Public Choice
https://doi.org/10.1007/s11127-024-01178-w

Ruled by robots: preference for algorithmic decision makers and perceptions of their choices

Marina Chugunova[1] · Wolfgang J. Luhan[2]

Received: 22 September 2023 / Accepted: 23 May 2024

Abstract

As technology-assisted decision-making is becoming more widespread, it is important to understand how the algorithmic nature of the decision maker affects how decisions are perceived by those affected. We use an online experiment to study the preference for human or algorithmic decision makers in redistributive decisions. In particular, we consider whether an algorithmic decision maker will be preferred because of its impartiality. Contrary to previous findings, the majority of participants (over 60%) prefer the algorithm as a decision maker over a human—but this is not driven by concerns over biased decisions. However, despite this preference, the decisions made by humans are regarded more favorably. Subjective ratings of the decisions are mainly driven by participants' own material interests and fairness ideals. Participants tolerate any explainable deviation between the actual decision and their ideals but react very strongly and negatively to redistribution decisions that are not consistent with any fairness principles.

Keywords Delegation · Decision-making for others · Algorithm aversion · Redistribution · Fairness

JEL Classification C91 · D31 · D81 · D9 · O33.

The authors gratefully acknowledge funding by the Nuffield Foundation (FR-000000326) and DFG CRC TRR 190 (Project number 280092119). We thank seminar participants at the University of Portsmouth and the Max Planck Institute for Innovation and Competition as well as participants at the 2019 European ESA Meetings and 2021 ESA Global Online Around-the-Clock Meetings for helpful comments and suggestions. We thank Nishan Lin and Georgia Buckle for their comments and assistance. Study materials can be found at OSF: https://osf.io/jucev/.

✉ Wolfgang J. Luhan
wolfgang.luhan@port.ac.uk

[1] Max Planck Institute for Innovation and Competition, Munich, Germany

[2] Faculty of Business and Law, University of Portsmouth, Richmond Building, Portland Street, Portsmouth, NH PO1 3DE, United Kingdom

Published online: 20 June 2024

1 Introduction

Algorithms and Artificial Intelligence (AI) have become an integral part of our decision making, not only for personal and professional, but also political or organizational decisions that systematically affect large groups of people. Examples include the COMPAS system determining recidivism risks of prisoners by the US justice system (Washington, 2018), predictive policing (Meijer & Wessels, 2019), fully automated taxation monitoring systems (Braun Binder, 2018) or the pre-screening of job applicants in organizations (Van Esch et al., 2019). While organizations and public bodies can decide if they want to delegate these decisions to technology or determine how much weight to give to their suggestions, those affected by the decisions cannot (directly) influence if AI decision supports are used. Nevertheless, they may perceive or react differently to a decision depending on whether it was made by a human or an algorithmic decision maker (Bai et al., 2021; Strobel, 2019). We consider a set of important, yet easy to overlook questions: Would those who are affected by the decision prefer an algorithm or a human to make it? How will the nature of the decision maker (DM) affect the perception of and reactions to the decision?

We study these questions in the context of income redistribution. We compare the extreme situations of either a human or an algorithmic DM, excluding the intermediate case of algorithm-augmented decisions to get a clear picture of the reaction to decisions taken by humans or AI.[1] The question of people's perception of and attitude towards algorithmic decisions and AI in general has become more important recently, with many industry leaders warning of the dangers of the threat AI poses and calls for regulation (Criddle, 2023). The stellar rise of AI chat-bots such as ChatGPT (Hu, 2023), however, means that the use of AI is increasingly widespread, for private use but also to support or replace workforce (Brynjolfsson & Raymond, 2023; Vallance, 2023; Dell'Acqua et al., 2023).

We focus on redistributive decisions made on behalf of others as these are common both in a large variety of economic and political decisions, ranging from taxation to social support. Whether people would want or accept an AI DM is particularly relevant for these types of decisions. Unlike in prediction tasks, where algorithms are widely employed and accepted (see, e.g., Humm et al., 2021), there are no objectively correct solutions. In this sense, redistributive decisions can be seen as a type of moral decision, where the definition of correct or fair depends on the observer's personal ideals and beliefs (in the spirit of Kolm, 1996). As a consequence, redistributive decisions may spark controversy and lead to societal tensions, workplace and even international conflicts (e.g., Brams, 2019; Greenberg & Alge, 1998; Klamler, 2019; Sznycer et al., 2017; Wakslak et al., 2007). Identifying which DM is preferred and whose decisions are perceived to be fairer can potentially improve the acceptance of such decisions or policies, and with it, the compliance and support independent of the decision itself. Acceptance and perceived legitimacy in general are fundamental to the democratic system and public acceptance of AI implementation will critically affect how and where it is deployed (Zerilli et al., 2019; de Fine Licht & de Fine Licht, 2020). From the perspective of fairness and acceptance of public choice decisions (see, e.g., Klamler, 2019), the use of AI decision support systems could make a significant difference, especially in the context of public bodies or even political decisions (e.g., Haesevoets et al., 2024). First empirical evidence shows that the use of AI for task allocation, dismissal and hiring affects reactions of decision subjects (Bai et al., 2021; Corgnet,

[1] In this paper, we employ the terms "algorithms" and "AI" interchangeably. This choice is made for the sake of clarity and simplicity, as these terms often overlap in the context of our discussion.

2023; Dargnies et al., 2022), therefore pointing to opportunities and challenges triggered by behavioral responses to AI applications.

The rapidly growing literature on how people perceive algorithmic decisions and engage with algorithms lays the ground for our study. People seem to be willing to outsource analytical tasks to an algorithm but are reluctant to do so with social tasks (Lee, 2018; Waytz & Norton, 2014; Hertz & Wiese, 2019; Buchanan & Hickman, 2023) and are particularly averse to algorithms in the moral domain (Gogoll & Uhl, 2018; Bigman & Gray, 2018). If algorithms are employed in "human tasks", their perceived lack of intuition and subjective judgment capabilities causes them to be judged as less fair and trustworthy (Lee, 2018) or reductionist (Newman et al., 2020). Claure et al. (2023) finds that tasking AI with allocation tasks changes perceptions of dominance and hierarchy. Yet, in general, algorithmic decisions are viewed as more objective (e.g., Cowgill et al., 2020). Our study contributes to this literature by performing a direct empirical test of whether people prefer a human or an algorithm to make redistributive decisions and how decisions made by different DMs are perceived. This allows us to develop clear-cut predictions from the literature while analyzing an economically relevant setting in a novel experimental design. Importantly, our results only indirectly speak to the discussion of whether people are generally averse to algorithms (Dietvorst et al., 2015), appreciate them (Logg et al., 2019), or even over-rely on them (for the overview of the literature see, Chugunova & Sele, 2022). While algorithm aversion has been found in situations where humans delegate tasks to algorithms, evidence on the preferences of *decision subjects* is only starting to emerge (e.g., Dargnies et al., 2022; Fumagalli et al., 2022). Our main focus is not on people who have the discretion to use or not use algorithmic aids, but on those who are affected by these decisions.

A priori, it is not clear if applying algorithms for redistributive decisions would increase or decrease perceived fairness and which DM would be preferred. While humans can arguably better apply ambiguous rules of morality, they can also apply different fairness principles for their own benefit. Equipped with different moral principles, people can always argue that a decision that benefits themselves (or their group) has the moral high ground (Batson & Thompson, 2001; Monin & Merritt, 2012; Epley & Dunning, 2000) or change the fairness principles they adhere to Luhan et al. (2019). As algorithms consistently and selflessly stick to a programmed set of rules, they might therefore score higher on procedural fairness (Hechter, 2013). If people are concerned with the potential bias of the DM, they might prefer an algorithm, even in the context of a moral, redistributive decision.

As an empirical investigation of these questions requires data that cannot be readily found in administrative or company records, we rely on the experimental method widely used for addressing public choice questions (for overviews see Razzolini, 2013; Schram, 2008). We implement an online experiment where a DM can redistribute earnings from three tasks between two players. The closest analogy would be individual team members who all provided effort for a project. Importantly, our setting allows for team members to bring different and often difficult-to-compare inputs to the team performance: e.g., coming up with an idea, putting long hours into implementation, or securing needed material resources. At the end of the project, a manager will decide on how the bonus is allocated. A similar logic would apply to some core questions of public choice, for example, redistributive politics such as taxation or social support policies that reallocate resources within a society or—if one considers not monetary outcomes but welfare—to decisions of public and social services provision such as child care, hospitals, etc. In our experiment, participants can choose if this DM is an algorithm or a human (who has no stake in the outcome) and subsequently express their satisfaction with the decision. The choice of the DM is a proxy for a preference over a DM type. We choose three specific tasks that allow a range of “fair” distributions, depending on the fairness principle applied (based on Konow (2003), see also Sect. 3). This reflects the ambiguity of fair decisions, and allows for a range of differing views on any decision taken. Additionally, depending on the treatment, we provide information on group affiliation, thus varying the potential bias of the DM. Boettke and Thompson (2022) provide an excellent overview regarding the importance of identity in politics. Importantly, when choosing the DM, participants cannot definitively anticipate what decision will be made and how it will affect them, acting under a quasi veil of ignorance (Buchanan & Tullock, 1965) which should lead to an increased desire for fairness.

The way the algorithm is implemented is mimicking the approach of Large Language Models (LLM), generating decisions based on training data. The algorithm makes decisions based on a data set from a specifically designed pre-study (i.e. training data); it is more likely to redistribute if the subjects in the pre-study were more likely to redistribute. In this sense, the algorithm is probabilistic and not rule-based but it is nevertheless impartial as it cannot intentionally change its decision. One could argue that people may prefer using an algorithm because they worry that an individual could make a random or arbitrary decision. We consider this perspective as complimentary to the argument that AI decisions are unbiased due to procedural fairness. However, much as LLMs can make errors (e.g., Buchanan et al., 2023; Roberts et al., 2023), the algorithm in our experiment can produce decisions that do not follow any of the established fairness principles.

We find a strong and robust preference for an algorithmic DM. Regardless of the potential bias of the human DM, more than 64% of participants prefer the algorithm across treatments. Participants are less likely to choose an algorithm if they have earned more than their opponent from effort or talent tasks. However, this preference does not seem to be driven by expected performance differences: the choice of the DM is not determined by the participant’s own fairness ideals. Interestingly, participant’s risk preferences do not contribute to explaining the choice of the DM, suggesting that it is the preference for the type of the DM and not an aversion towards idiosyncratic individual decisions that drives the result. Even though the majority of participants choose an algorithm, the analysis of fairness perceptions reveals that players are (slightly) more satisfied if decisions are made by a human. This result is independent of the actual redistribution decision imposed by the DM. The strongest decrease in satisfaction is triggered by decisions that do not follow a consistent fairness principle (e.g., egalitarian, meritocratic, etc.).

2 Theory and hypotheses

Consider one of the classic examples for distributional fairness in the public choice literature (Klamler, 2019; Brams, 2019): two people have each individually generated an income, which is then pooled, and a third party will decide how this pooled endowment is distributed. While the main theoretical evaluation rests on the comparison of outcomes, such as proportionality or efficiency, we take the perspective of the perceived quality of the decision and which decision maker would be preferred.

The primary question we ask is whether people will prefer a human or an algorithmic DM to make this decision when it affects them. While our study considers the preference for the type of the DM among decision subjects, the closest literature we can relate to analyzes the use and reliance on algorithms. It generally finds opposing results on whether people are averse (e.g., Dietvorst et al., 2015) or appreciative of algorithms (e.g., Logg et al., 2019), but there is no apparent consensus on the overall general preference. In moral contexts, however, such as in our experiment where decisions are driven by fairness principles and beliefs, people are found to have a particularly strong aversion to algorithms (Gogoll & Uhl, 2018; Bigman & Gray, 2018) while simultaneously seeing them as more objective and rational than a human advisor (Dijkstra et al., 1998) and with a "halo" of scientific authority (Cowgill et al., 2020). The perceived fairness of automated decisions may also be driven by the increased procedural fairness associated with the use of algorithms, as they decide "without regard for persons" (Weber, 1978, p. 975 on benefits of bureaucracy). Moreover, as most algorithms use large amounts of data, algorithmic choices are unlikely to be driven by outliers. In that, a preference for algorithmic decisions could reflect a reluctance to rely on decisions taken by a single individual, which might be prone to idiosyncrasies. A related argument can be found in Wilson (2012), who identifies fairness to be defined in a social context where individual concepts are embedded in a shared set of definitions (ultimately leading to rules). Algorithms, trained on large amounts of "fairness" data, would therefore be a better representation of what is fair than an individual's fairness principle. We test these opposing motives by systematically varying whether there is room for potential discrimination. It allows us to observe whether the mere *possibility* of discrimination affects preferences for the type of DM. Simply put, the human has the potential for discrimination, the algorithm can not change its decision at will.

H_1 If there is no scope for bias, a human DM will be preferred over an automated one.

H_2 If there is scope for bias an automated DM will be preferred.

The literature not only analyzes the preference to use and rely on algorithms but even more so how their decisions are perceived, specifically as compared to the decisions of humans. Although there is no unifying finding—for example decisions are viewed as more objective (Dijkstra et al., 1998) and fair (Bai et al., 2021) in some studies, yet as reductionist (Newman et al., 2020) and ignoring unique features of individuals (Longoni et al., 2019) in others—all of the literature finds that the nature of the DM matters for how decisions are perceived. Therefore, our hypothesis is non-directional.

H_3 The nature of the DM affects the perceptions of fairness and satisfaction with the decision.

A biased DM can disadvantage or favor a decision subject (in the following, *negative* and *positive discrimination* respectively). In our experiment, we define negative (positive) discrimination as a reduction (an increase) of earnings due to the revealed features of the affected person—specifically the choice of a painting (see Sect. 3).[2] If we only consider the potential monetary benefits of positive discrimination, we would expect that this will lead to a preference for the human DM over the impartial algorithm:

H_{4a} Expected *positive* discrimination will *increase* the choice of a human DM as compared to a situation without discrimination.

[2] These redistributions of earnings could be in line with some fairness ideal but would benefit an in-group. As we focus on the changes in behavior and perceptions based on expected discrimination, we do not consider if DMs indeed discriminate.

This view, however, neglects any form of social preferences and implies that people solely care about their own outcomes which contradicts ample empirical evidence (e.g., Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999). If we assume that the monetary incentive outweighs social preferences, hypothesis 4a will still hold. If their fairness preferences outweigh the monetary incentives, subjects might prefer a more equal outcome over potential positive discrimination and hence will prefer the algorithm if they believe the human would treat them favorably.

H_{4b} Expected *positive* discrimination will *decrease* the choice of human DM as compared to a situation without discrimination.

Expected negative discrimination should have a straightforward impact on the preference for the algorithm over a human DM. Irrespective of the starting point of the relative income distribution, the decision of the algorithm would always be strictly preferred to the one of a negatively biased human as a biased human would, in each circumstance, either reduce the fairness of the outcome and/or the income of the subject.

H_{4c} Expected *negative* discrimination will *decrease* the choice of human DM as compared to a situation without discrimination.

As an additional test for the validity of these effects (4a, 4b, and 4c), we expect to not observe any significant change between a situation where there is no discrimination possible (no information about the painting choice), and a situation where discrimination is possible, but not applicable (information provided but the painting choices are identical).

3 Design and procedure

We create a scenario where we can observe participants' preference for either a human or an algorithmic DM to redistribute income that they had previously generated. We incorporated the possibility of discrimination to examine whether this would increase the preference for the algorithm as an impartial DM. In addition, we measure, ceteris paribus, the satisfaction and the perceived fairness of the decision, depending on the DM and potential discrimination.

Income generation To start, participants individually earned an initial income by completing three tasks that mimic three potential determinants of income that are central to major fairness theories: luck, effort and talent (Konow, 2003).[3]

In the luck task, participants could earn 100 tokens via a coin toss. In the effort task, participants were given 15 s to count the zeros in two matrices of zeros and ones for 100 tokens each. In the talent task, participants earned 100 tokens for solving a matrix from the Raven fluid intelligence test correctly. In the description of effort and talent tasks,

[3] These principles are closely related to the requirements of proportionality, envy-freeness and efficiency for a fair division (Brams, 2019). All of Konow (2003) principles fulfill these notions, provided that the parties agree on the principle applied.

participants were told that attention to detail and innate abilities respectively are of major importance for performing well.[4]

Participants knew that the tokens would be exchanged for cash (Euro) at the end of the experiment and that token-earnings for each individual task could vary in the exchange rate from 1 to 6 cents per token. This design feature offers two benefits. First, the separate tokens earned via luck, effort and talent allow us to clearly distinguish the fairness principle behind any distributive decision. We focus on four principles (see, e.g., Cappelen et al., 2007; Cappelen et al., 2010; Konow, 2003, 1996; Luhan et al., 2019): egalitarian, choice egalitarian, meritocratic, and libertarian. The aim of our study is to shed light on which decision maker will be preferred in a situation where fairness principles play a role, but we are not primarily interested in the fairness principles of the DMs. The existence of an array of potentially fair behaviors and redistributions enables DMs to discriminate against one participant while still making a *fair* decision.[5] Being able to determine a fairness principle also allows us to consider the effect of the discrepancy between participants' own fairness ideals and those of the DM on satisfaction with the decision. Second, the fact that the monetary value of the tokens was not known ex-ante and could vary forces all participants to see all tasks as equally important and not focus on single income elements or just the total number of tokens.

Choice of a decision maker To test our H_1 on the general preference for a human DM or an algorithm to redistribute the earnings, we paired two participants and informed them of their own and the other person's token earnings from all three tasks. Both participants could then individually opt for a human or an algorithmic DM. In case of a unanimous choice of one DM, it would be implemented, in case of disagreement the decision of one participant would be chosen with equal probability.

The human DM was an anonymous and uninvolved third party. Participants were told that the person received the same explanation about the tasks that generated the incomes as they did. DMs received no other information about the two participants other than their income portfolios, nor were they given any instructions on how to decide other than to "make a fair decision". The actions of the DMs were not incentivized: they received a flat payment regardless of their choices. All of this was common knowledge.

As for the description of the algorithm, we deliberately did not reveal detailed information about the mechanics of the algorithm to keep the information status close to the real world where people are generally aware of, for example, how their sat-nav calculates routes, but are not informed about exact computations behind the recommendation. We therefore—truthfully—informed participants that the algorithm would choose a *"fair distribution based on data from a survey of several hundred participants. The survey participants were informed about the three tasks you completed in stage 1 and then determined what a fair distribution is. The algorithm will apply these decision patterns to the group's income and determine a fair distribution"*. The description mirrors the description of the

[4] The Raven test measures *fluid intelligence* that is considered to be innate. Since the task was performed in an online setting, we timed it to prevent cheating.

[5] Any analysis of fairness principles rests on the assumption that the tasks used for income generation do in fact represent luck, effort, and talent. Luhan et al. (2019) have analyzed the perception of a coin-toss, counting zeros in matrices of zeros and ones and in a language-based intelligence test. They find that the first two tasks are seen as being almost exclusively determined by luck and effort respectively, while the last was seen as mostly talent with a small element of effort. In this study we limit the role of effort in the "talent-task" by using only one question. We are therefore confident that the distribution of earnings can be classified using Konow (2003) fairness principles.

human DM as closely as possible (see the instructions here) and it clearly states that the data used by the algorithm is not historic[6], was specifically tailored to the tasks the participants faced and that the decision involved some transformation of the data. It could be argued that the preference for the algorithm might be driven by the concern that a single individual will make an arbitrary or unconventional decision. As most algorithms use data for their decisions and individual decisions are typical for many organizational settings, we believe it is a correct comparison. Ultimately, it can be viewed as an alternative manifestation of the impartiality of algorithmic decisions. Furthermore, even if the human DM is expected to make a fair decision, this could stem from any fairness principle, while the algorithm is more likely to align with the most prevalent fairness principle. The description of both DMs highlights the differences in possible approaches to a fair decision, which is at the core of our study. This is, of course not representative of all algorithmic decisions, but it is the focus of our research question.

To implement our decision algorithm, we conducted an online survey via Prolific.co, with 506 participants (253 male and 253 female) from the UK and Germany. The survey participants were asked to determine a fair redistribution of tokens for hypothetical pairs of players. They saw the same tasks as in the subsequent experiment with identical explanations and made separate decisions for tokens from each task. The nine situations the survey participants faced covered all initial token distributions that could occur in the experiment, with either one person earning more, or both starting with equal amounts for each task type. Based on the answers, we programmed an automated DM. It considered if the tokens to be redistributed stem from effort, luck or talent and if participants have an equal or unequal number of tokens. Next, it determined the redistribution using answers of the survey participants as probability weights. For instance, in the effort task if one participant in the pair had 100 tokens and another 0, 76.48% of survey respondents did not redistribute the tokens within the pair, therefore with 76.48% probability the algorithmic decision maker would not redistribute the points either. 21.94% of survey respondents split them equally which resulted in a 21.94% chance that the algorithm would do the same and so on.

To simplify the design and further interpretation, we did not allow for continuous redistribution for either type of the DM. The DM could redistribute the tokens of a certain task evenly, give them all to one of the players or keep unchanged.[7]

The experiment created a choice between an algorithm that was fair—based on the fairness principles held by several hundred people—and a human DM who was asked to make a fair decision. We discussed above that, based on the literature, generally human DMs are preferred in situations concerning moral questions. However, if the DM could be biased the preference might switch to the impartial algorithm.

Negative and Positive Discrimination To test the role of bias as formulated in our hypotheses H_2, H_{4a}, H_{4b} and H_{4c}, we introduced a source of potential discrimination for human DMs. We used a purely lab-induced feature to keep this source of discrimination free from the possible confounding effects of real-world biases. At the beginning of the experiment, all participants (including the human DMs) saw two paintings and were asked to select the one they preferred. This simple choice, if revealed to others, has been shown to induce perceptions of an in-group and an out-group amongst participants, which in the

[6] As elaborated in, for example, Cowgill and Tucker (2019) or Motoki et al. (2024) historical data may lead to amplifying existing stereotypes.

[7] Luck and talent tasks resulted in binary outcomes (100 or 0 points). The real effort task consisted of two screens and therefore allowed for three possible outcomes (200,100 or 0 tokens). Therefore, for the real effort task tokens DMs could redistribute in steps of 100 tokens.

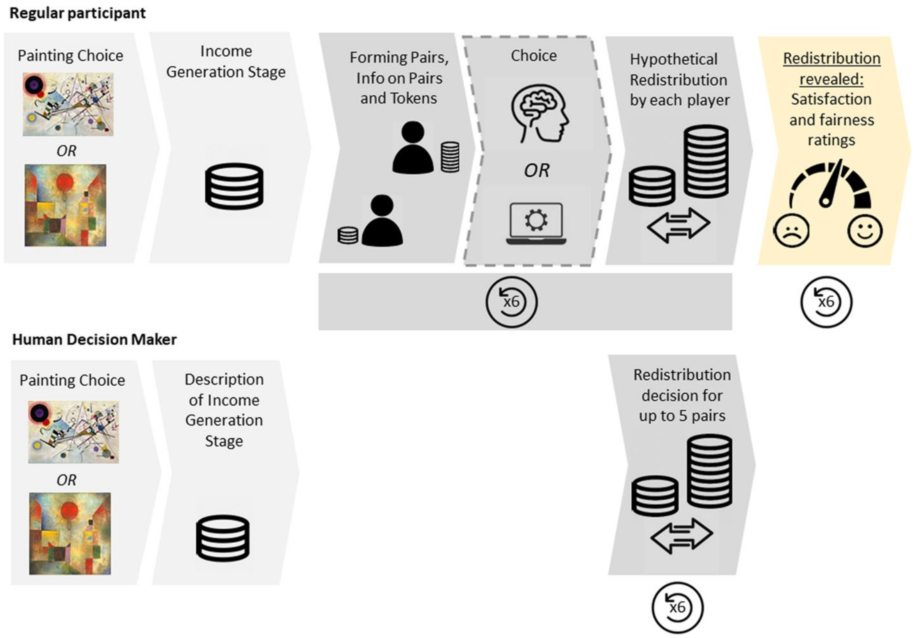
Regular participant
Painting Choice
OR
Income Generation Stage
Forming Pairs, Info on Pairs and Tokens
Choice
OR
Hypothetical Redistribution by each player
Redistribution revealed: Satisfaction and fairness ratings
x6
x6
Human Decision Maker
Painting Choice
OR
Description of Income Generation Stage
Redistribution decision for up to 5 pairs
x6

Fig. 1 Sequence of events in all treatments for regular participants and human DMs

absence of any other information can lead to discriminatory behavior (Tajfel, 1970). The DM might favor her in-group due to, for instance, homophily (McPherson et al., 2001; Chen & Li, 2009). By design, we do not incentivize any sort of discrimination, as the payment of the DM is independent of the decision, and therefore we test the lower bound of the effect. Even if the DM does not actually favor the members of the in-group, the introduction of the group information allows for discrimination and therefore may affect the preference of decision subjects.

Experimental Treatments Our experimental treatment varies if the group information is revealed. Participants choose a painting in all treatments. In *NoInfo* no further mention of this was made in the experiment and this choice was not revealed to anybody. In *Info,* the information about the painting choice is revealed within the matching group: participants in the pair knew the paintings of each other and of the (potential) DM and knew that the DM would have the same information.

Timeline of the experiment Figure 1 provides an overview of the timeline in all treatments. After choosing a painting, participants were randomly assigned to be *regular* participants or *DMs*.[8] Regular participants received explanations for the tasks in the *income generation stage* and performed them, the DMs received the same explanations with a note that only the regular participants perform the tasks. In the *redistribution stage*, participants were matched into pairs, learned about both participant's earnings and were asked to choose a DM who would redistribute the pair's income. In Info, the information on the painting choice was revealed alongside the information on token earnings. Each participant

[8] In the instructions they were called Type P and Type D to avoid any framing effects (the complete instructions can be found here).

faced one treatment only. The redistribution stage consisted of six repetitions with different random matching groups. In all treatments, participants were shown their own and the matched player's token portfolios and were informed that the tokens would be redistributed within the pair. To have a proxy of participants' own fairness preference and to be able to control for the differences between preferred and implemented decisions, participants were asked to make a hypothetical decision on what distribution they would think was fair for their pair. The DMs learned the token portfolio of the pair and could separately decide for each type of token if it should be redistributed. It was common knowledge that both the DMs as well as players were not aware of the value of each token at this point. Only after all repetitions of the *redistribution stage*, regular players were shown, one-by-one, all six pairs that they were part of and learned what redistribution decision was made for each of them. Participants were informed of the nature of the DM, the painting choices of all involved parties (in Info), and the outcome of the redistribution. Participants were asked to indicate on separate seven-point Likert scales how happy they were with the redistribution decision and if they considered it to be fair. A random draw determined the payoff relevant pair and the Euro value of the tokens from each task was revealed. Based on this information, participants were informed about how much they earned in the experiment.

After the experiment was completed, players filled out the questionnaire including basic demographic characteristics, self-evaluations of trust, risk, a shortened version of the readiness for technology scale (Neyer et al., 2012) and social justice orientation scale (Hülle et al., 2018) and asked several questions on their attitudes towards technology.

Procedures The experiment was implemented online using oTree (Chen et al., 2016) with participants recruited from the subject pool of the WiSo Laboratory of the University of Hamburg using hroot (Bock et al., 2014). Although implemented online, participants were required to conform to usual laboratory procedures such as the possibility of questions and simultaneous start of the sessions. All participants took part only once. In total, 212 participants took part in the experiment, 126 in the Info treatment and 86 in the Non Info treatment.[9] The sessions were gender-balanced and the average age of participants was 25.9. 98% of participants were students. By treatment, there are no differences on observable characteristics of the participants. The average payment was 8.72 Euro for 45 min.

In each treatment we randomly allocated two human DMs per session, each deciding for several pairs of regular participants. They received a flat payment of 10 Euro regardless of their decisions.

4 Results

4.1 Choice of the decision maker

Table 1 contains the absolute and relative frequencies of DM choices from NoInfo and Info along with the p-values from non-parametric inference tests. We find an overall preference for the AI DM. In the absence of information on the group membership (the chosen painting), the algorithm is preferred in 63.25% of all choices. We reject our first hypothesis that

[9] Our a priori power analysis documents that our sample size would allow us to detect a medium effect size (Cohen's d = 0.45) with 80% power (α-error probability of 0.05).

Table 1 Chosen decision maker

	AI	Human	Total	χ^2Info	χ^2NoInfo	BI AI = H = 0.5
Positive	156	90	246		p = 0.965	p < 0.001
(%)	(63.41)	(36.59)	(100.00)			
None	122	70	192		p = 0.943	p < 0.001
(%)	(63.54)	(36.46)	(100.00)			
Negative	159	87	246		p = 0.714	p < 0.001
(%)	(64.63)	(35.37)	(100.00)			
Total Info	437	247	684	p = 0.954	p = 0.824	p < 0.001
(%)	(63.89)	(36.11)	(100.00)			
Total NoInfo	296	172	468			p < 0.001
(%)	(63.25)	(36.37)	(100.00)			

Frequencies of choices in treatments Info (AI or human DM *with* the group info) and NoInfo (AI or human DM *without* the group info). Percentages in parentheses below absolute numbers of observations. Column χ^2Info displays the p value for the test of differences between the discrimination classes within Info. Column χ^2NoInfo contains the p values from the individual tests of the observations in the respective row against the observations in NoInfo. The final column contains the p values from binomial tests of the observations against a hypothetical 50% frequency of AI choices (or human choices respectively)

if there is no possibility of discrimination, the human DM is preferred. We find quite the contrary, that the AI is chosen significantly more frequently than 50% (two-sided binomial test $p < 0.001$).

Revealing the information on the choice of the painting for all parties—and introducing the potential for discrimination—does not change this preference and we find an almost identical 63.89% choice majority for the algorithm in Info. When comparing the choices between treatments, there is no significant difference in the preferred DM (χ^2 $p = 0.824$). We conclude from these findings that the general preference for the AI DM rather *prevails* than *appears* in Info (two-sided binominal test, $p < 0.001$). It is not the prospect of discrimination that drives this overall preference for the AI DM and we reject H_2.

As potential positive discrimination for one member of the pair means potential negative discrimination for the other in our setting, the aggregate result of no effect of potential discrimination could be due to the fact that choices under positive discrimination ($H_{4a,b}$) are balanced out by choices under negative discrimination (H_{4c}). We therefore split up the sample into the three classes of potential discrimination (positive, negative, and no discrimination) and analyze the effects of each type of discrimination separately. We again do not find any impact of potential discrimination on the choice of the DM. In all three cases, we observe a strong preference for the AI as a DM, and no significant difference to any of the other discrimination types in the information treatment ($\chi^2 p = 0.954$) or to the treatment without information (see column χ^2 NoInfo in Table 1). Irrespective of potential positive or negative discrimination, the majority of choices are for the AI DM and we reject our hypotheses H_{4a}, H_{4b}, and H_{4c}. As a final non-parametric test, we implement a trend test but do not find a significant trend in our observations when ranked by order of potential discrimination (two-sided Jonckheere-Terpstra test $p < 0.7784$).

We implement a series of pooled probit regressions with robust standard errors clustered on the individual level to identify ceteris paribus influences on the choice of the DM. We consider if the probability of choosing a human DM relates to the token portfolio of the paired participants, whether the group information was revealed, and the resulting potential

Table 2 Determinants: choice of decision maker. Pooled probit regression

Dep.: *Choice Human DM*	(1)	(2)	(3)	(4)	(5)
Variables					
Info	0.0786	0.0201	0.00881	0.00598	0.00494
	(0.155)	(0.158)	(0.157)	(0.159)	(0.158)
Positive discrimination	− 0.0230	− 0.0248	− 0.0197	− 0.00941	− 0.0177
	(0.144)	(0.149)	(0.149)	(0.149)	(0.149)
Negative discrimination	− 0.0518	− 0.0561	− 0.0549	− 0.0556	− 0.0519
	(0.157)	(0.160)	(0.160)	(0.160)	(0.160)
Tokens luck	− 0.000527				
	(0.00127)				
Tokens effort	0.00268***				
	(0.000801)				
Tokens talent	0.00257*				
	(0.00135)				
Tokens luck partner	0.00106				
	(0.000850)				
Tokens effort partner	0.00115**				
	(0.000520)				
Tokens talent partner	− 0.000232				
	(0.000933)				
Distance earnings luck		− 0.000773			
		(0.000710)			
Distance earnings effort		0.000727			
		(0.000491)			
Distance earnings talent		0.00131			
		(0.000812)			
More luck			− 0.0969		− 0.161
			(0.116)		(0.254)
More effort			0.213**		0.210*
			(0.107)		(0.108)
More talent			0.283**		0.262*
			(0.143)		(0.153)
Lose tokens egalitarian				0.336	
				(0.319)	
Lose tokens choice				0.216	
				(0.225)	
Lose tokens meritocratic				− 0.178	
				(0.136)	
Count lose					0.0486
					(0.159)
Human fair	0.107*	0.102*	0.102*	0.0988*	0.103*
	(0.0553)	(0.0557)	(0.0556)	(0.0563)	(0.0559)
Unbiased	0.131*	0.122*	0.125*	0.125*	0.124*
	(0.0691)	(0.0689)	(0.0686)	(0.0686)	(0.0685)

Table 2 (continued)

Dep.: *Choice Human DM* Variables	(1)	(2)	(3)	(4)	(5)
Constant	− 0.974**	− 0.296	− 0.580	− 0.616	− 0.643
	(0.422)	(0.374)	(0.422)	(0.491)	(0.479)
Controls	✓	✓	✓	✓	✓
Observations	1,152	1,152	1,152	1,152	1,152
pseudo R^2	0.041	0.022	0.025	0.018	0.025

Observations from NoInfo and Info, Robust standard errors, clustered at the individual level in parentheses. The base category for the discrimination dummies is "no discrimination", which corresponds to either the NoInfo treatment or both members of the pair had identical information.

Additional controls in all specifications (non-significant): age, gender, technical readiness scale by Neyer et al. (2012), and a trust score measured via three items in the questionnaire.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

discrimination for the participant. Additionally, we consider implications from the various fairness principles and a small range of control variables from the questionnaire (Table 2). First, we reconfirm that neither the availability of the group information (variable *Info*), nor the expected direction of discrimination drive the choice of the DM.

Next, people may be more or less likely to prefer human or AI DMs depending on the differences in the earnings between themselves and the paired partner. In separate specifications (columns 1 and 2) we capture differences in token earnings within the pair. In column 1, we include the tokens earnings from all three tasks by the participant and their partner as individual variables. We find that earning more from effort and talent tasks increases the likelihood of choosing a human DM. In column 2, we calculate the absolute distance between the two participants' earnings. As none of these distances significantly affect the probability of choosing a human DM, this does not seem to reflect how participants considered the earnings when choosing the DM. In column 3, we include a binary variable that captures whether the focal participant had *more* tokens of each kind than the partner. We again find a significant positive impact of the earnings from effort and talent, but not the luck task on the choice of the human DM.[10]

The fact that an advantage in some token types is more important than in others might be due to the expectation that a human DM would hold a fairness principle that is more favorable to their (higher) earnings. In column 4 of Table 2, we determined whether the participant would lose tokens (these would be redistributed to the other participant) if the DM held one of the four fairness ideals (see Sect. 3). We find no impact of this prospect of losing tokens under one of the fairness principles. However, this specification assumes that the participants are aware of these principles and mentally process the displayed earning tables in a very sophisticated way. To relax this assumption in column 5, we simplify this approach by creating a variable that counts under how many of the fairness ideals the participant would lose tokens to the partner. This variable ranges from 0 to 3 and is a simple representation of how likely it is that a fair DM will redistribute money away from the

[10] Alternative specifications, e.g., with lower earnings or measured in absolute and relative distances, showed no significant effect. The reported specifications were chosen based on goodness-of-fit statistics.

participant.[11] We find that even this simple specification does not yield a significant impact on the choice of the DM, and we can conclude that participants consider fairness principles only in a very limited way.

In addition, we control for the participants' age, gender, whether they are classified as technology-ready, whether they are trusting, and two opinion items on fair and unbiased decision making from our questionnaire. Only the two opinion items[12] have a significant impact on the choice of the DM. *Human Fair* askes for a rating of who is better at making fair and just decisions, the AI or humans. As expected, the higher participants rate this ability for humans, the more likely they are to opt for a human DM. *Unbiased* records whether people believed that it is hard for humans to make unbiased decisions. Unsurprisingly, the more participants believe that this would be easy for humans, the more they pick the human DM.[13] In the questionnaire we also elicit risk preferences of individuals, but these are not significantly correlated with the choice of the human DM (Pearson correlation $= -0.01, p = 0.7$) and do not prove to be contributing to explanatory power or the goodness of fit of the model. This result may be regarded as suggestive evidence that the preference for the AI DM is not driven by the risk associated with entrusting the decision to a single individual DM.

4.2 Satisfaction with the decision and perceived fairness

To abstract from individual differences that might affect the level of perceived fairness and satisfaction irrespective of the treatment, we run a fixed effect regression, controlling for several parameters of the decision situation. For each type of tokens, we consider if the number of tokens increased or decreased after the redistribution as compared to the initial earnings (variable *Before-After*), whether a person has the same fairness ideals as the DM (variable *Hyp-Actual*) and, in line with several fairness theories (see, e.g., Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999), the token-difference after the redistribution (*Own-Partner (After)*). Additionally, we introduce dummy variables that capture whether the DM was human, if the player lost tokens overall, and what type of discrimination (positive/negative or none) the player could expect in the pair. Importantly, we add a dummy variable for "non-ideal" redistributions (*Non-Ideal*). This captures whether the implemented redistribution does not correspond to any of the major fairness ideals and therefore may be regarded as inconsistent. As probabilities for the decisions of the algorithm were drawn independently *per task category*—following the results from the survey—the algorithm inevitably ended up being less consistent with the applied principles: 13% of AI decisions were inconsistent, i.e., not following one principle, as compared to only 3% of human ones (t-test, $p < 0.001$).[14] In total 9.2% of all redistribution decisions were classified as inconsistent (Non-Ideal equals to 1). We consider fairness and satisfaction rankings

[11] By definition, no redistribution can take place under the libertarian principle therefore this is not included in the analysis.

[12] We use a 14-question battery concerning decision-making abilities of humans and AI—all on five-point Likert scales. From the responses to the individual questions we generate two-factor variables and two opinion scales that are used in the regressions. Other variables elicited in the questionnaire do not contribute to the explanatory power of the model and are not reported.

[13] Although these variables seemingly capture very related concepts, the correlation between them is rather low and insignificant.

[14] Appendix A.1 provides further details on the decisions made by algorithms and human decision makers.

Table 3 Determinants of satisfaction and fairness ratings: Fixed effects panel regression

	(1)	(2)	(3)	(4)
VARIABLES	Satisfaction	Fairness	Satisfaction	Fairness
Non-ideal (0/1)	− 0.781***	− 1.376***	− 0.729***	− 1.340***
	(0.207)	(0.246)	(0.213)	(0.240)
DM human (0/1)	0.248**	0.241*	0.266**	0.254**
	(0.107)	(0.123)	(0.107)	(0.120)
Non-ideal#DM human			− 0.352	− 0.248
			(0.418)	(0.664)
Lost tokens (0/1)	− 0.519**	− 0.899***	− 0.525**	− 0.903***
	(0.210)	(0.250)	(0.209)	(0.249)
Preferred DM (0/1)	− 0.0220	0.0412	− 0.0222	0.0410
	(0.0962)	(0.124)	(0.0963)	(0.124)
No discrimination	0.00301	0.00406	0.00337	0.00431
	(0.164)	(0.175)	(0.164)	(0.176)
Neg. discrimination	− 0.166	− 0.0552	− 0.167	− 0.0556
	(0.134)	(0.163)	(0.135)	(0.164)
Luck: own-partner (After)	0.00571***	0.00463**	0.00568***	0.00460**
	(0.00185)	(0.00214)	(0.00185)	(0.00214)
Talent: own-partner (After)	0.00313**	0.000247	0.00317**	0.000279
	(0.00128)	(0.00140)	(0.00128)	(0.00140)
Effort: own-partner (After)	0.00371***	0.00147*	0.00371***	0.00147*
	(0.000821)	(0.000885)	(0.000818)	(0.000880)
Luck: before-after	0.00806***	− 0.00155	0.00803***	− 0.00158
	(0.00247)	(0.00313)	(0.00246)	(0.00313)
Talent: before-after	0.0143***	0.00266	0.0143***	0.00262
	(0.00347)	(0.00417)	(0.00346)	(0.00416)
Effort: before-after	0.0116***	0.00367	0.0115***	0.00365
	(0.00233)	(0.00329)	(0.00233)	(0.00330)
Luck: hyp-actual	− 0.00495*	− 0.00310	− 0.00499*	− 0.00313
	(0.00264)	(0.00271)	(0.00264)	(0.00272)
Talent: hyp-actual	− 0.00223	− 0.000756	− 0.00219	− 0.000725
	(0.00264)	(0.00273)	(0.00264)	(0.00273)
Effort: hyp-actual	− 0.00108	0.000245	− 0.00105	0.000263
	(0.00147)	(0.00150)	(0.00147)	(0.00149)
Constant	1.196***	1.139***	1.190***	1.134***
	(0.161)	(0.194)	(0.162)	(0.194)
Observations	1,152	1,152	1,152	1,152
R^2	0.366	0.187	0.366	0.187
Id.	192	192	192	192

Standard errors in parentheses, “(0/1)” indicates a dummy variable, Non-Ideal indicates decision not consistently adhering to one fairness ideal, Lost Tokens indicates an overall loss after the redistribution, Own-Partner variables contain the difference in tokens between the two participants, Before-After is the difference in tokens after the redistribution, Hyp-Actual is the difference in fairness ideals of the participant and the DM—both indicated in the token distribution. Id reports on the number of unique participants. ***$p <$ 0.01, **$p < 0.05$, *$p < 0.1$

separately—while the two are highly correlated (0.76, $p < 0.001$), they are not identical, which explains why regression results vary slightly.

When the redistribution is inconsistent (*Non-Ideal*), the satisfaction with the decision is reduced by 0.78 points and the perceived fairness by 1.38 points of a 7 point likert scales which correspond to approximately 10% and 20% decreases respectively. Reactions to inconsistent decisions do not depend on the nature of the DM (specifications 3 and 4 in Table 3). We observe some "flexibility" in the notion of fairness in our participant sample. A deviation from the participant's own fairness ideals did not impact their fairness rating (*Hyp-Actual*). The open-field comments confirm that participants were aware of different fairness ideals and tolerated deviations as long as one fairness principle was followed consistently.

Having fewer tokens in total after the redistribution is the second largest driver of satisfaction and fairness (*Lost Tokens*). In Table A5 (Appendix A), we consider if the inconsistency of a certain type of DM is particularly harmful to fairness and satisfaction ratings. We find suggestive evidence that losing tokens as a result of a human decision might negatively affect satisfaction ratings. Having more tokens after the redistribution (*Before-After*) and more than the partner (*Own-Partner (After)*) increases satisfaction but not fairness ratings.

In line with the findings of Gogoll and Uhl (2018) and Bigman and Gray (2018), we observe that if a moral decision is made by a human DM it is rated as about a quarter of a point more fair and participants report higher satisfaction (*DM Human*). We, therefore, fail to reject our H_3 on the DM's nature and the impact on satisfaction and perceived fairness.

Receiving the preferred DM type does not affect fairness and satisfaction ratings (variable *Preferred DM* in Table 3). We also tested whether *having a choice* impacts the participant's perceived satisfaction with and fairness of the decisions. We show in Sect. A.2 in the appendix that this had no significant impact and our results remain unchanged.

Finally, to test if the group information affected the satisfaction and fairness ratings, we run a pooled OLS regression including largely the same controls as in the fixed effect estimation, but adding several demographic variables and the treatment dummy *Info*. The regression results in Table A6 in appendix suggest that revealing group affiliation—and therefore introducing the possibility of discrimination—significantly reduces both the perceived fairness and satisfaction by about a third of a point. Importantly, potential discrimination *per se* decreases the ratings. This result also confirms that the participants paid attention to the group information and the resulting threat of discrimination, yet, it did not affect their choices of the DM in our first set of results. This finding is also in line with the results of Dargnies et al. (2022), who document that removing gender information from an algorithm increases preference for an algorithm in all participants.

5 Discussion and conclusion

We study whether people prefer a human or an algorithm to decide how earnings are redistributed and analyze the impact of discrimination on this preference. We also examine how the nature of the decision maker affects the perceived fairness of and satisfaction with the decision.

Our experiment provides two sets of results. First, with over 60% of participants choosing a redistributive algorithm over a human, we find a strong preference for algorithmic DMs among *decision subjects*. In our experiment, the risk preferences do not seem to contribute to the choice of AI DM either, suggesting that it is not the risk associated with individual decisions that drives the majority choice of algorithmic DMs. This preference for algorithmic DMs persists regardless of the potential discrimination. Therefore, it appears that it is not the perceived impartiality of the algorithm that drives the result. However, the

Table 4 Determinants of satisfaction and fairness ratings: Pooled OLS regression

Rating of	(1) Satisfaction	(2) Fairness	(3) Satisfaction	(4) Fairness
Non-ideal (0/1)	– 0.756***	– 1.375***	– 0.767***	– 1.370***
	(0.140)	(0.152)	(0.141)	(0.154)
DM human (0/1)	0.174*	0.124	0.201*	0.109
	(0.0955)	(0.102)	(0.105)	(0.112)
Lost tokens (0/1)	– 0.524***	– 0.720***	– 0.473**	– 0.747***
	(0.166)	(0.186)	(0.192)	(0.217)
DM human#Lost tokens			– 0.105	0.0564
			(0.213)	(0.219)
Choice (0/1)	0.0413	– 0.0237	0.0404	– 0.0232
	(0.120)	(0.129)	(0.119)	(0.128)
Info (0/1)	– 0.315**	– 0.293*	– 0.316**	– 0.293*
	(0.155)	(0.158)	(0.156)	(0.158)
No discrimination	– 0.142	– 0.0335	– 0.145	– 0.0320
	(0.152)	(0.172)	(0.153)	(0.173)
Neg. discrimination	– 0.309**	– 0.136	– 0.310**	– 0.136
	(0.134)	(0.155)	(0.134)	(0.155)
Luck: own-partner (After)	0.00605***	0.00444***	0.00605***	0.00444***
	(0.00144)	(0.00160)	(0.00144)	(0.00160)
Talent: own-partner (After)	0.00507***	0.00251**	0.00509***	0.00250**
	(0.000915)	(0.000981)	(0.000917)	(0.000981)
Effort: own-partner (After)	0.00519***	0.00336***	0.00520***	0.00335***
	(0.000568)	(0.000610)	(0.000568)	(0.000609)
Luck: before-after	0.00584***	– 0.00193	0.00584***	– 0.00193
	(0.00198)	(0.00233)	(0.00198)	(0.00233)
Talent: before-after	0.00942***	0.000212	0.00942***	0.000215
	(0.00230)	(0.00281)	(0.00230)	(0.00281)
Effort: before-after	0.00772***	0.00210	0.00759***	0.00217
	(0.00170)	(0.00212)	(0.00168)	(0.00210)
Luck: hyp-actual	– 0.00527**	– 0.00309	– 0.00525**	– 0.00311
	(0.00213)	(0.00236)	(0.00213)	(0.00236)
Talent: hyp-actual	– 0.00624***	– 0.00380*	– 0.00623***	– 0.00380*
	(0.00193)	(0.00203)	(0.00193)	(0.00203)
Effort: hyp-actual	– 0.00174	0.000193	– 0.00175	0.000202
	(0.00117)	(0.00114)	(0.00117)	(0.00114)
Female (0/1)	– 0.00726	0.0941	– 0.00618	0.0936
	(0.112)	(0.120)	(0.112)	(0.120)
Age	– 0.0383***	– 0.0480***	– 0.0383***	– 0.0480***
	(0.0131)	(0.0142)	(0.0130)	(0.0142)
Trust	0.183**	0.251***	0.183**	0.251***
	(0.0768)	(0.0837)	(0.0768)	(0.0836)
Constant	2.495***	2.616***	2.488***	2.619***
	(0.383)	(0.412)	(0.382)	(0.412)
Observations	2,004	2,004	2,004	2,004
R-squared	0.333	0.184	0.333	0.184

Pooled OLS regression. "(0/1)' indicates a dummy variable Robust standard errors in parentheses

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Table A5 Continuation: determinants of satisfaction and fairness ratings. Fixed effects panel regression

Variables	(1) Satisfaction	(2) Fairness
Non-ideal (0/1)	− 0.810***	− 1.400***
	(0.207)	(0.246)
DM human (0/1)	0.366***	0.337**
	(0.122)	(0.138)
Lost tokens (0/1)	− 0.351	− 0.761***
	(0.226)	(0.275)
DM human#Lost tokens	− 0.428*	− 0.350
	(0.233)	(0.251)
Preferred DM (0/1)	− 0.0255	0.0383
	(0.0958)	(0.123)
No discrimination	− 0.00164	0.000251
	(0.166)	(0.176)
Neg. discrimination	− 0.164	− 0.0538
	(0.134)	(0.164)
Luck: own-partner (After)	0.00585***	0.00474**
	(0.00183)	(0.00212)
Talent: own-partner (After)	0.00315**	0.000266
	(0.00127)	(0.00139)
Effort: own-partner (After)	0.00372***	0.00148*
	(0.000824)	(0.000890)
Luck: before-after	0.00803***	− 0.00158
	(0.00246)	(0.00311)
Talent: before-after	0.0142***	0.00254
	(0.00348)	(0.00417)
Effort: before-after	0.0109***	0.00313
	(0.00233)	(0.00328)
Luck: hyp-actual	− 0.00468*	− 0.00288
	(0.00262)	(0.00270)
Talent: hyp-actual	− 0.00235	− 0.000853
	(0.00264)	(0.00273)
Effort: hyp-actual	− 0.00122	0.000131
	(0.00148)	(0.00150)
Constant	1.169***	1.117***
	(0.162)	(0.196)
Observations	1,152	1,152
R-squared	0.369	0.189
Number of Obs.	192	192

"(0/1)" indicates a dummy variable. Robust standard errors in parentheses

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

potential for discrimination was acknowledged by the participants and manifested itself in variations in the satisfaction and perceived fairness of the decision. One potential reason for the preference for an algorithm could be that it may fare better on procedural justice,

Table A6 Pooled OLS regression. The effect of revealing the group affiliation on fairness and satisfaction

Variables	(1) Satisfaction	(2) Fairness
Non-ideal (0/1)	− 0.892***	− 1.421***
	(0.198)	(0.212)
DM human (0/1)	0.182	0.136
	(0.118)	(0.128)
Lost tokens (0/1)	− 0.368*	− 0.692***
	(0.214)	(0.248)
Info (0/1)	− 0.399**	− 0.348**
	(0.177)	(0.173)
No discrimination	− 0.136	− 0.0112
	(0.171)	(0.198)
Neg. discrimination	− 0.299*	− 0.0815
	(0.153)	(0.181)
Luck: own-partner (After)	0.00737***	0.00575***
	(0.00193)	(0.00211)
Talent: own-partner (After)	0.00495***	0.00226*
	(0.00127)	(0.00133)
Effort: own-partner (After)	0.00568***	0.00364***
	(0.000773)	(0.000831)
Luck: before-after	0.00713***	− 0.000819
	(0.00264)	(0.00306)
Talent: before-after	0.00668**	− 0.00293
	(0.00321)	(0.00400)
Effort: before-after	0.00875***	0.00267
	(0.00230)	(0.00313)
Luck: hyp-actual	− 0.00513*	− 0.00219
	(0.00285)	(0.00302)
Talent: hyp-actual	− 0.00940***	− 0.00661**
	(0.00277)	(0.00301)
Effort: hyp-actual	− 0.00114	0.000889
	(0.00168)	(0.00163)
Female (0/1)	0.0542	0.153
	(0.154)	(0.166)
Age	− 0.0345**	− 0.0379**
	(0.0152)	(0.0175)
Trust	0.108	0.220**
	(0.102)	(0.111)
Constant	2.433***	2.306***
	(0.496)	(0.545)
Observations	1,152	1,152
R-squared	0.351	0.189

"(0/1)" indicates a dummy variable

Standard errors clustered at the individual level in parentheses

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

i.e., treating all cases the same regardless of discriminable characteristics. Nevertheless, this is speculation and more research is needed to provide a definitive answer.

Second, and somewhat in contrast to the first result of preferring the algorithm, people are less satisfied with algorithmic decisions and they find them less fair than human decisions. Two main factors contribute to lower satisfaction and fairness ratings. Most importantly, decisions have to be consistent with a fairness principle. Participants react very negatively to “mistakes" of both human DMs and algorithms, that is, if they apply fairness principles inconsistently. We do not observe a difference in reactions to mistakes by humans or algorithms, which has been reported in previous studies as one of the reasons for algorithm aversion in delegation settings (e.g., Dietvorst et al., 2015). This result leads us to believe that a more sophisticated algorithm that does not allow for inconsistencies and makes fewer “mistakes" (as e.g., developed by Koster et al., 2022) could elicit a more positive reaction. A smaller, but nevertheless significant factor is indeed the nature of the DM. Decisions made by a human, regardless of the decision itself, are rated better. Based on a recent study by Hidalgo et al. (2021), one might speculate that it might be due to the lack of intentions of algorithmic DMs. Future research could also further explore the role of fairness expectations. The lower satisfaction with AI decisions could stem from the fact that humans apply the expected fairness principles, but the algorithm follows other fairness principles than expected.

Considering the populations affected by redistributive decisions, our results give reasons for optimism for advocates of technology adoption. While technological advancement has always offered clear advantages in terms of operational efficiency (e.g., Solow, 1957; Stiroh, 2001), in the case of redistributive decisions it appears to also align with the preferences of the affected. From a public choice perspective, this means that decisions could be perceived as more fair and therefore increase acceptance and welfare. While the question of our research might appear futuristic at first glance, some companies (e.g., IBM, see Guenole & Feinzig, 2018) are already using AI for compensation planning and in political decisions and public bodies – to determine policing and parole strategies (see examples in the introduction). The people affected by these decisions, even in the moral domain, prefer algorithmic decision makers. While we observe a slight drop in satisfaction with the algorithmic decision as compared to a human one, more sophisticated algorithms that produce internally consistent decisions are likely to overcome it.

A Appendix

A.1 Additional analysis: algorithmic decisions

As explained in the design section, the algorithm was generated using the data of the survey participants from Prolific. The decision makers in the actual experiment differed in their redistributive decisions from the Prolific participants and thus the decisions produced by experimental participants (Type D) and the algorithm systematically differed. Decision makers in the actual experiment tended to make more egalitarian choices for tokens of all colors (ttest $p < 0.001$ in all three cases).

We do not consider differences in performance between human decision makers and the algorithm to be a concern for addressing the research question. First, the difference in performance could not have affected the choice of the preferred decision maker, as the decisions were revealed to the participants after they chose of decision maker. Potentially, the difference in decisions could have affected the satisfaction and fairness scores, but we control for the type of decision in the analysis.

The difference in human and algorithmic decisions could have stemmed from the fact that in some treatments the human decision makers have an opportunity to discriminate based on the group. Yet, we find no evidence that the decision makers in our experiment discriminate the outgroups or make decisions favorable to the ingroups (Luck Tokens: χ^2 (4, N = 190)= 3.05, p = 0.5; Talent Tokens: χ^2(4, N = 205)= 4.05, p = 0.399; Effort Tokens: χ^2(4, N = 217) = 3.6, p = 0.461).

We do not analyze how different conditions affect the behavior of the decision makers because of the very small number of decision makers in our sessions.

A.2 Additional analysis: having a choice

Our focus lies on whether participants prefer—and therefore choose—a human or an algorithm to make the decision for them and whether getting the decision maker they choose additionally influences their satisfaction with the decision. The mere fact of *having a choice* could, of course, also impact on the satisfaction with the decision. Having a choice has been shown to bear an intrinsic value (Bartling et al., 2014). Even closer to our question, Mellizo et al. (e.g., 2014 and Sausgruber et al. 2021) find the so-called endogeneity premium in different domains which states that if certain policies or institutions are chosen and not exogenously imposed, people appear to like them more. In line with this literature, we expect that having the option to make a choice will overall increase the satisfaction with a decision. Interestingly, recent findings by Gallier (2020) suggest that even if one's preference is overruled in the vote, compliance with the new rules is higher if they were endogenously chosen.

We therefore introduced three more treatments with exogenously determined decision makers and no choice for the participants to test these theories and to control for possible interaction effects: in Treatment AI Only (66 participants) the decision maker was an algorithm; in in Human Only Info (34 particpants) the decision maker was human and the information of the painting choice was revealed; in Human Only No Info (34 participants), finally, the decision maker was human and there was no information on the painting choice. We implemented the same estimation approach as in Sect. 4.2, the results can be found in Table A4. Comparing treatments with a choice (No Info and Info) and with exogenously determined decision makers we can conclude that being able to choose the type of the decision maker does not contribute to satisfaction with the decision or its perceived fairness (variable *Choice*).

A.3 Additional tables

References

Bai, B., Dai, H., Zhang, D., Zhang, F., & Hu, H. (2021). The impacts of algorithmic work assignment on fairness perceptions and productivity. *Academy of Management Proceedings, 2021*(1), 12335.

Bartling, B., Fehr, E., & Herz, H. (2014). The intrinsic value of decision rights. *Econometrica, 82*(6), 2005–2039.

Batson, C. D., & Thompson, E. R. (2001). Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science, 10*(2), 54–57.

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

Bock, O., Baetge, I., & Nicklisch, A. (2014). Hroot: Hamburg registration and organization online tool. *European Economic Review, 71*, 117–120.

Boettke, P. J., & Thompson, H. A. (2022). Identity and off-diagonals: How permanent winning coalitions destroy democratic governance. *Public Choice, 191*(3), 483–499.

Bolton, G. E., & Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review, 90*(1), 166–193.

Brams, S. J. (2019). Fair division in dispute resolution. In *The Oxford Handbook of Public Choice.* (Vol. 1). Oxford University Press.

Braun Binder, N. (2018). *Ai and taxation: Risk management in fully automated taxation procedures*. Available at SSRN 3293577.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work (tech. rep.)*. National Bureau of Economic Research.

Buchanan, J., & Tullock, G. (1965). *The calculus of consent: Logical foundations of constitutional democracy* (Vol. 100). University of Michigan press.

Buchanan, J., & Hickman, W. (2023). Do people trust humans more than chatgpt?.

Buchanan, J., Hill, S., & Shapoval, O. (2023). Chatgpt hallucinates non-existent citations: Evidence from economics. *The American Economist, 69*(1), 80–87.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review, 97*(3), 818–827.

Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Responsibility for what? Fairness and individual responsibility. *European Economic Review, 54*(3), 429–441.

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

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review, 99*(1), 431–57.

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

Claure, H., Kim, S., Kizilcec, R. F., & Jung, M. (2023). The social consequences of machine allocation behavior: Fairness, interpersonal perceptions and performance. *Computers in Human Behavior, 146*, 107628.

Corgnet, B. (2023). An experimental test of algorithmic dismissals. In *Working paper 2302, groupe d'analyse et de théorie economique Lyon at-Étienne (GATE Lyon St-Étienne), Université de Lyon*.

Cowgill, B., Dell'Acqua, F., & Matz, S. (2020). The managerial effects of algorithmic fairness activism. *AEA Papers and Proceedings, 110*, 85–90.

Cowgill, B., & Tucker, C. E. (2019). *Economics, fairness and algorithmic bias*. Available at SSRN: https://ssrn.com/abstract=3361280.

Criddle, C. (2023). *AI executives warn its threat to humanity rivals 'pandemics and nuclear war'*. Accessed from 06 Jul 2023.

Dargnies, M.-P., Hakimov, R., & Kübler, D. (2022). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence.

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & Society, 35*, 917–926.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. In *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114.

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155–163.

Epley, N., & Dunning, D. (2000). Feeling" holier than thou": Are self-serving assessments produced by errors in self-or social prediction? *Journal of Personality and Social Psychology, 79*(6), 861.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868.

Fumagalli, E., Rezaei, S., & Salomons, A. (2022). Ok computer: Worker perceptions of algorithmic recruitment. *Research Policy, 51*(2), 104420.

Gallier, C. (2020). Democracy and compliance in public goods games. *European Economic Review, 121*, 103346.

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

Greenberg, J., & Alge, B. J. (1998). Aggressive reactions to workplace injustice.

Guenole, N., & Feinzig, S. (2018). *The business case for AI in HR: With insights and tips on getting started*. IBM Smarter Workforce Institute, IBM Corporation.

Haesevoets, T., Verschuere, B., Van Severen, R., & Roets, A. (2024). How do citizens perceive the use of artificial intelligence in public sector decisions? *Government Information Quarterly, 41*(1), 101906.

Hechter, M. (2013). *Alien rule*. Cambridge University Press.

Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine? *Journal of Experimental Psychology: Applied, 25*(3), 386.

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Hu, K. (2023). *ChatGPT's explosive growth shows first decline in traffic since launch*. Accessed from 06 Jul 2023.

Hülle, S., Liebig, S., & May, M. J. (2018). Measuring attitudes toward distributive justice: The basic social justice orientations scale. *Social Indicators Research, 136*(2), 663–692.

Humm, B. G., Bense, H., Fuchs, M., Gernhardt, B., Hemmje, M., Hoppe, T., Kaupp, L., Lothary, S., Schäfer, K.-U., Thull, B., et al. (2021). Machine intelligence today: Applications, methodology, and technology. *Informatik Spektrum, 44*(2), 104–114.

Klamler, C. (2019). 715Fairness Concepts. *The Oxford Handbook of Public Choice* (Vol. 1). Oxford University Press.

Kolm, S.-C. (1996). Moral public choice. *Public Choice, 87*(1), 117–141.

Konow, J. (1996). A positive theory of economic fairness. *Journal of Economic Behavior & Organization, 31*(1), 13–35.

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature, 41*(4), 1188–1239.

Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., Williams, D., Campbell-Gillingham, L., Thacker, P., Botvinick, M., et al. (2022). Human-centered mechanism design with democratic AI. *Nature Human Behaviour, 6*(10), 1398–1407.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 2053951718756684.

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

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.

Luhan, W. J., Poulsen, O., & Roos, M. W. (2019). Money or morality: Fairness ideals in unstructured bargaining. *Social Choice and Welfare, 53*(4), 655–675.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.

Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration, 42*(12), 1031–1039.

Mellizo, P., Carpenter, J., & Matthews, P. H. (2014). Workplace democracy in the lab. *Industrial Relations Journal, 45*(4), 313–328.

Monin, B., & Merritt, A. (2012). Moral hypocrisy, moral inconsistency, and the struggle for moral integrity. In *Working paper*.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring chatgpt political bias. *Public Choice, 198*(1), 3–23.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167.

Neyer, F. J., Felber, J., & Gebhardt, C. (2012). Entwicklung und validierung einer kurzskala zur erfassung von technikbereitschaft. *Diagnostica*.

Razzolini, L. (2013). Experimental public choice. *The Elgar Companion to Public Choice.* (2nd Ed., pp. 415–426).

Roberts, J., Lüddecke, T., Das, S., Han, K., & Albanie, S. (2023). Gpt4geo: How a language model sees the world's geography. arXiv preprint arXiv:2306.00020.

Sausgruber, R., Sonntag, A., & Tyran, J.-R. (2021). Disincentives from redistribution: Evidence on a dividend of democracy. *European Economic Review, 136*, 103749.

Schram, A. J. H. C. (2008). Experimental public choice. *Readings in public choice and constitutional political economy* (pp. 579–591). Springer.

Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics, 39*(3), 312–320.

Stiroh, K. J. (2001). What drives productivity growth?. *Economic Policy Review*, *7*(1).

Strobel, C. (2019). The hidden costs of automation. In *Working paper*.

Sznycer, D., Lopez Seal, M. F., Sell, A., Lim, J., Porat, R., Shalvi, S., Halperin, E., Cosmides, L., & Tooby, J. (2017). Support for redistribution is shaped by compassion, envy, and self-interest, but not a taste for fairness. *Proceedings of the National Academy of Sciences, 114*(31), 8420–8425. https://doi.org/10.1073/pnas.1703801114

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223*(5), 96–103.

Vallance, C. (2023). *AI could replace equivalent of 300 million jobs - report*. Accessed from 06 Jul 2023.

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior, 90*, 215–222.

Wakslak, C. J., Jost, J. T., Tyler, T. R., & Chen, E. S. (2007). Moral outrage mediates the dampening effect of system justification on support for redistributive social policies. *Psychological Science, 18*(3), 267–274.

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ, 17*, 131.

Waytz, A., & Norton, M. I. (2014). Botsourcing and outsourcing: Robot, british, chinese, and german workers are for thinking – not feeling – jobs. *Emotion, 14*(2), 434.

Weber, M. (1978). *Economy and society: An outline of interpretive sociology* (vol. 1). Univ of California Press.

Wilson, B. J. (2012). Contra private fairness. *American Journal of Economics and Sociology, 71*(2), 407–435.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*, 661–683.