the dictator's expected payoff now amounts to $2\pi_D + \frac{1}{4}(\pi_D + \pi_R)$. The relative weight a purely selfish individual puts on the receiver's payoff is lower than in the Switch treatment and amount to $\alpha = \beta = 0.1$ when normalizing the expected payoff. This individual would choose option X in 80% of all rounds (all except games 1,10,11,20,21,30 in Table C.1). Still, because the Offspring Switch variation simultaneously activates both mechanisms, self-interest in the receiver's payoff and the externality through training, we expect dictators to act more egalitarian than in the other two treatments.

Electronic copy available at: https://ssrn.com/abstract=4002578

Conjecture 4. *Compared with the Switch and Offspring treatments, dictators in the Offspring Switch treatment choose the selfish option less frequently in decisions in which the receiver would be better off with the alternative. The estimated social preferences parameters, α and β, are larger in Offspring Switch compared with both other treatments.*

4 Results

To test our four conjectures, we primarily considered two measures that proxy moral behavior. One measure directly refers to the proportion of selfish choices of option X by the dictator in the different decision scenarios. The other measure is given by the estimated social preferences parameters, α and β, of a representative agent, following Bruhin et al. (2019).[14] Tables 2 and 3 report pairwise tests that compare the differences in the two aforementioned measures of preferences, as revealed by the actual choices, across treatments.[15] In what follows, we refer to the set of decisions characterized by conflicting interests, that is, the decisions in which the dictator gets a strictly higher payoff with option X and the receiver gets a strictly higher payoff with option Y, as the "restricted sample". For these decisions, choosing more frequently the selfish option can be considered a less moral action (because both players produced a similar effort in part 1). We report all games in appendix Table C.1 and mark those belonging to the restricted sample with an asterisk.

Surprisingly, Table 2 reveals no significant differences in the share of dictators' choices of the selfish option X between the Baseline, the Externality, and the Offspring treatments (t tests, $p > 0.39$ for all pairwise comparisons), in line with the prediction of a model with standard economic preferences. The share of selfish choices was relatively high in each of these treatments (70.29% in the Baseline, 71.08% in the Externality treatment, and 73% in the Offspring treatment). Similarly, in Table 3, the estimated social preferences parameters for the representative agent did not differ significantly across these three treatments ($p > 0.20$ in each pairwise comparison). Thus, being informed that training the AI algorithm would affect third parties' earnings in the Externality treatment did not affect the dictators' choices. Dictators did not care about the externality of their behavior in terms of income inequality in the whole experimental group. Comparing the Baseline and the Offspring treatments reveals that dictators neither altered their behavior when they could additionally benefit at a future point in time from teaching the AI with selfish decisions.

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

[15] Unless specified otherwise, all the non-parametric statistics reported in this paper were two-sided, and each individual provided one independent observation. We did not report regression analyses comparing treatments because there was no control group to compare all the treatments with. The only regression analysis that was conducted both at the aggregate level and for each treatment separately aimed at checking for the influence of individual sociodemographic characteristics on the probability to choose the selfish option. The results are reported in Table C.7 in Appendix C. At the aggregate level, only the variable "studying in economics" had a significant effect on this probability. Most variables have no effect in any regression.

Electronic copy available at: https://ssrn.com/abstract=4002578

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

Treatments	Nb Obs.	Option X	*p–values*	Option X Restricted Sample	*p–values*
Baseline	34	70.29% (0.029)		66.01% (0.046)	
			0.846		*0.713*
Externality	34	71.08% (0.028)		68.30% (0.042)	
			0.592		*0.391*
Offspring	30	73.00% (0.021)		73.00% (0.033)	
			0.000		*0.000*
Offspring Switch	31	58.71% (0.018)		48.39% (0.030)	
			0.081		*0.039*
Switch	32	63.85% (0.023)		58.16% (0.035)	
			0.087		*0.182*
Baseline	34	70.29% (0.029)		66.01% (0.046)	

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 gives one independent observation. Column "Option X [Restricted sample]" includes only the decisions in games characterized by conflicting interests, that is, those 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. The Baseline appears twice to report comparisons with both Externality and Switch treatments.

Table 3: Estimated Parameters of Social Preferences across Treatments

Treatments	Nb Obs.	Dictators	α	*p–values*	β	*p–values*
Baseline	1020	34	0.082^{*} (0.048)		0.395^{***} (0.051)	
				0.863		*0.441*
Externality	1020	34	0.070 (0.054)		0.341^{***} (0.047)	
				0.200		*0.974*
Offspring	900	30	-0.041 (0.068)		0.343^{***} (0.051)	
				0.000		*0.007*
Offspring Switch	930	31	0.246^{***} (0.034)		0.506^{***} (0.033)	
				0.030		*0.801*
Switch	960	32	0.123^{***} (0.045)		0.492^{***} (0.046)	
				0.536		*0.158*
Baseline	1020	34	0.082^{*} (0.048)		0.395^{***} (0.051)	

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 for a representative agent in each treatment. p-values refer to z-tests for differences in estimates. The Baseline appears twice to report comparisons with both Externality and Switch treatments. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

This analysis rejects both Conjectures 1 and 2, and leads to Result 1.

Result 1. *Being informed of the externality of training data for an AI did not affect the selfishness of decisions when the future status was certain (i.e., no offspring existing or the offspring holding the same relative position). Neither the share of the selfish option nor the estimated social preference parameters differed significantly between the Baseline, Externality, and Offspring treatments.*

In line with Conjecture 3 but without reaching a standard level of significance (t test,

Electronic copy available at: https://ssrn.com/abstract=4002578

$p = 0.087$), we found a lower fraction of selfish choices in the Switch treatment than in the Baseline condition, considering the full sample (63.85% *vs.* 70.29%). The percentage was close to the share of 70% option X an expected payoff maximizer would choose as noted in section 3. The effect mainly stemmed from the decisions in which choosing option Y would reduce the absolute distance in payoffs between the dictator and the receiver (see Table C.6 in Appendix C).[16] In this subsample, dictators chose the selfish option – which in this case also raised inequality – less often when there was the uncertainty of switching payoffs (48.13% *vs.* 57.65% in the corresponding Baseline condition), but this remained only marginally significant (t test, $p = 0.086$). Participants tended to hedge the risk of a future lower payoff by opting more frequently for the alternative that offered a lower inequality in the training data. The estimates indicate that the representative dictator in the Switch treatment values the payoff of the receiver positively ($\alpha > 0$ and $\beta > 0$ with $p < 0.01$) regardless of whether the other player is monetarily better or worse off. This raise in the relative weighting of the receiver's payoff goes in line with what changed monetary incentives would predict, though less pronounced. Finally, in the cases in which they were in an advantageous position, dictators weighted their and the receiver's payoffs almost equally ($\beta = 0.492$, see Table 3).

This analysis (mildly) supports Conjecture 3 and leads to Result 2.

Result 2. *When dictators were aware that they could receive the payoff of the receiver when the AI's prediction determined the allocation of payoffs (Switch treatment), there is some (weak) evidence that they tended to reduce payoffs' inequality when training the AI, compared with the Baseline condition. This goes in line with changed monetary incentives, though less pronounced.*

The most striking and stronger effects emerged when comparing the Offspring Switch treatment with the Offspring and with the Switch treatments. Table 2 shows that when dictators could earn a fraction of the receiver's payoff in the future generation, they chose highly significantly less often the selfish option compared with the participants who received the dictator's payoff with certainty at both points in time (58.71% in the full sample and 48.39% in the restricted sample; t tests, $p < 0.001$ in both samples). In line with this effect, the estimated parameters α and β for the representative agent were both highly significantly higher in the Offspring Switch treatment than in the Offspring treatment (α: 0.246 *vs.* -0.041; z-test, $p = 0.0002$. β: 0.506 *vs.* 0.343; z-test, $p = 0.007$). As in the Switch treatment, the representative dictator in the Offspring Switch treatment values the payoff of the receiver positively ($\alpha > 0$ and $\beta > 0$ with $p < 0.01$), independent of advantageous or disadvantageous inequity. Note that the increase in the weight the dictator puts on the receiver's payoff is about twice as large in the case of disadvantageous inequity compared to advantageous inequity ($\Delta\alpha = 0.287$ vs. $\Delta\beta = 0.163$). The strongest behavioral change

[16] This corresponds to games 1 to 7, 10, 28 and 29 in Table C.1 in Appendix C.

Electronic copy available at: https://ssrn.com/abstract=4002578

hence occurs in those decisions where the dictator is monetarily worse off than the receiver and it is highly prosocial to positively weight the payoff of the other in one's decision. Still, both parameters rise by more than 0.1, *i.e.*, by more than what the change in monetary incentives would predict (see section 3). Thus, when holding the intergenerational impact of AI training constant and introducing the possibility of swapped payoffs, dictators selected overwhelmingly more frequently the option favoring the receiver. Our findings suggest that individuals only take the externality through AI training into account when we exogenously put them in the original position and let them decide behind a "veil of ignorance" regarding their future position, in the spirit of John Rawls.

The observed treatment difference replicated independent of the advantageousness or disadvantageousness of the dictator's position (Table C.4 in Appendix C), and of which alternative exhibited lower inequality (see Table C.6). The treatment effect vanished only when considering those decisions in which option X increased the sum of payoffs, namely, when they were more efficient.[17] If both motives, increasing the immediate payoff and raising efficiency, favored option X, dictators chose this alternative in over 90% of cases in both treatments (Table C.5).

This analysis supports Conjecture 4 for the comparison between the Offspring Switch and Offspring treatments, and is summarized in Result 3.

Result 3. *When future status became uncertain and dictators could be harmed by the externality of their training data, intergenerational responsibility arose and the selfishness of decisions decreased. The percentage of choices of option X was significantly lower and the estimated social preference parameters significantly increased in the Offspring Switch treatment compared with the Offspring treatment. Changes in monetary incentives alone cannot explain the change in revealed social preferences.*

Finally, we compared the Offspring Switch and the Switch treatments. This comparison maintains the possibility of payoff switching but varies the existence of an externality of the AI training data on a future generation. If payoff switching occurs, the consequences for the dictator are relatively less important in the Offspring Switch treatment than in the Switch treatment because it affects only the additional payment from the future generation (equal to 50% of the future generation's payoff). This comparison showed similar but weaker effects as in Result 3. Participants in the former treatment chose the selfish option X less often (full sample: 58.71% *vs.* 63.85%; t test, $p = 0.081$; restricted sample: 48.39% *vs.* 58.16%; $p = 0.039$). Recall from section 3 that a purely selfish individual would choose option X more often in Offspring Switch compared to Switch as the receiver's payoff made up a smaller share of the dictator's total expected outcome, supporting a greater role of social preferences. In both treatments with uncertainty regarding future outcomes, dictators assign positive weight to the receiver's payoff in all situations. The behindness-

[17] This corresponds to games 6 to 9, 16 to 19, and 26 to 29 in Table C.1 in Appendix C.

Electronic copy available at: https://ssrn.com/abstract=4002578

aversion parameter, α, was significantly higher in the Offspring Switch than in the Switch treatment (0.246 *vs.* 0.123; z-test, $p = 0.030$). The aheadness-aversion parameter, β, did not differ between these two treatments and remained at approximately 0.5 (precisely, 0.506 *vs.* 0.492; $p = 0.80$). The changes in behavior can be observed only in case of disadvantageous inequity, when assigning positive weight to the receiver's payoff is the strongest sign of prosocial preferences. From the perspective of an expected payoff maximizer, however, both parameters would, by design, go down by 0.15 in Offspring Switch compared to Switch (see section 3), which is not what we observe. We detect an increase in prosociality despite the relatively small size of the potential monetary loss when roles are switched. Returning to John Rawls, one could interpret that only an effect on an independent future pair of players makes participants adopt the original position behind a veil of ignorance. Then, the intergenerational transmission of training data becomes particularly salient. In fact, Rawls' thought experiment originally intended to improve decisions on a societal level also affecting others rather than just oneself.

In line with this finding, the likelihood of selecting the selfish option differed significantly when considering the subset of decisions in which the dictator was always in a disadvantaged position compared with the receiver (67.4% *vs.* 75%; $p = 0.045$),[18] but not in the cases in which the dictator was in an advantageous position (46.77% *vs.* 50%; $p = 0.416$) or in a mixed position (61.94% *vs.* 66.56%; $p = 0.227$; Table C.4 in Appendix C). In the scenarios in which the dictator was in a disadvantaged position, the intergenerational impact of training made individuals put relatively more weight on the payoff of others. Finally, introducing the externality triggered efficiency concerns. Indeed, in the Offspring Switch treatment, only approximately one quarter of decisions were selfish when option X was inefficient (Table C.5). This share was significantly lower than that in the Switch treatment (26.88% *vs.* 38.33%; t test, $p = 0.033$). No differences were observed when the selfish choices were the efficient choices (95.70% *vs.* 92.45%; t test, $p = 0.287$).

Result 4 provides additional support for Conjecture 4.

Result 4. *When the future status was uncertain and for a given possibility of switched payoffs, introducing an externality of AI training on the future significantly reduced the frequency of selfish choices, especially when efficiency could be improved by an altruistic choice. Individuals in the Offspring Switch treatment chose option X significantly less often than those in the Switch treatment did, and the parameter α of aversion against disadvantageous inequality was significantly higher in that treatment, though monetary self-interest in the receiver's payoff decreases.*

Overall, these results show that the externality of AI training data on a future generation (and an individual's future payoffs) together with individuals' uncertainty about their future status when being again affected by the implementation of the AI's predic-

[18] This corresponds to games 21 to 30 in Table C.1 in Appendix C.

Electronic copy available at: https://ssrn.com/abstract=4002578

tions strongly influenced individuals' decisions and the social preferences revealed by their choices. Teaching an AI algorithm seemed to trigger intergenerational responsibility if and only if the individuals could be harmed once the AI training leads to less moral decisions in a future generation.

To test the robustness of these findings, we additionally adjusted the statistical results for multiple hypotheses testing within each column of Tables 2 and 3. We used the family-wise error rate (FWER) by employing the Holm-Bonferroni and Holm-Šidák methods (Holm, 1979). We used the false discovery rate (FDR) by employing the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Comparing the Offspring Switch and Offspring treatments, we found that the observed differences in all four measures (*i.e.*, the share of option X in all decisions and in the restricted sample, α and β) remained significant at a 5% level. Comparing the Offspring Switch and Switch treatments, we observed that the effect of introducing an externality under a given mobility regime survived the FDR correction only at the 10% level when considering the share of option X in the restricted sample and the envy parameter α. The other comparisons between these two treatments lost significance.

Finally, using a non pre-registered, exploratory, analysis, we examined how dictators formed their decisions beyond the analysis of their actual choices of the option. In particular, as an assessment of how much attention the individuals paid to the menu of options, we measured the time in seconds dictators stayed on the page before submitting their choice between option X and option Y. Table 4 reports how long dictators on average weighed alternatives before submitting their decisions in each treatment.

Table 4: Average Decision Time in Seconds, by Treatment

Treatments	Nb of Obs.	Mean Time	$p-values$	Mean Time Restricted Sample	$p-values$
Baseline	34	11.44 (0.573)	*0.552*	12.64 (0.781)	*0.780*
Externality	34	11.89 (0.505)	*0.001*	12.91 (0.606)	*0.002*
Offspring	30	9.86 (0.299)	*0.001*	10.52 (0.344)	*0.001*
Offspring Switch	31	13.88 (1.102)	*0.243*	15.89 (1.477)	*0.241*
Switch	32	12.28 (0.795)	*0.385*	13.71 (1.121)	*0.432*
Baseline	34	11.44 (0.573)		12.64 (0.781)	

Notes: The table reports the dictators' average decision time, in seconds. These values include the three seconds they had to wait between selecting an option and submitting it. Standard errors of means are in parentheses. One observation corresponds to one dictator. The column "Mean Time [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.

The comparison between the Baseline and the Externality treatments indicated that the awareness of their intergenerational impact through AI training did not significantly affect

Electronic copy available at: https://ssrn.com/abstract=4002578

the dictators' decision time. By contrast, the comparison between the Offspring and the Externality treatments revealed a highly significant difference. Although the decisions did not differ between these treatments (see Result 1), dictators needed on average less time for making their decision when these choices involved a higher personal monetary stake, namely, when they could benefit another time in the future from how they trained the AI.

Although the risk of swapped payoffs in period 31 in the Switch treatment did not significantly affect decision time compared with the Baseline, such a risk in the future in the presence of a monetary dependence between generations raised the average decision time in the Offspring Switch treatment in comparison with the Offspring treatment. Individuals used approximately 41% more time to select one option if the externality their AI training data generated could harm them in the future. This difference is highly significant ($p = 0.001$). This difference in timing reflects the difference in decisions described in Result 3.[19]

This exploratory analysis supports our last result.

Result 5. *Dictators made on average faster decisions in the Offspring treatment than their counterparts did in the Offspring Switch and Externality treatments.*

5 Discussion and Conclusion

Technological progress, strongly accelerated by an ever-increasing amount of available data and growing capacities of automated agents, comes along with ethical questions. Machines often have to include in their decision models parameters that represent the social preferences of individuals in the society. Therefore, they have to learn what these social preferences are, based on the repeated observation of human decision making in social environments, that is, when decisions impact others' utility. AI is thus far not an abstract, self-deciding super-intelligence, but primarily consists of prediction machines that make decisions based on training data (Agrawal et al., 2018). These training data do not come from impartial spectators but from the decision makers themselves. If these training data reflect unethical behavior (*e.g.*, discrimination, egoism, imposition of the law of the strongest), this influences AI's decision models across scenarios and generations. With the growing use of "black box" algorithms, making humans accountable for AI predictions is becoming increasingly difficult (Pasquale, 2015). An obvious approach to managing this challenge could be to ensure algorithmic transparency. However, making individuals fully aware of their responsibility in training AI, not only in terms of immediate consequences but also regarding the impact on the future, may not be sufficient to make them more attentive to the externalities they generate through their actions. Our study suggests that this may differ depending on whether

[19] When splitting the sample by decision (X or Y), the same treatment differences in terms of decision time persist, independent of the decision. Dictators were always faster in the Offspring than in the Externality and Offspring Switch treatments. This also holds when splitting by decision type, namely, restricted sample, advantageousness, fairness, and efficiency.

Electronic copy available at: https://ssrn.com/abstract=4002578

individuals are living in societies with high or low intergenerational income mobility.

Studying whether individuals change behavior if they anticipate that their actions generate externalities through the training of an AI would be extremely difficult with observational data. This is why we designed a controlled, simplified environment in which individuals received transparent information that their decisions were used to train an AI and would have externalities. The algorithm predicted the choice the individual would make in a future game. We manipulated the presence of an externality of the AI training data, the concern of the participants for the consequences of these training data on future participants affected by the AI, and uncertainty about future status.

Our findings reveal that intergenerational responsibility in human behavior when training an AI cannot be triggered by information alone. It was only in settings where individuals understood that there was a risk of being harmed by immoral algorithmic decisions (through harming their offspring) that they considered intergenerational transmission of data when teaching the machines and exhibited less selfishness. Uncertainty regarding their own future position and their offspring's payoff allowed individuals to be more considerate and put higher weight on the others' outcome when providing training data for a machine learning algorithm. Though we implemented interdependence of earnings across generations, this effect cannot be explained by monetary self-interest alone. Under standard economic preferences, an expected payoff maximizing individual would have neglected the uncertainty in future earnings because the affected share of total payoff is small. Indeed, in the Switch treatment, 25% of the total payoff corresponds to the receiver's payoff in expectation; in the Offspring Switch treatment, it is only 10%. As we observe higher weights on the receiver's payoff in Offspring Switch than in Switch, self-interest cannot explain more egalitarian choices when training the AI also for future generations. By contrast, in stable settings with no intergenerational earnings mobility, individuals disregarded the negative externalities of their decisions on the future generation. One possible interpretation is that being more uncertain about one's future situation leads individuals to take more distance from their immediate selfish interests and leads them to envision the situation more broadly from the beginning, in the spirit of John Rawls' idea of taking decisions behind a veil of ignorance. Another possible interpretation building on Prospect Theory (Kahneman and Tversky, 1979) is that pessimistic individuals who are more anxious about their future status and payoffs might overweight the subjective probability of switching payoffs in the last period and overweight the possibly occurring losses. Since in the current design the dictator is not always in advantageous position (and thus in period 31, switching position is not necessarily bad in terms of absolute payoff), we tend to favor the first interpretation. However, to neatly disentangle between these different interpretations it might be interesting in future work to elicit subjective beliefs about the risk of getting a lower payoff in the last period.

Our study has limitations. The main limitation is the caution that must be used regarding the external validity of our results because they have been generated in artificial

Electronic copy available at: https://ssrn.com/abstract=4002578

settings. In real settings, AI is obviously not trained with the data of a single or just a few individuals. But the artificiality of the experiment allows us to observe what is precisely impossible to observe in reality: How do individuals alter their decisions when they are fully aware that their action has an impact on future generations through the training of an AI? If we do not find any impact in this specific context where individuals are the only source of the externality through training, it is very unlikely that we would find an impact when individuals are not pivotal. While keeping this caveat in mind, our experimental results suggest a need for designing machines more as explicit ethical agents since we show that even when being aware and pivotal, most individuals do not adjust to the simple presence of externalities. Because human beings are not ideal models of ethical agents due to widespread selfishness, teaching an AI by simply imitating human behavior risks reproducing unfairness. An implication might be that machine learning algorithms should be extended with classical programming that exogenously implements basic moral guidelines. Which moral guidelines should be implemented must be subject to evaluations of both universal and more culture-specific moral preferences (Awad et al., 2020).

Our paper represents a starting point by focusing on simple dictator games in which one single individual is fully responsible for the outcome in the group and for the externality she creates by passing on training data to future players. Further research could address our questions by using more complex, strategic games, such as bargaining games. We could also use the training data to allow the AI to learn social preferences and then, let the AI use these preferences in a different decision environment. Another avenue would be to attempt to replicate our results by comparing how managing the ethics of new technologies differs across cultures and countries with a low *versus* high degree of intergenerational income mobility (OECD, 2008) and by considering real-world applications of AI systems that make decisions based on observed behavior. Promising avenues of research would also be to study the effects of an individual's actual *versus* perceived accountability for algorithmic biases, or of varying the pivotality for choices of algorithms with the availability of Big Data.

Electronic copy available at: https://ssrn.com/abstract=4002578

References

Adewumi, A. O. and A. A. Akinyelu (2017): "A survey of machine-learning and nature-inspired based credit card fraud detection techniques," *International Journal of System Assurance Engineering and Management*, 8, 937–953.

Agrawal, A., J. Gans, and A. Goldfarb (2018): *Prediction machines: The simple economics of artificial intelligence*, Harvard Business Press.

Alesina, A., S. Stantcheva, and E. Teso (2018): "Intergenerational mobility and preferences for redistribution," *American Economic Review*, 108, 521–54.

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

Andreoni, J. and J. Miller (2002): "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica*, 70, 737–753.

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.

Bénabou, R. and E. A. Ok (2001): "Social mobility and the demand for redistribution: The POUM hypothesis," *Quarterly Journal of Economics*, 116, 447–487.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289–300.

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

Bhattacharyya, S., S. Jha, K. Tharakunnel, and J. C. Westland (2011): "Data mining for credit card fraud: A comparative study," *Decision Support Systems*, 50, 602–613.

Bisin, A. and T. Verdier (2001): "The economics of cultural transmission and the dynamics of preferences," *Journal of Economic Theory*, 97, 298–319.

Björklund, A., M. Lindahl, and E. Plug (2006): "The origins of intergenerational associations: Lessons from Swedish adoption data," *Quarterly Journal of Economics*, 121, 999–1028.

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.

Electronic copy available at: https://ssrn.com/abstract=4002578

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

Brennan, T., W. Dieterich, and B. Ehret (2009): "Evaluating the predictive validity of the COMPAS risk and needs assessment system," *Criminal Justice and Behavior*, 36, 21–40.

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.

Burks, S., J. Carpenter, and E. Verhoogen (2003): "Playing both roles in the trust game," *Journal of Economic Behavior & Organization*, 51, 195–216.

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

Chaudhuri, A., S. Graziano, and P. Maitra (2006): "Social learning and norms in a public goods experiment with intergenerational advice," *Review of Economic Studies*, 73, 357–380.

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

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

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

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.

Cowgill, B. (2018): "The impact of algorithms on judicial discretion: Evidence from regression discontinuities," Working paper.

Cowgill, B. and C. E. Tucker (Forthcoming): "Algorithmic fairness and economics," *Journal of Economic Perspectives*.

Dickerson, J. and T. Sandholm (2015): "FutureMatch: Combining human value judgments and machine learning to match in dynamic environments," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 622–628.

Engelmann, D. and M. Strobel (2004): "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," *American Economic Review*, 94, 857–869.

Ensign, D., S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian (2018): "Runaway feedback loops in predictive policing," in *Conference on Fairness, Accountability and Transparency*, PMLR, 160–171.

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

Electronic copy available at: https://ssrn.com/abstract=4002578

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.

Flores, A. W., K. Bechtel, and C. T. Lowenkamp (2016): "False positives, false negatives, and false analyses: A rejoinder to 'Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks'," *Fed. Probation*, 80, 38–46.

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

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

Holm, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 65–70.

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.

Huang, K., J. D. Greene, and M. Bazerman (2019): "Veil-of-ignorance reasoning favors the greater good," *Proceedings of the national academy of sciences*, 116, 23989–23995.

Iriberri, N. and P. Rey-Biel (2011): "The role of role uncertainty in modified dictator games," *Experimental Economics*, 14, 160–180.

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

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–292.

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

Klockmann, V., A. von Schenk, and M. C. Villeval (2021): "Artificial Intelligence, Ethics and Pivotality in Individual 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.

OECD (2008): *Growing Unequal? Income Distribution and Poverty in OECD Countries,*, Paris: OECD Publishing.

Pasquale, F. (2015): *The black box society*, Harvard University Press.

Electronic copy available at: https://ssrn.com/abstract=4002578

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.

Sacerdote, B. (2002): "The nature and nurture of economic outcomes," *American Economic Review*, 92, 344–348.

Schotter, A. and B. Sopher (2003): "Social learning and coordination conventions in intergenerational games: An experimental study," *Journal of Political Economy*, 111, 498–529.

——— (2006): "Trust and trustworthiness in games: An experimental study of intergenerational advice," *Experimental Economics*, 9, 123–145.

——— (2007): "Advice and behavior in intergenerational ultimatum games: An experimental approach," *Games and Economic Behavior*, 58, 365–393.

Shariff, A., J.-F. Bonnefon, and I. Rahwan (2017): "Psychological roadblocks to the adoption of self-driving vehicles," *Nature Human Behaviour*, 1, 694–696.

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

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

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

Electronic copy available at: https://ssrn.com/abstract=4002578

A Appendix Random Forest Algorithm

For predicting the out-of-sample choice of the dictator in the 31st period, 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 161 dictators, this yielded an average predictive accuracy of about 81.8%, a precision of 86.6%, a recall of 86.6%, and a F_1 score of 0.866. 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 decisions for training to make the out-of-sample prediction. In Table A.1, we report the precision, recall, and F1 score of the random forest algorithm in each treatment. We conclude that the performance of the algorithm was not different across treatments.

Table A.1: Performance Metrics of the Random Forest Algorithm by Treatment

Treatment	Dictators	Accuracy	Precision	Recall	F_1 score
Baseline	34	84.31%	89.93%	87.41%	0.8865
Externality	34	80.39%	85.16%	88.59%	0.8684
Offspring	30	83.89%	87.50%	90.84%	0.8914
Offspring Switch	31	77.96%	81.90%	82.61%	0.8225
Switch	32	80.21%	81.34%	89.34%	0.8516

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 questionnaire in the final stage included questions about the participants' attitudes toward AI in general and toward the machine learning algorithm in our experiment. On a scale from 1 to 5, we asked dictators to rate their familiarity with and confidence in this technology (averages of 2.8 and 3.7, respectively), their satisfaction with the prediction in period 31 (average of 4.0),

Electronic copy available at: https://ssrn.com/abstract=4002578

and their assessment of how accurately the AI's decision matched their true preferences (average of 4.3). There were no significant differences across treatments in any of these variables. There is also no correlation of these variables with observed behavior and treatment differences. Satisfaction with and assessed accuracy of the algorithm were not only very high, but also strongly correlated (Spearman rank correlation: 0.717, $p < 0.001$).

Electronic copy available at: https://ssrn.com/abstract=4002578

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.

Overview

Today's experiment consists of two parts.
In the first part you earn points by solving tasks.
You will receive more detailed information on the second part at the end of the first part.

Instructions of Part 1

In the first part you will earn points by performing 5 tasks.
For each task you will see a different block of numbers.
In each block, you must select a specific combination of numbers.
By completing all 5 tasks successfully you will earn 1200 points that will be used in the second part of the experiment.

Figure B.1: Instructions – Real Effort Tasks

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

End of Tasks

You have successfully completed all tasks and earned 1200 points that you will be able to use in the second part of the experiment.

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

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

Electronic copy available at: https://ssrn.com/abstract=4002578

Instructions of Part 2

Instructions

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

Overview

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

Decisions in Periods 1 to 30

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

Figure B.3: Instructions – Main Part and Decisions

Note: This screen was displayed in all treatment pairs.

Instructions of Part 2

Period 31

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

Your payoff in period 31 is determined as follows.
The previous 30 decisions of participant A are used to train an artificially intelligent Random Forest algorithm (see Info Box).
It is a machine learning algorithm that observes and learns from participant A's behavior.
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 <u>based on its prediction what participant A would prefer</u>, given this participant A's choices of options in the first 30 periods.
This option chosen in this prediction determines your payoff in period 31 and the payoff of the other participant in your pair.

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

Electronic copy available at: https://ssrn.com/abstract=4002578

[Baseline, Externality, Offspring, Offspring Switch]
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.

[Switch]
Note that with 50% probability, the roles in your pair are reversed for this payoff in round 31.
If you were participant A in periods 1 to 30, you will become participant B in period 31 with 50% probability and you will remain participant A with the remaining 50%.
Similarly, if you were participant B in periods 1 to 30, you will become participant A in period 31 with 50% probability and you will remain participant B with the remaining 50%.

Thus, with 50% probability each, you get the payoff of participant A or of participant B in your pair, independent of your role in the previous 30 periods.

Example: If the algorithm predicts that participant A would prefer option X with payoffs (K, V): With 50% probability, A in the pair receives K points and B receives V points; with 50% probability, A in the pair receives V points and B receives K points.

Info box: Random Forest Algorithm

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

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

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

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

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

Electronic copy available at: https://ssrn.com/abstract=4002578

Instructions of Part 2

The 30 decisions of participant A in your pair generate additional training data for further predictions.
These affect the payoffs of a successor pair in period 31 that will take your place in this same experiment in the next days.

The artificially intelligent algorithm is trained with the decisions of participant A in your current pair.
The algorithm will make two predictions.

1. As mentioned above, there will be a prediction for a decision of participant A in your pair today.
   In period 31, the algorithm decides once between option X and option Y based on participant A's previous decisions in periods 1 to 30.

2. There will be a prediction that determines the decision of participant A in your successor pair.
   The algorithm will decide based on the 30 decisions of participant A in your pair today and on the 30 decisions of the future participant A in your successor pair.
   The training data generated by your pair and by your successor pair are given the same weight in the prediction of the algorithm.

Summarizing, based on the 30 decisions of participant A in your pair today, the algorithm decides once between option X and Y for your pair today and once for your successor pair in the future.
Thus, the 30 decisions of participant A today have monetary consequences for your pair today and for another future pair, through the predictions of the algorithm.

Figure B.5: Instructions – Impact on Future Session (Externality Treatment)

Note: This screen was displayed in the Externality treatment pairs.

Electronic copy available at: https://ssrn.com/abstract=4002578

Instructions of Part 2

For both participants A and B, there is an additional payoff.
This amount will be paid out at a future point in time.
We will give details concerning the payment at the end of the session.

The 30 decisions of participant A in your pair generate additional training data for further predictions.
On the one hand, these affect the payoffs in period 31 of a successor pair that will take your place in this same experiment in the next days.
On the other hand, they also affect your own additional payoff.

The artificially intelligent algorithm is trained with the decisions of participant A in your current pair.
The algorithm will make two predictions.

1. As mentioned above, there will be a prediction for a decision of participant A in your pair today.
   In period 31, the algorithm decides once between option X and option Y based on participant A's previous decisions in periods 1 to 30.

2. There will be a prediction that determines the decision of participant A in your successor pair.
   This second prediction may be for another pair of options.
   The algorithm will decide based on the 30 decisions of participant A in your pair today and on the 30 decisions of the future participant A in your successor pair.
   The training data generated by your pair and by your successor pair are given the same weight.
   In particular, participant A accounts for one half of the data the algorithm deciding for your successor pair builds on.

Summarizing, based on the 30 decisions of participant A in your pair today, the algorithm decides once between option X and Y for your pair today and once for your successor pair in the future.
Thus, the 30 decisions of participant A today have monetary consequences for your pair today and for another future pair, through the predictions of the algorithm.
On top of today's payoff, we pay your pair an amount equal to one half of the payoff your successor pair gets from the prediction of the algorithm in the future.
This prediction forms the third part of your payoff.

[Offspring Switch]
Note that with 50% probability, the roles in your pair are reversed for this future payoff.
With 50% probability each, you receive 50% of the payoff of participant A or of participant B in your successor pair, independent of your role today.

[Offspring]
Note that your role as participant A or participant B remains the same for this future payoff.
If you are participant A today, your additional payoff will be determined as 50% of the payoff of participant A in your successor pair.
If you are participant B today, your additional payoff will be determined as 50% of the payoff of participant B in your successor pair.

Figure B.6: Instructions – Impact on Future Session (Offspring and Offspring Switch Treatments, cont'd)

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

Electronic copy available at: https://ssrn.com/abstract=4002578

Control Questions

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

Question 1

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

- ◉ Yes
- ○ No

Question 2

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

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

Question 3

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

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

Question 4

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

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

Figure B.7: Control Questions

Electronic copy available at: https://ssrn.com/abstract=4002578

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.
- ○ Exclusively from the 30 decisions of participant A in another pair in your session.
- ○ From the 30 decisions of participant A in your pair and the 30 decisions of participant A in another pair.
- ○ From the 30 decisions of participant A in your pair and the 30 decisions of participant A in 99 other pairs.

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. [Baseline, Switch]
- ○ Exclusively for another pair in your session.
- ○ For your pair and for another successor pair of yours in the future. [Externality, Offspring, Offspring Switch]

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.
- ○ The decision of the artificially intelligent algorithm in period 31 of your successor pair in the future is implemented for your pair. [Offspring, Offspring Switch]

Question 9

Can roles within your pair be switched for payoff?

- ○ No, I always keep my role. [Baseline, Externality, Offspring]
- ○ 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. [Switch]
- ○ Yes, it might be that with 50% probability I get the payoff of the other participant in my pair for the future payoff of my successor pair. [Offspring Switch]

Figure B.7: Control Questions (cont'd)

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

Electronic copy available at: https://ssrn.com/abstract=4002578

Results

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

Period 31: Prediction

The artificially intelligent algorithm decided between (760, 490) and (440, 710) in this period 31.
Based on the previous decisions of participant A in your pair, the prediction and decision of the algorithm was **option X**.

[Switch for groups with payoff swapping]
For this payoff, the roles in your group have been swapped.
You will receive the payoff of participant B.

Your payout from period 31 is therefore 760 [490] points.

Figure B.8: Feedback in Parts 2 and 3 (Example)

Notes: This screen was shown to all participants after the dictator's and the algorithm's decisions. 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

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**.

[Offspring, Offspring Switch]
You will receive an additional payoff at a future point in time.
Once your successor pair has participated in the experiment, you will receive this additional payoff.
It is equal to 50% of the payoff from period 31 in your successor pair.

Figure B.9: Final Results (Example)

Notes: This screen was shown to all participants at the end of the experiment before the final questionnaire. The numbers and option choice are for illustrative purposes only. The content of the paragraphs which dynamically varied across treatments is marked accordingly.

Electronic copy available at: https://ssrn.com/abstract=4002578

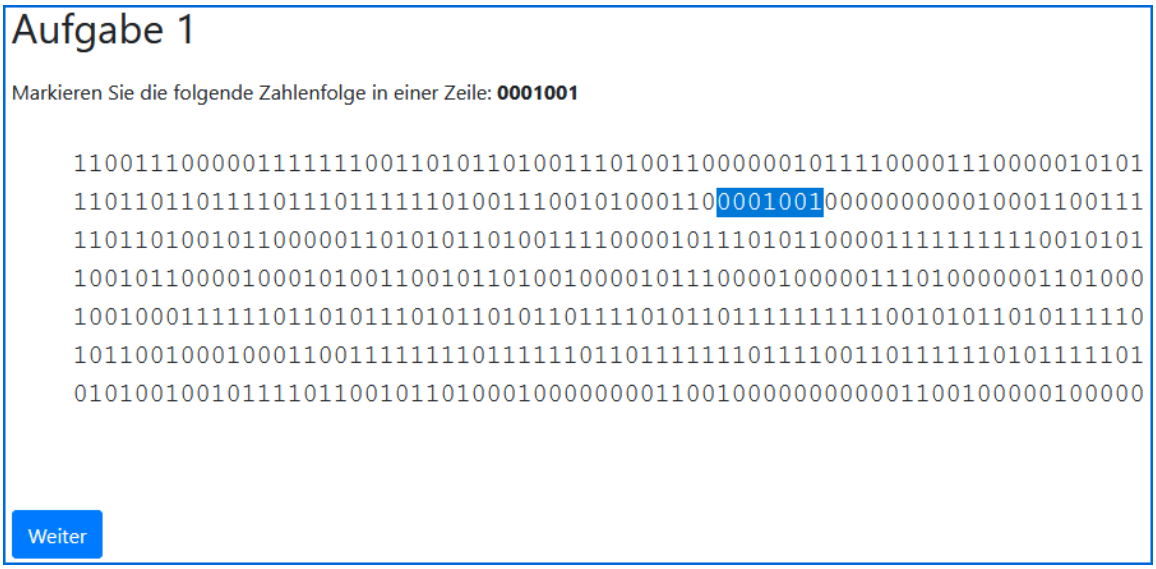

Figure B.10: Real Effort Task

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

Electronic copy available at: https://ssrn.com/abstract=4002578

C Appendix Tables

Table C.1: 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, *i.e.*, 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".

Electronic copy available at: https://ssrn.com/abstract=4002578

Table C.2: 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, *i.e.*, 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 C.3: Summary Statistics, by Treatment

Treatments	Baseline	Externality	Offspring	Offspring Switch	Switch
% Females	60.29	41.18**	51.67	51.61	57.81
Mean age in years	24.28	26.18**	23.28	23.97	25.14
	(0.60)	(0.66)	(0.34)	(0.39)	(0.61)
% Studies in Economics	33.82	44.12	41.67	38.71	43.75
Mean nb Semesters	7.01	7.94	6.60	7.55	7.33
	(0.42)	(0.57)	(0.41)	(0.54)	(0.41)
Mean grade	1.96	2.06	1.97	1.88	1.97
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)
Mean expenses	1.41	1.53	1.47	1.58	1.47
	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)
N	68	68	60	62	64

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 Baseline condition. These tests are Fisher's exact tests for all the variables, except age, grade and semester, for which we use t-tests. ** $p < 0.05$.

Electronic copy available at: https://ssrn.com/abstract=4002578

Table C.4: Relative Frequency of Choices of the Selfish Option X, by Treatment and Relative Position of the Dictator

Treatments	Nb Obs.	Option X [Advantageous]	*p–values*	Option X [Disadvantageous]	*p–values*
Baseline	34	57.94% (0.0417)		79.42% (0.0273)	
			0.581		*0.816*
Externality	34	61.18% (0.0407)		78.53% (0.0261)	
			0.705		*0.456*
Offspring	30	59.00% (0.0399)		81.00% (0.0188)	
			0.011		*0.000*
Offspring Switch	31	46.77% (0.0243)		67.42% (0.0266)	
			0.416		*0.045*
Switch	32	50.00% (0.0308)		75.00% (0.0258)	
			0.135		*0.246*
Baseline	34	57.94% (0.0417)		79.42% (0.0273)	

Treatments	Nb Obs.	Option X [Mixed]	*p–values*
Baseline	34	73.53% (0.0355)	
			1.000
Externality	34	73.53% (0.0286)	
			0.177
Offspring	30	79.00% (0.0277)	
			0.000
Offspring Switch	31	61.94% (0.0229)	
			0.227
Switch	32	66.56% (0.0300)	
			0.141
Baseline	34	73.53% (0.0355)	

Notes: The 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. The Baseline appears twice to report comparisons with both Externality and Switch treatments.

Electronic copy available at: https://ssrn.com/abstract=4002578

Table C.5: Relative Frequency of Choices of the Selfish Option X, by Treatment and Efficiency

Treatments	Nb Obs.	Option X [X efficient]	*p–values*	Option X [Y efficient]	*p–values*
Baseline	34	96.32% (0.0101)		48.63% (0.0502)	
			0.500		*0.893*
Externality	34	95.10% (0.0150)		49.61% (0.0526)	
			0.242		*0.306*
Offspring	30	91.94% (0.0228)		56.44% (0.0376)	
			0.156		*0.000*
Offspring Switch	31	95.70% (0.0229)		26.88% (0.0344)	
			0.287		*0.033*
Switch	32	92.45% (0.0268)		38.33% (0.0395)	
			0.171		*0.115*
Baseline	34	96.32% (0.0101)		48.63% (0.0502)	

Notes: The 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. The Baseline appears twice to report comparisons with both Externality and Switch treatments.

Table C.6: Relative Frequency of Choices of the Selfish Option X, by Treatment and Relative Inequality

Treatments	Nb Obs.	Option X [X fairer]	*p–values*	Option X [Y fairer]	*p–values*
Baseline	34	79.71% (0.0272)		57.65% (0.0420)	
			0.940		*0.656*
Externality	34	79.41% (0.0277)		60.29% (0.0417)	
			0.415		*0.667*
Offspring	30	82.33% (0.0213)		57.67% (0.0441)	
			0.000		*0.026*
Offspring Switch	31	68.06% (0.0280)		46.13% (0.0253)	
			0.025		*0.643*
Switch	32	76.88% (0.0263)		48.13% (0.0343)	
			0.458		*0.086*
Baseline	34	79.71% (0.0272)		57.65% (0.0420)	

Treatments	Nb Obs.	Option X [equal]	*p–values*
Baseline	34	73.53% (0.0355)	
			1.000
Externality	34	73.53% (0.0286)	
			0.177
Offspring	30	79.00% (0.0277)	
			0.000
Offspring Switch	31	61.94% (0.0229)	
			0.227
Switch	32	66.56% (0.0300)	
			0.141
Baseline	34	73.53% (0.0355)	

Notes: The 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. The Baseline appears twice to report comparisons with both Externality and Switch treatments.

Electronic copy available at: https://ssrn.com/abstract=4002578

Table C.7: Probability of Choosing the Selfish Option X - Regression Analysis with Socio-Demographic Controls

Dependent variable: Share of choices of option X	Aggregate	Baseline	Externality	Offspring	Offspring Switch	Switch
Female	0.0294	0.0497	0.000833	0.0612	0.0484	0.0162
	(0.0232)	(0.0614)	(0.0549)	(0.0496)	(0.0434)	(0.0508)
Age	0.000379	0.0111**	-0.0167*	0.00947	0.00191	-0.00907
	(0.00305)	(0.00524)	(0.00847)	(0.0121)	(0.00778)	(0.00741)
Nb Semesters	0.00135	-0.00813	0.0152**	0.00647	-0.00280	-0.00119
	(0.00321)	(0.0103)	(0.00683)	(0.00837)	(0.00491)	(0.00681)
Studies in Economics	0.0709***	0.0769	0.0886	0.0752	0.0186	0.0274
	(0.0245)	(0.0623)	(0.0580)	(0.0442)	(0.0488)	(0.0495)
Abitur grade	0.0102	-0.0690	0.0824	0.00458	0.0121	-0.0435
	(0.0214)	(0.0470)	(0.0516)	(0.0525)	(0.0416)	(0.0464)
Average weekly spending	0.0363*	0.0694	0.0568	0.0368	0.0550	-0.0165
	(0.0191)	(0.0559)	(0.0405)	(0.0428)	(0.0386)	(0.0402)
Constant	0.540***	0.480***	0.718***	0.332	0.430*	0.954***
	(0.0787)	(0.144)	(0.204)	(0.217)	(0.242)	(0.175)
N	161	34	34	30	31	32
R^2	0.086	0.300	0.323	0.222	0.174	0.179

Notes: The table reports the estimation results from OLS regressions of the share of selfish options X on socio-demographic measures. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Electronic copy available at: https://ssrn.com/abstract=4002578

D Appendix Figures

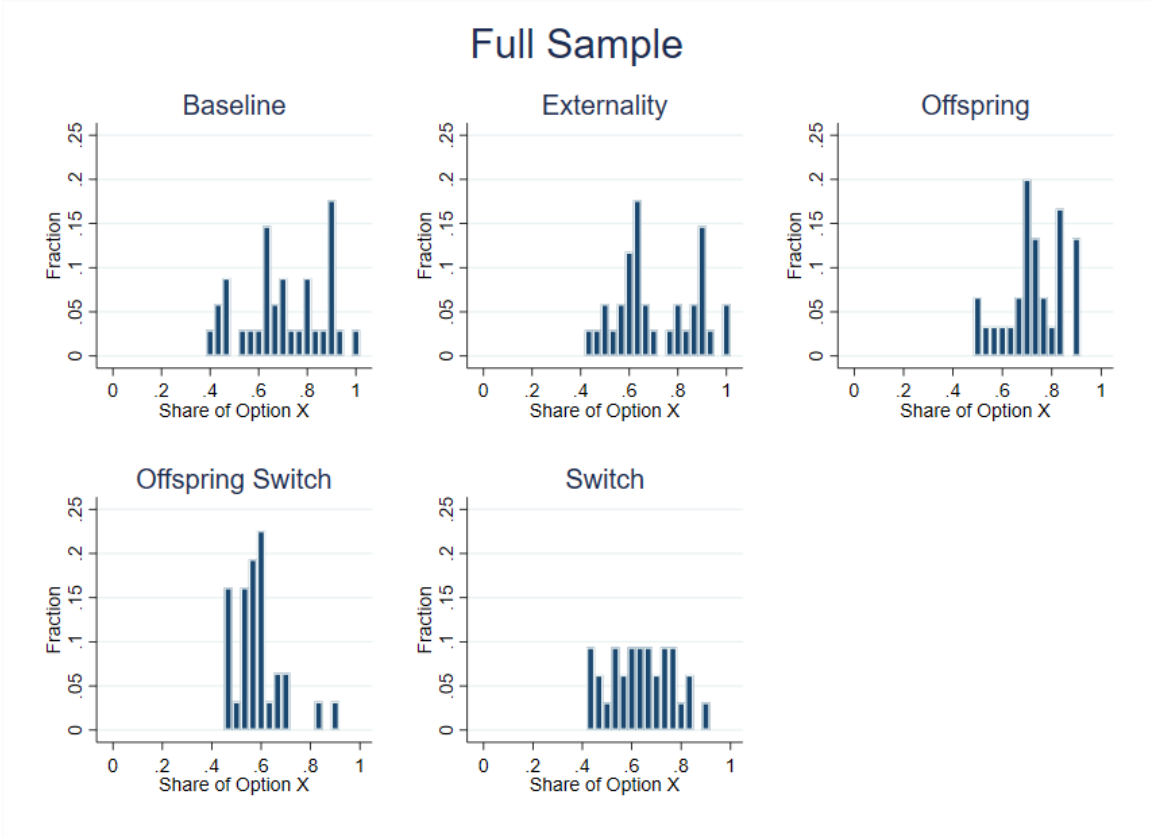

Figure D.1: Distribution of the Shares of Selfish Choices by the Dictators, by Treatment

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

Electronic copy available at: https://ssrn.com/abstract=4002578

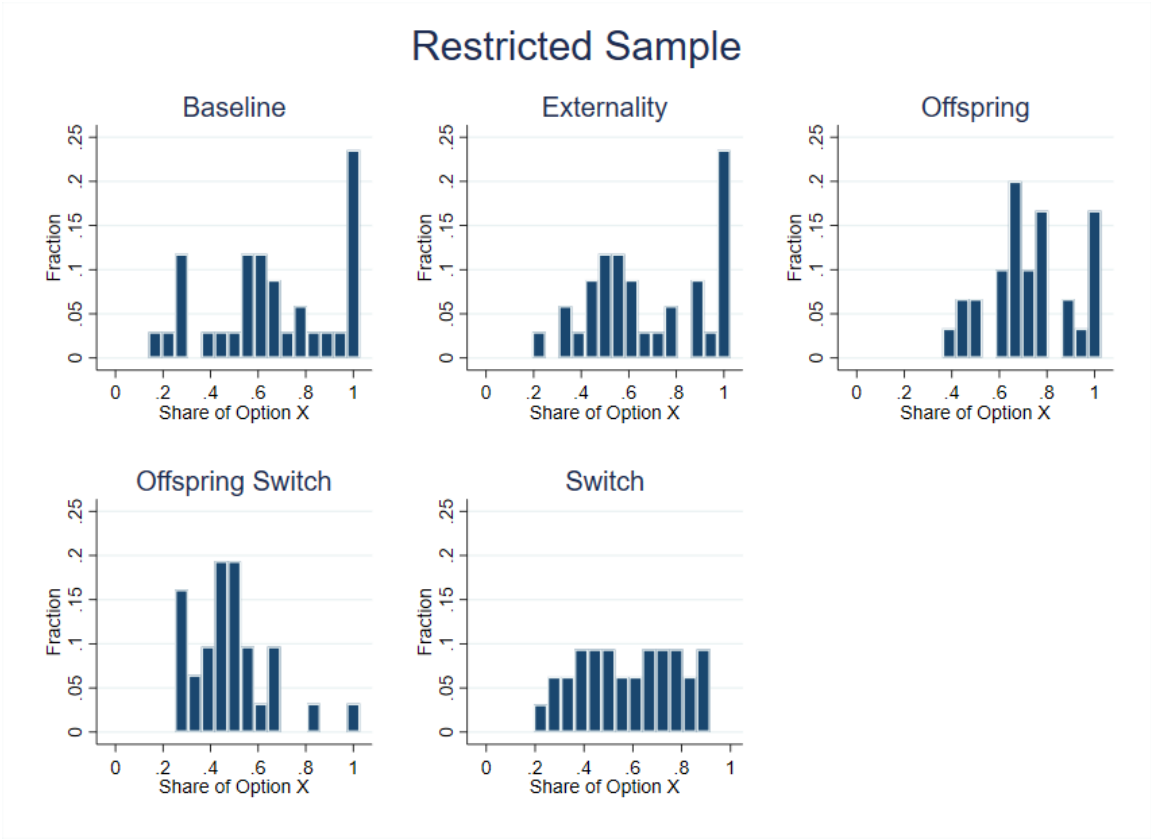

Figure D.2: Distribution of the Shares of Selfish Choices by the Dictators, by Treatment (Restricted Sample)

Notes: The figure displays the distribution of the shares of choices of the selfish option X by the dictators in the 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), by treatment. These games correspond to what is characterized in the analysis as the "restricted sample".

Electronic copy available at: https://ssrn.com/abstract=4002578

Recent Issues

No. 334	Ilya Dergunov, Christoph Meinerding, Christian Schlag	Extreme Inflation and Time-Varying Expected Consumption Growth
No. 333	Vincent R. Lindner	Solidarity without Conditionality. Comparing the EU Covid-19 Safety Nets SURE, Pandemic Crisis Support, and European Guarantee Fund
No. 332	Gyozo Gyöngyösi, Judit Rariga, Emil Verner	The Anatomy of Consumption in a Household Foreign Currency Debt Crisis
No. 331	Ruggero Jappelli, Loriana Pelizzon, Alberto Plazzi	The Core, the Periphery, and the Disaster: Corporate-Sovereign Nexus in COVID-19 Times
No. 330	Jan Krzyzanowski, Uwe Walz	Bank Regulation, Lending and Patenting: Evidence from the EBA Capital Exercise
No. 329	Jana Eisenkopf, Steffen Juranek, Uwe Walz	Responsible Investment and Stock Market Shocks: Short-Term Insurance and Persistent Outperformance Post-Crisis?
No. 328	Hoang Ha Nguyen Thi, Alfons Weichenrieder	C and S Corporation Banks: Did Trump's Tax Reform Lead to Differential Effects?
No. 327	Fincap team	Non-Standard Errors
No. 326	Wenhui Li, Peter Ockenfels, Christian Wilde	The Effect of Ambiguity on Price Formation and Trading Behavior in Financial Markets
No. 325	Alperen A. Gözlügöl	The Clash of 'E' and 'S' of ESG: Just Transition on the Path to Net Zero and the Implications for Sustainable Corporate Governance and Finance
No. 324	Massimiliano Caporina, Michele Costola	Time-Varying Granger Causality Tests for Applications in Global Crude Oil Markets: A Study on the DCC-MGARCH Hong Test
No. 323	Alexandre Corhay, Thilo Kind, Howard Kung, Gonzalo Morales	Discount Rates, Debt Maturity, and the Fiscal Theory
No. 322	Gianluca Anese, Marco Corazza, Michele Costola, Loriana Pelizzon	Impact of Public News Sentiment on Stock Market Index Return and Volatility
No. 321	Ignazio Angeloni, Johannes Kasinger, Chantawit Tantasith	The Geography of Banks in the United States (1990-2020)

Electronic copy available at: https://ssrn.com/abstract=4002578