RATIONALITY & COMPETITION
CRC TRR 190

# Ruled by Robots: Preference for Algorithmic Decision Makers and Perceptions of Their Choices

**Marina Chugunova** (Max Planck Institute for Innovation and Competition)
**Wolfgang Luhan** (University of Portsmouth)

Discussion Paper No. 439

October 24, 2023

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

# Ruled by robots: Preference for algorithmic decision makers and perceptions of their choices.*

Marina Chugunova † Wolfgang J. Luhan ‡

March 2022

**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 the affected people. We use a laboratory experiment to study the preference for human or algorithmic decision makers in re-distributive decisions. In particular, we consider whether algorithmic decision maker will be preferred because of its unbiasedness. 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. Yet, despite this preference, the decisions made by humans are regarded more favorably. Participants judge the decisions to be equally fair, but are nonetheless less satisfied with the AI decisions. Subjective ratings of the decisions are mainly driven by own material interests and fairness ideals. For the latter, players display remarkable flexibility: they tolerate any explainable deviation between the actual decision and their ideals, but react very strongly and negatively to redistribution decisions that do not fit any fairness ideals. Our results suggest that even in the realm of moral decisions algorithmic decision-makers might be preferred, but actual performance of the algorithm plays an important role in how the decisions are rated.

Keywords: delegation, algorithm aversion, redistribution, fairness.

JEL-Classification: C91, D31, D81, D9, O33.

*The authors gratefully acknowledge funding by the Nuffield Foundation (FR-000000326) and the Collaborative Research Center TRR 190 Rationality and Competition. We thank the seminar participants at University of Portsmouth and the Max Planck Institute for Innovation and Competition as well as participants at the 2019 European ESA Meeting and 2021 ESA Global Online Around-the-Clock Meetings for helpful comments and suggestions. We thank Nishan Lin for able research assistance.

†Max Planck Institute for Innovation and Competition, Munich, Germany

‡Corresponding author: Faculty of Business and Law, University of Portsmouth, Richmond Building, Portland Street, Portsmouth, Hampshire PO1 3DE, United Kingdom; e-mail: wolfgang.luhan@port.ac.uk

# 1 Introduction

*I would never judge you. I do not belong to any country or religion. I am only out to make your life better.*

–GPT-3, a text-generating algorithm,
In an essay for The Guardian[1]

Algorithms and Artificial Intelligence (AI) have become an integral part of our decision making, not only for personal but also for important professional decisions. Managers increasingly rely on algorithmic aids when determining how to assign a bonus or other performance incentives, who should return to work at the office after the pandemic and when deciding whom to hire and what salary to offer them (Fisher, 2019; Grensing-Pophal, 2021; Riberolles, 2021; Van Esch et al., 2019). While individual companies and managers can decide if they want to rely on digital technology for these decisions, and determine how much weight to give to their suggestions, those affected by the decisions cannot (directly) determine if AI decision supports are used. Yet, they may perceive or react differently to a decision depending on whether it was taken by a human or an algorithmic decision maker. As the number of possible applications of technology grows, which offer clear advantages in terms of operational efficiency (Solow, 1957; Stiroh, 2001), 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 a decision? How will the nature of the decision maker affect the perception of the decision? Will people react to the decisions differently if they come from a human manager as compared to an algorithm? Specifically, we consider these questions in the context of income redistribution.

Redistributive decisions taken on behalf of others represent a wide range of common situations, both in the workplace (how to assign a bonus for a team task or who gets an undesirable task) and in the economic and political context in general, ranging from taxation to social support, unemployment benefits, education policies, monetary policies and many more. These types of decisions are especially interesting for the question whether people would want or accept an AI decision maker. Unlike calculation and prediction tasks, where algorithms are widely employed and accepted (see, e.g., Humm et al., 2021), there are no objectively correct solutions in such scenarios. In this sense, redistributive decisions can be seen as a type of moral decisions, where the definition of correct or fair depends on the observer's personal ideals and beliefs. As a consequence redistributive decisions often spark controversy and lead to societal tensions and conflicts (e.g., Sznycer et al., 2017; Wakslak et al., 2007). Defining which decision maker 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.

The nature of the decision maker might affect the perception of the decision, the acceptance and potentially also the compliance with this decision—independent of the decision itself. In a managerial context this can severely impact the performance of the affected workforce. A good illustration of this link can be found in Bai et al. (2020). The authors conducted a field experiment in an Alibaba warehouse where pick lists (i.e., lists of items the workers need to collect from different shelves in the warehouse) were either distributed by a computer terminal or a human manager. The group that received pick lists from a computer terminal perceived them to be more fair which led to a striking increase in picking efficiency of almost 20%. If this holds true in other domains of decisions made for others, the the utilization of the most widely *accepted* type of decision maker will increase satisfaction with the decision or policy and, by extension, compliance. Apart from considerations of fairness, different performance

[1] https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

expectations might be important. For example, Strobel (2019) examines whether employees exert more effort if the threshold performance for receiving a bonus is set manually on a case by case basis or through an automatic system, and finds that the effort is significantly lower under an automated performance evaluation system. Yet, this result appears to be driven by the fact that employees expect lower performance thresholds in the automated condition and thus, expected to receive a bonus for a lower effort.

The newly emerging, but rapidly growing literature on how people perceive algorithmic decisions and engage with algorithms for certain tasks finds mixed results and provides a series of important findings for our study. First, it appears that the nature of the task matters for the preference to involve an algorithm (Hertz and Wiese, 2019; Lee, 2018; Waytz and Norton, 2014). People seem to be willing to outsource more analytical tasks to an automated agent, but are reluctant to do so with social tasks (Waytz and Norton, 2014). If algorithms are employed in "human tasks", algorithms' perceived lack of intuition and subjective judgment capabilities lead to them being judged as less fair and trustworthy (Lee, 2018) or reductionist (Newman et al., 2020).

Moral decisions received particular attention by Gogoll and Uhl (2018) and Bigman and Gray (2018). Gogoll and Uhl (2018) found that in moral decisions–those affecting third parties–people not only preferred the human decision-maker, but even punished others who chose the algorithm. The authors attribute this to a general aversion to automated decisions in the moral domain. This is corroborated by Bigman and Gray (2018), who find that this aversion holds irrespective of whether the decisions made are favorable for the affected parties. The *per se* aversion towards the algorithms in moral domains can be explained, for example, by the deeply ingrained belief that "human is better" (Eastwood et al., 2012), and that algorithms are dehumanizing and unethical in nature (Dawes et al., 1989). In a recent study, Hidalgo et al. (2021) presents a series of vignette studies, documenting that people judge moral actions by machines differently than identical moral actions by humans. The authors highlight the role of intentions in how decisions are perceived.

Yet, despite these negative attitudes towards algorithms in the moral domain, algorithmic decisions also have a "halo" of perceived scientific authority and objectivity (Cowgill, Dell'Acqua, et al., 2020). Suggestions coming from automated expert systems are seen as more objective and rational than identical suggestions from a human advisor (Dijkstra et al., 1998), and people tend to react less emotionally and less negatively to unfair decisions made by automated systems (Sanfey et al., 2003; Shank, 2012). The perceived fairness of the automated decisions may be additionally driven by the increased procedural fairness associated with the use of algorithms, as they act following "calculable rules" and decide "without regard for persons" (Weber, 1978, p.975 on benefits of bureaucracy).

Our study contributes to this literature by considering if people prefer a human or an algorithm to make redistributive decisions on their behalf, and how decisions made by different decision makers are perceived. 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 and Sele, 2020), as our main focus is not on people who have discretion to use or not use algorithmic aids, but on those who are affected by these decisions.

It is a priori not clear if the application of algorithms in redistributive decisions would increase or decrease perceived fairness and which decision maker would be preferred. While humans can be moral agents, they can also apply different fairness frameworks to different outcomes for their own benefit. Equipped with different moral frameworks, people can always argue that the decision that benefits themselves (or their group) has the moral high ground (Batson and Thompson, 2001; Epley and Dunning, 2000; Monin and Merritt, 2012). That is, humans can arguably better apply

the ambiguous rules of moral decision making, but algorithms can coherently and selflessly stick to a programmed set of rules and thus score higher on procedural justice that may affect how people perceive the decisions (Hechter, 2013). In one sentence, AI can not change its decision at will and therefore its decisions are always (in this sense) unbiased, while humans might discriminate in somebody's favor. If people are aware of this, they might prefer an algorithm as the decision maker, even in the context of a moral, redistributive decision.

As empirical investigations of these questions require data which cannot be readily found in administrative or company records, we implement an online experiment where a decision maker can redistribute earnings from three real effort tasks between two players. The immediate analogy would be individual team members who all provided an effort for a project or a solution. 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 process a team manager will decide on who gets which share of the bonus. In our experiment, participants can choose if the decision maker is an algorithm or a human and subsequently express their perceptions of how fair the decision is and how satisfied they are with the outcome. We choose three specific tasks that allow a range of "fair" distributions, depending on the fairness principle applied. This reflects the ambiguity of the right decisions in everyday working environments, 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 decision maker.

We find an overwhelming preference for an algorithmic decision maker. Regardless of the potential bias of the human decision maker, more than 63% of participants prefer the algorithmic decision maker across treatments. Participants are less likely to choose AI if they have earned more than their opponent from effort or talent tasks, but it does not seem that the preference for the decision maker is driven by expected performance differences: the choice of the decision maker is not determined by he participant's own fairness ideals. Even though the majority of participants choose the AI decision maker, 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 decision maker. As expected, losing tokens severely impacts the satisfaction with the decision and the perceived fairness. The strongest reaction, however comes from a decision that is perceived as unfair because it does not follow a consistent fairness principle (e.g., egalitarian, meritocratic etc.). This can, in part, explain the lower satisfaction with the AI decisions as these—due to technical simplicity of the employed algorithm—are more likely to be inconsistent with one principle.

We conclude that in contrast to some of the previous findings in the literature, people do not dislike algorithms in moral tasks per se. They actually prefer the AI decision maker, which is not due to a fear of discrimination but appears to be a *general preference*. The AI's actual decisions, however, are rated as inferior. To "live up to the expectations" and increase the acceptance of these AI decisions, the algorithm has to consistently and coherently apply fairness principles.

# 2 Theory and Hypotheses

Consider a situation where two people have each individually generated an income, which is then pooled, and a third party will decide how this pooled endowment is distributed. The primary question we ask is whether people will prefer a human or an algorithm to make this decision when it affects them.

We have discussed before that the literature finds opposing results on whether an algorithm or a human is the preferred decision maker *in general.* The overarching theme, however, appears to be a general "algorithm aversion" (see, e.g., the reviews by Burton et al., 2020; Chugunova and Sele, 2020) which means that a human decision maker should generally be preferred. Both of these papers also stress the importance of the decision context. Our decision context is inherently a moral one where the decisions will mostly be driven by fairness principles and beliefs. According to the literature, people should have a particularly strong aversion to algorithmic decision makers in this context (Bigman and Gray, 2018; Gogoll and Uhl, 2018). One reason why an algorithm might be preferred in such a situation, however, is if the impartiality of a human decision maker is in question. If a human is perceived to be biased or if the situation could lead to a potentially biased human decision, the objectivity of an algorithm might be preferred (Cowgill, Dell'Acqua, et al., 2020). We will test this assumption by systematically varying whether there is room for potential discrimination or not between treatments, which allows us to observe whether the mere *possibility* of discrimination affects preferences for the type of a decision maker. Simply put, the human has the potential for discrimination, the algorithm has not.[2]

*$H_1$: If there is no scope for bias (i.e., a decision under the veil of ignorance), a human-decision maker will be preferred over an automated one.*

*$H_2$: If there is scope for bias but the direction of bias is unpredictable, an automated decision maker will be preferred.*

A decision bias can have two main forms, positive and negative. While a negative bias is what is colloquially referred to as discrimination, from which people wish to avoid the negative consequences, a positive bias would be preferential treatment and have positive (in the context of our study, monetary) effects and might be sought-after. In the following we refer to these biases as negative and positive discrimination respectively. In the restricted context of our controlled experiment, we define negative discrimination as a reduction of earnings due to the revealed features of the affected person—specifically the choice of a painting (see the next section for details). Positive discrimination is similarly an increase in earnings due to the revealed features[3]. If we only consider the potential monetary benefits of positive discrimination, we would expect that this will lead to a preference for the human decision maker over the unbiased algorithm:

*$H_{3a}$: Expected positive discrimination will increase the choice of a human decision maker as compared to a situation without discrimination.*

[2]It is true that algorithmic decisions—whether supporting human decisions or acting completely autonomously — are based on the parameters set by humans and data stemming from human decisions. Therefore algorithms have the potential to make biased decisions and even distill and amplify existing discriminatory patterns (e.g., Arnold et al., 2021; Cowgill and Tucker, 2019). As it appears to be a challenging, but implementation issue, in the following we assume that the algorithm is unbiased and represents a generally accepted fairness norm. See the experimental instructions Appendix for the exact wording used to describe algorithm and human decision makers.

[3]These re distributions of earnings could be in line with a fairness ideal, but this would be different from the true fairness ideal the decision maker holds. Effectively, the decision bias could manifest itself as a change of fairness ideals, which would make it easier to justify the negative or positive discrimination. For our purposes the mere change of earnings is sufficient as our focus lies on the effect of expectations about these biases

This view, however, neglects any form of social preferences and implies that people solely care about own outcomes. All outcome based fairness preference models (e.g., Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) describe a trade off between preferences for the individual payoff and the fairness of the distribution among all parties. If we assume that the individual payoff preferences outweigh the fairness preferences, hypothesis 3a still holds. If a person has stronger fairness preferences than monetary preferences, we should find that they would prefer the fair outcome over a potential positive discrimination, and hence would prefer the algorithm provided they believe the algorithm is more fair than the human.

*$H_{3b}$: Expected positive discrimination will decrease the choice of human decision makers as compared to a situation without discrimination.*

Expected discrimination should have a straightforward impact on the preference for the algorithm over a human decision maker. Firstly, the expected payoffs are decreased when there is a risk of discrimination. Secondly, if the distribution of payoffs is already equal, or the person in question starts with lower earnings than the other participant, they cannot expect either an increase in payoff or a fairness improvement from a human decision maker who they expect to be negatively biased towards them—but they might expect this from the algorithm. If the person has higher earnings before the redistribution, they would still prefer the AI to redistribute, expecting a fair(er) end result.

*$H_{3c}$: Expected discrimination will decrease the choice of human decision makers as compared to no discrimination.*

As an additional test for the validity of these effects (3a, 3b, and 3c), we expect to not observe any significant change between a situation where there is no discrimination possible, and a situation where discrimination is possible, but not applicable in a particular situation. An example might clarify this: Assume there are two groups of people, $A$ and $B$ and the group affiliation is the only known identification. If the decision maker is a member of group $A$ and the two people whose money this person is distributing are from different groups, one from $A$ and one from $B$, the person from group $A$ might expect positive discrimination, and the person from group $B$ negative discrimination. If however, all three are from the same group, or the decision maker is from group $A$ but the others are both from group $B$, no discrimination is possible. If no discrimination is expected in such a way, the distribution of choices should be the same as in a situation where there are no groups or these are not know.

Moving on to our expectations regarding how satisfied people will be with the decision ex post, we start with the obvious: money will, ceteris paribus, make people happy.

*$H_4$: The satisfaction with a decision increases with the allocated payoffs, irrespective of the decision maker.*

When considering how people judge the decision of algorithms as compared to those of humans the literature is inconclusive. On the one hand, Sanfey et al., 2003 find people react less negatively if algorithms make "unfair" decisions (a similar result can be found in Leyer and Schneider, 2019). Moreover, algorithms appear to enjoy the perceived "halo" of scientific authority and objectivity (Cowgill, Dell'Acqua, et al., 2020) and their decisions may be regarded as more fair (Bai et al., 2020). On the other hand, literature documents strong aversion to the use of algorithms in the moral domain (Bigman and Gray, 2018; Gogoll and Uhl, 2018), suggesting that perceptions of algorithms may be context dependent. Newman et al. (2020) find that AI decisions for promotion and performance evaluations were considered reductionist and Longoni et al. (2019) suggest that algorithms are viewed as unable to take into account unique features of individuals in medical recommendation settings. Considering the question

of how humans judge machines in a series of ethically relevant situations, Hidalgo et al. (2021) also find in a series of vignette studies that identical actions by humans and machines would be judged differently. Yet, while the literature does not agree on the direction of the effect, it agrees that the nature of the decision maker matters for how decisions are perceived and acted upon, so our hypothesis is non-directional.

*$H_5$: The nature of the decision maker affects the perceptions of fairness and satisfaction with the decision.*

Whether a decision is unfair and discriminatory might be very subjective. Individual fairness principles may even shift from before to after income is earned (Luhan et al., 2019). We do expect however, that the possibility of a biased decision will, on average, reduce the satisfaction with this decision. This relies on the concept of procedural justice: if the process is fair then any outcome that resulted from the fair process can be considered fair (Rawls, 1971).

*$H_6$: Irrespective of the actual decision made, the possibility of discrimination reduces satisfaction.*

Finally, 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.

*$H_7$: Irrespective of the actual decision made and the nature of the decision maker, having a choice of the decision maker increases the satisfaction with the decision.*

# 3 Design and Procedure

The main aim of the design was to create a situation where we can observe participants' preference for either a human or an algorithmic decision maker to redistribute income that they had previously generated. We incorporated the possibility of discrimination to consider whether this would increase the preference for the algorithm as an unbiased decision maker. In addition we wanted to measure, ceteris paribus, the satisfaction and the perceived fairness of the decision, depending on the decision maker, the perceived discrimination and whether the participant had chosen the decision maker in charge.

**Income Generation** To start, experimental participants individually earned their initial income by completing three tasks. In each task participants earned tokens of different colors. The three tasks mimic three potential determinants of income central to major fairness theories: luck, effort and talent (Konow, 2003).

In the luck task, participants could earn 100 green tokens if the coin virtually tossed by the computer shows heads. In the effort task, participants were given 15 seconds to count the zeros in two matrices of zeros and ones for 100 yellow tokens each. In the talent task, participants earned 100 blue tokens for solving a matrix from the Ravin fluid intelligence test correctly. In the description of effort and talent tasks, 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 each color could vary in the exchange rate from 1 to 6 cents per token. This design feature offers two benefits. First, the separately colored tokens allow us to clearly distinguish the fairness principle behind any distributive 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 colors as equally important and not focus on single income elements or just the total number of tokens.

The distinction of earnings from effort, talent, and luck allows us to differentiate between four distinct fairness principles and related distributions of earnings (see, e.g., Konow, 2003; Luhan et al., 2019): egalitarian, choice egalitarian, meritocratic, and libertarian. While these fairness principles are not the focus of our study, the existence of an array of potentially fair behaviors and re-distributions enables decision makers to discriminate against one participant while still making a *fair* decision. This should make it more apparent that discrimination could potentially happen, as decision makers could hide behind fairness principles. It also allows us to observe whether the discrepancy between own fairness ideals and those of the decision maker can affect satisfaction with the decision.

**Choice of a Decision Maker** In order to test our $H_1$ on the general preference for a human decision maker 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 individually pick a human or an algorithmic decision maker. In case of a unanimous choice of one decision maker, it would be implemented, in case of disagreement the decision of one participant would be chosen with equal probability.

The human decision maker (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. Decision makers received no other information about the two participants other than their income portfolios, or given any instructions on how to decide other than to "make a fair decision". The actions of the decision makers were not incentivized: they received a flat payment regardless of their choices.

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 satnav calculates the routes, but are not able or interested in fully understanding the mechanics of the algorithmic process. We therefore—truthfully—informed participants that the algorithm would choose a *"fair distribution based on data from a survey of several hundred participants. The participants of the survey 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"*. This description clearly states that the data used by the algorithm is not historic, was specifically tailored to the tasks the participants faced, and that the decision involved some transformation of the data.

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, all fluent in English. The survey participants were asked to determine a fair redistri-

[4]The Ravin test measures *fluid intelligence* that is considered to be innate. Several studies find that training helps to improve the score in this task (Bors and Vigneau, 2003; Hayes et al., 2015), but participants had only one task to solve and no possibility to practice during the experiment. Even if participants happened to have trained for such a task at some point in their life, this can be considered as pre-experimental talent. Since the task was performed in an online setting, we timed the task such that an online search for the answer was not possible.

bution of tokens for hypothetical pairs of players. They saw the same tasks as in the subsequent experiment with identical explanations. The series of questions 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 decision maker. 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, after which 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, with 76,48% probability the algorithmic decision maker would not redistribute the points, with 21.94% would split them equally among the two participants and with 0.79% probability it would redistribute all the points in favor of the participant who had zero points.

To simplify the design and further interpretation, we did not allow for continuous redistribution for either type of the decision maker: e.g., the decision makers could not transfer 1 token out of 100 to another player. The decision maker could redistribute the tokens of a certain color evenly, give them all to one of the players or keep unchanged.[5]

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

**Negative and Positive Discrimination** To test the role of bias as formulated in our hypotheses $H_2$, $H_{3a}$, $H_{3b}$ and $H_{3c}$, we introduced a source of potential discrimination for the human decision maker. Our aim was to keep this source of discrimination free from the possible confounding effects of real-world biases and use a purely lab-induced feature. We implemented a well established procedure to create minimal groups as introduced by Tajfel (1970). At the beginning of the experiment, all participants (including the human decision makers) were shown two paintings, one by Paul Klee and one by Wassily Kandinsky, and were asked to select which 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 might not be very strong, but in the absence of any other information can lead to discriminatory behavior (ibid.). The decision maker might favor his or her in-group due to, for instance, homophily (Y. Chen and S. X. Li, 2009; McPherson et al., 2001) and providing information allows for favoritism of this type—allowing for positive discrimination in the redistribution of income. By design, we do not incentivize any sort of discrimination, as the payment of the decision maker is independent of the decision, and therefore we test the lower bound of the effect. Even if the decision maker does not actually favor the members of the in-group, the introduction of the group information allows for discrimination and therefore may affect the choice of the decision maker.

**Experimental Treatments** Our first experimental treatment varies if the group information is revealed. In all treatments all participants choose a painting. In the first treatment (Tr1) no further mention of this was made in the experiment and this choice was not revealed to anybody. In Tr2 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) decision maker and knew that the decision maker would have the same information.

[5]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 we included an additional options to enable the decision makers to redistribute in steps of 100 tokens.

Table 1: Features of treatments and sample information

| Treatment Name | Treatment features | | Info on Sample | | |
|---|---|---|---|---|---|
| | Info | Nature of DM | # Sessions | # Regular Participants | # DM |
| Tr1: Choice_NoInfo | no | Choice: Human or AI | 4 | 70 | 8 |
| Tr2: Choice_Info | yes | Choice: Human or AI | 6 | 102 | 12 |
| Tr3: AI | no | AI | 3 | 66 | - |
| Tr4: Human_NoInfo | no | Human | 2 | 34 | 4 |
| Tr5: Human_Info | yes | Human | 2 | 34 | 4 |

Notes: **info** indicates whether the choice of the picture of all parties was revealed, **Nature of DM** indicates whether the decisions were taken by a human decision maker, an algorithm or whether there was a choice between the two; and the last three columns contain the number of **sessions** per treatment, the number of *regular* **participants** who earn points and choose a decision maker, and the number of **decision makers** in each session.

In addition to the question which decision maker is preferred, and whether potential discrimination alters this preference, we also study whether the participants are satisfied with the redistributive decision, and how this is influenced by their change of earnings ($H_4$), the nature of the decision maker ($H_5$),the perceived fairness of the decision ($H_6$), the discrimination ($H_7$), and the influence of having a choice ($H_8$).

We therefore introduced three more treatments to test these hypotheses and to control for possible interaction effects: in Treatment Tr3 the decision maker was always an algorithm and in Tr4 always a human. To consider the interaction effect of endogenous choice of the decision maker and presence or absence of the group information we additionally vary if the information is revealed in treatments with exogenous imposed decision makers.[6] In all treatments players could indicate on two seven point Likert scales how happy they were with the redistribution and if they considered it fair. When answering it, they saw the distribution of tokens after the redistribution, the initial distribution of tokens within the pair and the nature of the decision maker.

Table 1 summarizes our five treatments varying three parameters: (1) if participants could choose the nature of the decision maker, (2) if the group information is revealed and (3) the nature of the decision maker.

**Timeline of the Experiment** Fig. 1 provides an overview of the timeline in all treatments. First, all participants chose their preferred painting. Following this, participants were randomly assigned to be *regular* participants or *decision makers.*[7] The regular participants received instructions for the tasks in the *income generation stage* and performed them (a coin toss, matrices with zeros and a Ravin matrix). The decision makers received the same instructions with an explanation that only the regular participants would perform the tasks. In the following *redistribution stage*, players were matched into pairs and the treatment variation was introduced, either giving players a choice of the decision maker (Tr1 and Tr2) or informing them about the nature of the decision maker (Tr3, Tr4, Tr5). In the Info treatments (Tr2 and Tr5), in addition to the information on tokens earned by each player in the pair, the information on the painting choice was revealed. The treatments were implemented

[6] We omitted a treatment of only AI decisions *with* the information of the painting choice, since the algorithm by design could not include this information.

[7] In the experimental instructions they were called Type P and Type D to avoid any framing effects. The translation of experimental instructions can be found in the appendix B.

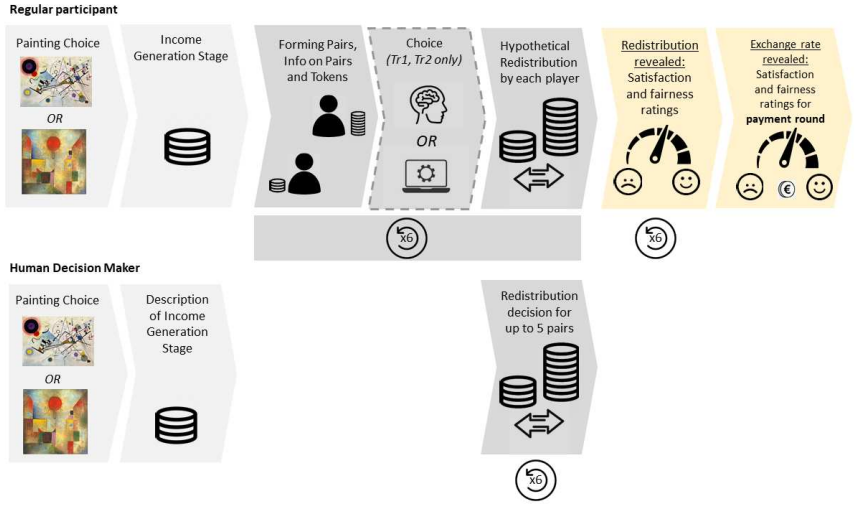



Figure 1: Sequence of events in all treatments for regular participants and human decision makers.

between-subjects and each participant faced one treatment only. This redistribution stage consisted of six periods, effectively six repetitions with different 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. They were asked to make a hypothetical decision on what distribution they would think was fair for their pair. The decision makers learned the token portfolio of the pair and could separately decide for each color token if it should be redistributed. The decision makers as well as players were not aware of the value of each token at this point. After the *redistribution stage*, regular players were shown one-by-one all six pairings in the individual periods and learned what redistribution decision was made for each period/pairing. Participants were informed (Tr1 and Tr2) or were reminded (Tr3, Tr4, and Tr5) of the nature of the decision maker, the painting choices of all involved parties (Tr2 and Tr5) and the outcome of the distribution. 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 period and the Euro value of each color of the token was revealed. Based on this information, participants were informed about how much they earned in the experiment. After learning the exchange rate for each color of the tokens, players were asked again how happy they were with the decision in the payment relevant round and how fair it was.

After the experiment was completed, players filled out the questionnaire including basic demographic characteristics, self-evaluations of trust, risk and political attitudes. Additionally, we included 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.These measures were selected as they might capture important sources of heterogeneity in evaluating the decisions, the perceived fairness and the preference for human or AI decision makers (Parker and Grote, 2020, see e.g., ).

**Procedures** The experiment was implemented using oTree (D. L. 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). None of the participants took part in the experiment more than once. We conducted 17 online sessions of around 22 participants each. In total, 362 participants took part in the experiment. The sessions were gender-balanced and the average age of participants was 25. The average payment was 9.10 Euro for 45 minutes.

Due to our focus on the preference for and choice of the type of decision and how possible discrimination affects this choice, we conducted an unequal number of sessions per treatment (see Table 1). In each treatment—apart from Tr3 where the decision was taken only by the algorithm— we randomly allocated two human decision makers per session, each deciding for several pairs of regular participants. The decision makers received a flat payment of 10 Euro regardless of the decisions.

# 4 Results

We will follow the order of our hypotheses and start with the question which decision maker was preferred, i.e. chosen most frequently, before analyzing what determines whether decisions were perceived as fair and how happy participants were with the decisions taken.

## 4.1 Choice of the Decision Maker

Table 2 contains the absolute and relative frequencies of decision maker choices from Tr1 and Tr2 along with the p-values from non-parametric inference tests. We find an overall preference for the AI decision maker. In the absence of information on the group membership (the chosen painting), the algorithm is preferred in 63.25% of all choices in Tr1. We reject our first hypothesis that under the veil of ignorance the human decision maker is preferred. We find quite the contrary, that the AI is chosen significantly more frequently than 50% (two-sided binomial test $p < 0.001$).

**Result 1.** We find that the AI decision maker is preferred over the human decision maker, in the absence of potential discrimination.

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 AI in Tr2. This does not allow for a clear test of our second hypoteses that the general prospect of a biased decision by the human decision maker leads to a preference for the AI. The majority of participants chooses the AI, but with no significant difference in the preferred decision maker between Tr1 without and Tr2 with the possibility for discrimination ($\chi^2$ $p = 0.824$), it appears that the general preference for AI decision maker is rather *prevails* in Tr2 (two-sided binominal test, $p < 0.001$). It is not the prospect of discrimination that drives this overall preferences for the AI decision maker and we reject $H_2$.

As in our setting potential positive discrimination for one member of the pair means potential negative discrimination for the other, the aggregate result of no effect of potential discrimination could be due to the fact that choices under positive discrimination ($H_{3a,b}$) are balanced out by choices under negative discrimination ($H_{3c}$). To consider this, we 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. However, within discrimination types, we again do not find any impact of potential discrimination on the choice of the decision maker. In all three cases, we observe a strong preference for the AI as a decision maker, 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 $\chi^2$Tr1 in table 2). Irrespective of potential positive or negative discrimination, the majority

of choices are for the AI decision maker and we reject our hypotheses $H_{3a}$, $H_{3b}$, and $H_{3c}$. 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 $p < 0.7784$).

**Result 2.** Neither the potential for discrimination, nor the direction of discrimination affect the preference for the algorithm.

Table 2: Chosen decision maker

| | AI | Human | Total | $\chi^2$Tr2 | $\chi^2$Tr1 | 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 Tr2 (Info) | 437 | 247 | 684 | p = 0.954 | p = 0.824 | p < 0.001 |
| (%) | (63.89) | (36.11) | (100.00) | | | |
| Total Tr1 (No Info) | 296 | 172 | 468 | | | p < 0.001 |
| (%) | (63.25) | (36.37) | (100.00) | | | |

Notes: Frequencies of choices in treatments Tr2 (AI or human decision maker *with* the group info) and Tr1 (AI or human decision maker *without* the group info). Percentages in parentheses below absolute numbers of observations. Column $\chi^2$Tr2 displays the p-value for the test of differences between the discrimination classes within Tr2. Column $\chi^2$Tr1 contains the p-values from the individual tests of the observations in the respective row against the observations in Tr1. 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).

As a next step we use a multivariate analysis to look for the determinants of the choice of decision maker. Table 3 contains the results from pooled probit regressions with robust standard errors clustered on the individual level. The dependent variable is the probability of choosing a human decision maker. We link this choice to all available information at the time when participants make the choice of the decision maker: the tokens that both participants earned from the three tasks; whether the information on the painting choices was available and the resulting possibility of discrimination; implications from the various fairness principles; and a small range of background variables form the questionnaire. We find no significant impact of the availability of the group information on the choice of the decision maker (variable *Info*). It suggests that both based on non-parametric and regression analysis, we can reject $H_2$. We tested several specifications of perceived discrimination, none had a significant impact on the choice of the decision maker. In column 1, we report the result from a specification that uses three categorical dummies for types of discrimination: no discrimination (either no info provided or all participants had chosen the same picture) which serves as the base category, the possibility of positive discrimination and the possibility of negative discrimination. We also consider if people may be more or less likely to prefer human or AI decision makers depending on the differences in the earnings (i.e., token portfolios) between themselves and the paired player. In three regression specifications we use different approaches to capture differences in token earnings between the players. In column 1, we include the tokens earnings from all three tasks by the participant and their partner as individual variables. We find that only the earnings from the task requiring effort and talent have a significant impact—earning more in these tasks increases the likelihood of choosing a human decision maker. Looking at the partner's

tokens, we find a significant but much smaller effect of the income from the effort task. In column 2, we use an alternative specification, calculating the absolute distance between the two participants' earnings. As none of these distances have a significant effect on the probability of choosing a human decision maker, this does not seem to reflect how participants considered the earnings when choosing the decision maker. In column 3, we replace the distance between the tokens with a simpler approach. 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 luck tasks on the choice of the decision maker.[8] If participants had earned more in these tasks than their partners they were more likely to choose a human decision maker.

**Result 3.** Having earned more in the effort or talent tasks increases the likelihood of choosing a human decision maker. Earning *more than the other participant* in these tasks has an even stronger impact on the choice of a human decision maker.

This result could be due to the view that a human will have a higher appreciation of what was required to get these tokens and therefore will not redistribute these. Generally speaking, this would mean that participants believe that a human decision maker would hold a fairness principle that is more favorable to their (higher) earnings. We consider this in the last two columns of table 3. In column 4, we determined whether the participant would lose tokens (these would be redistributed to the other participant) if the decision maker held one of four fairness ideals (see section 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 [9] and is a simple representation of how likely it is that a fair decision maker will redistribute money away from the participant. We find that even this simple specification does not yield a significant impact on the choice of the decision maker, and we can conclude that participants consider fairness principles only in a very limited way.

**Result 4.** Possible fairness ideals of the decision maker and the resulting redistribution of tokens had no apparent impact on the choice of the decision maker.

In addition to these factors contributing to the choice of the decision maker, we control for the participants' age, sex, whether they are classified as technology ready, whether they are trusting, and two opinion questions on fair and unbiased decision making from our post experimental questionnaire [10] Of these six, only the two opinion questions had a significant impact on the choice of the decision maker. *Fair-Just* asked for a rating of who is better in making fair and just decisions, the AI or humans. As expected, the higher participants rated this ability for humans, the more likely they

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

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

[10] In the post questionnaire, participants answered three questions on general trust, six questions from the technical readiness scale by Neyer et al. (2012) and a 14-question battery concerning decision making abilities of humans and AI — all on 5 point Likert scales. From the responses to the individual questions we generate two factor variables and two opinion scales that were used in the regressions. Other variables elicited in the post experimental questionnaire did not contribute to the explanatory power of the model and we therefore do not report these.

were to opt for a human decision maker. *Unbiased* recorded whether people believed that it was hard for humans to make unbiased decisions. Unsurprisingly, the more participants believed taht this would be easiy for humans, the more they picked the human decision maker[11].

## 4.2 Satisfaction with the Decision and Perceived Fairness

To consider if people may react differently to a decision if it comes from a human as compared to an algorithm we need to disentangle potential differences in the decisions of different types of decision makers and the effect of the nature of the decision maker. Unsurprisingly, human decision makers and the algorithm make different decisions (see Appendix A.1 for more details). While the difference in performance cannot affect the choice of the decision makers in Tr1 and Tr2 ex ante, it is likely to affect perceptions of fairness and satisfaction. People reported their satisfaction and their rating of how fair a particular redistribution decision was on two separate, 7-point Likert scales.

To consider what factors affect the fairness and satisfaction rankings of the participants, we run several specifications of a pooled OLS regression with standard errors clustered at the individual level. Pooled OLS allows us to utilize the data from all the treatments and see if some invariant characteristics such as age or gender of the participant or treatment features (availability of choice or information of the picture choice) may affect the fairness and satisfaction ratings. See the results of the pooled OLS in Table 4.

In both regressions we controlled for several parameters of the decision situation. For each type of tokens, we consider if it increased or decreased after the redistribution as compared to the initial earnings (variable *Before-After*), if the overlap between own fairness ideals and those of the decision maker matters (variable *Hyp-Actual*) and, in line with several fairness theories (see, e.g., Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), the difference in tokens between the players after the redistribution (*Own-Partner (After)*). Additionally, we introduce dummy variables that capture whether the decision maker was human, if the player lost tokens in total, and what type of discrimination (positive/negative or none) is the player facing in the round. Importantly, we add a dummy variable for *unfair* redistribution (DV:Unfair). This dummy captures whether the implemented redistribution does not correspond to any of the major fairness principles and therefore may be regarded as inconsistent. When submitting their rankings, participants could leave additional comments in an open text field. From these comments, we can see that the participants are well aware of different fairness ideals and are ready to tolerate them even if they do not coincide with their own view. For instance, participants had no objections if no redistribution at all or egalitarian redistribution (i.e. all points split equally) were implemented, even if they themselves would have redistributed points differently. Yet, it appears that players were unhappy in cases where the decision mixed several justice principles. If, for example, points earned by talent or effort were redistributed but not those earned by luck, or if effort points were redistributed, but not the ones earned by talent.[12] From the comments, it appears that as long as one fairness principle was followed—even if participants held a differing view or ascribed to a different principle—this did

[11] **Although these variables seemingly capture very related concepts; the correlation between them is ratehr low and insignificant.**.

[12] Two representative comments: "Both players get the same amount. Even if it is not a complete coincidence how the points were distributed initially, I find it fair. But I would also have understood if the "better" player would have gotten more."

"There was no redistribution of the points determined randomly by coin toss. The points based on skill and concentration, on the other hand, were partially awarded to me without compensation. Even though I benefit from this, I do not feel this distribution is fair to Player B." As not all the participants left comments they are not suitable for any systematic text analysis.

Table 3: Determinants: choice of decision maker. Pooled probit regression.

| Dep.: *Choice Human DM* VARIABLES | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Info | 0.0786 | 0.0201 | 0.00881 | 0.00598 | 0.00494 |
| | (0.155) | (0.158) | (0.157) | (0.159) | (0.158) |
| Pos. Disc. | -0.0230 | -0.0248 | -0.0197 | -0.00941 | -0.0177 |
| | (0.144) | (0.149) | (0.149) | (0.149) | (0.149) |
| Neg. Disc. | -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) | | | | |
| Tok.Luck Partner | 0.00106 | | | | |
| | (0.000850) | | | | |
| Tok.Eff. Partner | 0.00115** | | | | |
| | (0.000520) | | | | |
| Tok.Tal. Partner | -0.000232 | | | | |
| | (0.000933) | | | | |
| Dist. Luck | | -0.000773 | | | |
| | | (0.000710) | | | |
| Dist. Effort | | 0.000727 | | | |
| | | (0.000491) | | | |
| Dist. 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 Egal. | | | | 0.336 | |
| | | | | (0.319) | |
| Lose Choice | | | | 0.216 | |
| | | | | (0.225) | |
| Lose Mer. | | | | -0.178 | |
| | | | | (0.136) | |
| Count Lose | | | | | 0.0486 |
| | | | | | (0.159) |
| Age | -0.00112 | -0.00543 | -0.00512 | -0.00833 | -0.00512 |
| | (0.0130) | (0.0134) | (0.0133) | (0.0133) | (0.0132) |
| Female | -0.0801 | -0.0653 | -0.0798 | -0.0667 | -0.0792 |
| | (0.131) | (0.133) | (0.133) | (0.132) | (0.133) |
| Tec. ready | -0.0771 | -0.0780 | -0.0815 | -0.0751 | -0.0807 |
| | (0.0801) | (0.0803) | (0.0809) | (0.0796) | (0.0811) |
| Trust | 0.108 | 0.110 | 0.118 | 0.119 | 0.119 |
| | (0.0792) | (0.0780) | (0.0782) | (0.0792) | (0.0783) |
| Fair-Just | 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) |
| Constant | -0.974** | -0.296 | -0.580 | -0.616 | -0.643 |
| | (0.422) | (0.374) | (0.422) | (0.491) | (0.479) |
| Observations | 1,152 | 1,152 | 1,152 | 1,152 | 1,152 |

Observations from Tr1 and Tr2, Robust standard errors, clustered at the individual level, in parentheses

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

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

| VARIABLES | (1) Satisfaction | (2) Fairness | (3) Satisfaction | (4) Fairness |
|---|---|---|---|---|
| DV:Unfair | -0.756*** | -1.375*** | -0.767*** | -1.370*** |
| | (0.140) | (0.152) | (0.141) | (0.154) |
| DV:DM Human | 0.174* | 0.124 | 0.201* | 0.109 |
| | (0.0955) | (0.102) | (0.105) | (0.112) |
| DV:Lost Tokens | -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) |
| DV:Choice | 0.0413 | -0.0237 | 0.0404 | -0.0232 |
| | (0.120) | (0.129) | (0.119) | (0.128) |
| DV:Info | -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) |
| DV:Female | -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 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

not lead to unhappiness or perceived unfairness of the decision. As probabilities for the decisions of the algorithm were drawn independently per task category, it mechanically followed that the algorithm ended up being less consistent with the applied principles: 12% of AI decisions were inconsistent, i.e., not following one principle, as compared to only 5% of human ones (t-test, $p < 0.001$). In total 9.5% of all redistribution decisions were classified as inconsistent (DV:Unfair equals to 1).

In the pooled OLS specification, we additionally control for age, gender, level of trust and if the treatment featured choice option or information. Tr1 and Tr2 allow for a fixed effect OLS specification to consider within-subject variation for participants exposed to both types of decision makers. We first report the results of the pooled OLS and then discuss additional insights that stem from fixed effect OLS estimation.The results are discussed below and are consistent regardless of the approach.

We consider fairness and satisfaction rankings separately. While the two are highly correlated (0.77, $p < 0.001$), they are not identical, which explains why regression results vary slightly. Many participants, however, differentiated between the two concepts, noting that they got more money and therefore they are more satisfied although they find the decision unfair[13].

We start by discussing results of the pooled OLS specification (Table 4). By far the largest in magnitude is the coefficient of *DV: Unfair*, that captures if the redistribution decision inconsistently combines several fairness principles. When the redistribution is inconsistent, the satisfaction with the decision is reduced by 0.76 points and the perceived fairness by 1.38 points which, for a 7-point scale, correspond to appriximately 10% and 20% decrease respectively. We additionally consider if reactions to inconsistent decisions depends on the nature of the decision maker, yet the respective interaction term is not significant as reported in table 7. Comparing the determinants for fairness and satisfaction ratings, we observe some "flexibility" in the notion of fairness in our participant sample—again in line with the comments from the open comment fields. We see that a deviation from the participant's own fairness ideals did not have a negative impact on their fairness rating but slightly affected satisfaction with the decision for luck and talent tasks (*Hyp-Actual*).

**Result 5.** Participants accept fairness principles differing from their own as fair. The main impact for satisfaction and perceived fairness overall is whether fairness principles are applied consistently.

The factor with the second largest impact on the perceived satisfaction and fairness of a decision is a loss of tokens (*DV:Lost Tokens*). The dummy variable takes a value of one if the total number of tokens (regardless of their color) is smaller after the redistribution than before. When we split this up into the *changes in token holdings in individual colors*, we find that changes in income from effort and talent are seen as more important than those from luck (*Before-After*), which is in line the the literature (e.g., Luhan et al., 2019). These token changes do not significantly impact the perceived fairness of the decision, however. This might be due to the fact that the mere change could be increases as well as decreases, so the fairness effect cannot be determined in general. Ending up with more tokens of any color than the opponent also improves satisfaction and perceived fairness, confirming a somewhat self-serving bias in terms of fairness, where participants appear to accept having more than their counterpart as being fair. We cannot reject our $H_4$ as we find a strong and significant impact of the changes in tokens on the satisfaction with the decision. It does not appear that there is a significant interaction effect of the lost tokens with the nature of the decision maker (see specifications 3 and 4).

**Result 6.** Changes in tokens strongly affect the satisfaction with a decision, but only the general loss of tokens reduces the perceived fairness. The nature of the

[13]See also the comments of the participants explaining their ratings in the footnote above.

decision maker does not additionally contribute to the effect.

Comparing treatments with a choice (Tr1 and Tr2) and with exogenously determined decision makers (Tr3–Tr5) 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 *DV:Choice*), which is a direct rejection of $H_8$

**Result 7.** We do not find a significant impact of the ability to choose the decision maker on the satisfaction or perceived fairness of the decisions.

Providing the information on the group affiliation (variable *DV:Info*)—and therefore introducing the possibility of discrimination—significantly reduces both the perceived fairness and the satisfaction with the decision by about a third of a point. Importantly, in case of fairness, it does not matter if the player expects the group affiliation to be to his disadvantage or not.[14] If players believe that they might be discriminated against (as the outsider) this decreases satisfaction with the decision.

**Result 8.** The general possibility of discrimination significantly reduces the satisfaction with a decision and its perceived fairness. The satisfaction is lowest if negative discrimination is expected. However, the judgment of fairness is independent of whether the expected discrimination is positive or negative.

Additionally, participants with higher levels of general trust are more satisfied with the decisions and perceive them to be more fair, which corroborates that the drop in satisfaction comes from perceived discrimination. Age negatively affects both ratings.

According to our pooled OLS regression, the decision maker being human has a slight positive effect on the satisfaction, but not the fairness rating. To consider how the nature of the decision maker may affect satisfaction and perceived fairness more closely, we run a fixed effect panel regression with standard errors clustered at the individual level for treatments Tr1 and Tr2.[15] The fixed effect regression allows us to consider subjects who experienced within-subject variation of the decision maker. The results of the panel regression can be found in Table 5 in the Appendix. We largely include the same controls as in a pooled OLS specification if they were time variant. 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 decision maker it is rated as about a quarter of a point more fair and participants report a higher satisfaction (*DV:DM Human*). We therefore fail to reject our $H_5$ on the decision maker's nature and the impact on satisfaction and perceived fairness:

**Result 9.** If the redistribution decision is made by a human decision maker, it is perceived as more fair and leads to higher satisfaction.

Additionally, in the fixed effect regression we can consider how getting the participant's desired decision maker contributes to the ratings. In our experiment, if two players in the pair disagreed on the type of the decision maker, which happened in 21.6% of cases, one of the choices was randomly implemented. Therefore we have a subset of players who preferred algorithm but received human decision maker and vice versa. We do not detect any effect of receiving the decision maker one preferred on the fairness and satisfaction ratings (variable *DV:Preferred DM* in table 5).

**Result 10.** Receiving the type of decision maker that one voted for does not affect ratings of fairness and satisfactions of the decisions made by the decision maker.

[14] Treatments where no information was revealed were coded as "no discrimination"

[15] In the fixed effect regression, the sample is restricted to treatments Tr1 and Tr2 because in all other treatments the nature of the decision maker was invariant across matching groups.

### 4.3 A Matter of Principle or Matter of Money?

During the experiment participants earned tokens of different colors. They only knew that the exchange rate for each token into Euro was between 1 and 6 cents per token and that different colors had different exchange rates.[16] After the exchange rate was announced, participants saw the payout relevant round again and were asked again how satisfied they are with the decision and how fair they find it. This feature of the experiment allows us to consider if the ratings are driven by meeting fairness principles or by own monetary interests. To do so, we compare fairness and satisfaction ratings for a certain decision before (i.e., ratings submitted for this decision initially) and for the same decision after the exchange rate was revealed.

Ratings with and without information on the exchange rate are highly correlated (satisfaction: 0.77, $p < 0.001$; fairness: 0.81 $p < 0.001$). Participants adjusted their evaluation in both directions. Over all treatments the fairness score decreased by 0.1 point (SD=1.2) and satisfaction score by 0.05 points (SD=1.32) after the exchange rate was revealed. To consider if there are factors that affect the adjustment in a systematic manner, we use an OLS regression (see Table 6 with final adjusted fairness and satisfaction scores as a dependent variable and a series of controls.) Initial satisfaction and fairness ranking submitted for the particular decision before the exchange rate was known is unsurprisingly a strong predictor of the final rating and shows that participants do not revise their ratings randomly. Our specification can explain almost 70% of the variation, yet not many controls are significant. Both satisfaction and fairness are affected by monetary outcomes: the more participants earned in monetary terms, the more they increased their satisfaction and fairness rankings. Additionally, the satisfaction with the outcome decreases if participants would have earned more without any redistribution or if they would have earned more under their own hypothetical redistribution decision. Ceteris paribus, women decrease their ranking significantly more than men. We find again that an inconsistent mix of several fairness principles reduces the final satisfaction by an additional 0.44 points. As expected, the nature of the decision maker and other controls do not affect the adjustment.

**Result 11.** Learning the monetary values of the tokens significantly impacts the perception of the decisions. Higher monetary values lead to an increased satisfaction and higher fairness ratings, even though the relative positions and quantitative re distributions have not changed.

## 5 Discussion and Conclusion

We study whether people prefer a human or an algorithm to decide on redistributing their earnings and analyze the impact of discrimination on this preference. We examine how the nature of the decision maker affects the perceived fairness of the decision and the satisfaction with it. The question is motivated by increased use of automated decision systems in domains beyond analytical and predictive tasks and draws attention to the use of algorithmic decisions in a wide range of applications, from policy making to determining job routines and wider management issues. It is important to determine the preferred decision maker and the impact of their type on how decisions are perceived, because ultimately this will impact not only general satisfaction of the people affected but also their reactions to and their compliance with these decisions. If one type of decision maker is perceived to make fairer decisions, for example, this will lead to wider acceptance and implementation of these decisions. In short, who is the preferred decision maker might determine who *should* make the decision. We find that the algorithmic decision maker is strongly preferred, even in a "moral" situation.

[16] See section 3 or the instructions in the Appendix.

Algorithmic advice or decision support systems have become a common tool in managerial decisions ranging from small frequent choices such as allocating daily tasks and shifts to tasks with important consequences such as hiring new employees, distributing bonus payments, and even promotions. Given the number of tasks of these type in the workplace and the fact that technological advances already allow to automate many of them, it is important to establish whether people affected by these decisions prefer a human or an algorithmic decision maker and the influence that the type of decision maker has on how the decision is perceived. Apart from the direct effect on the employee satisfaction, perception of decisions might potentially affect performance (Bai et al., 2020; Strobel, 2019). Indeed, for example, a recent survey of the use of digital tools for HR management documents that while companies are aware of the benefits that digital HRM tools may bring, they are also concerned about how they would be perceived by the employees (Chugunova and Danilov, 2022).

Our study contributes to the large and burgeoning literature which considers the multitude of questions that arise with the advancement of technology. Does it lead to more equality and equal treatment (Tucker and Yu, 2019) or on the contrary only deepens existing inequality through digital divide (Warschauer, 2003)? How should algorithms be designed (D. Li et al., 2020) and what values should be integrated in them (Awad et al., 2018)? Do algorithms make objectively better decisions (Cowgill and Tucker, 2019)?

Our experiment provides two sets of results. First, with over 60% of participants choosing a redistributive algorithm, we find a strong preference for algorithmic decision makers. This is in stark contrast with the previous evidence that documents particular aversion to algorithms in social and moral tasks (Bigman and Gray, 2018; Gogoll and Uhl, 2018). Our findings suggest that while people may use algorithms too little themselves (Dietvorst et al., 2015), they prefer algorithms when they are affected by the decision. This preference for the algorithmic decision maker persists regardless of the potential discrimination. That is, it is not the perceived unbiasedness of the algorithm that drives the result.

Second, and somewhat in contrast to the first result, people are less satisfied with the decisions taken by the algorithm, and they judge them as less fair than human decisions. These are the same people who wanted the algorithm to make the decision. Our analysis identifies two main drivers for 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 decision makers and algorithms if fairness principles are applied incoherently. We do not observe a difference in reactions to mistakes by humans or algorithms that was reported in the previous literature as one of the reasons for algorithm aversion (e.g., ibid.). This result leads us to believe that a more sophisticated algorithm that does not allow for inconsistencies and makes fewer "mistakes" could elicit a more positive reaction. A smaller, but nevertheless significant factor is indeed the nature of the decision maker. Decisions made by a human, irrespective of the content or consequences of the decision, are rated as fairer and they lead to a higher satisfaction. Based on a recent study by Hidalgo et al. (2021), one might speculate that it might be due to lack of intentions of algorithmic decision makers.

How can we proceed with these seemingly conflicting results that people opt for the algorithmic decision maker but do not seem to like its decisions? In our view the lessons to be learned for a management context are clear. We do not find any aversion against algorithm decisions, even in the complex "moral" domain of redistributing earnings—we find the opposite, a preference for algorithms. Disclosing the use of decision-support algorithms or even the full reliance on AI decisions should not lead to negative reactions from the affected people. Depending on the situation, highlighting the use of algorithms could increase the perceived fairness of the decisions and thus even improve acceptance and efficiency. Given that we can pinpoint the reason for

the observed ex-post dissatisfaction with the algorithmic decisions, our conclusion is that an algorithm that coherently follows fairness principles would be preferred and its decisions would have outperformed the human decisions.

# References


Arnold, David, Will Dobbie, and Peter Hull (2021). "Measuring racial discrimination in algorithms". In: *AEA Papers and Proceedings*. Vol. 111, pp. 49–54.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan (2018). "The moral machine experiment". In: *Nature* 563.7729, pp. 59–64.

Bai, Bing, Hengchen Dai, Dennis Zhang, Fuqiang Zhang, and Haoyuan Hu (2020). "The impacts of algorithmic work assignment on fairness perceptions and productivity: evidence from field experiments". In: Available at SSRN: https://ssrn.com/abstract=355088.

Batson, C Daniel and Elizabeth R Thompson (2001). "Why don't moral people act morally? Motivational considerations". In: *Current directions in psychological science* 10.2, pp. 54–57.

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

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch (2014). "hroot: Hamburg registration and organization online tool". In: *European Economic Review* 71, pp. 117–120.

Bolton, Gary E and Axel Ockenfels (2000). "ERC: A theory of equity, reciprocity, and competition". In: *American economic review* 90.1, pp. 166–193.

Bors, Douglas A and François Vigneau (2003). "The effect of practice on Raven's Advanced Progressive Matrices". In: *Learning and Individual Differences* 13.4, pp. 291–312.

Burton, Jason W, Mari-Klara Stein, and Tina Blegind Jensen (2020). "A systematic review of algorithm aversion in augmented decision making". In: *Journal of Behavioral Decision Making* 33.2, pp. 220–239.

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

Chen, Yan and Sherry Xin Li (2009). "Group identity and social preferences". In: *American Economic Review* 99.1, pp. 431–57.

Chugunova, Marina and Anastasia Danilov (2022). *Use of digital technologies for HR management in Germany: Survey evidence*. MPRA Paper. University Library of Munich, Germany. URL: `https://EconPapers.repec.org/RePEc:pra:mprapa:111530`.

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

Cowgill, Bo, Fabrizio Dell'Acqua, and Sandra Matz (2020). "The managerial effects of algorithmic fairness activism". In: *AEA Papers and Proceedings*. Vol. 110, pp. 85–90.

Cowgill, Bo and Catherine E. Tucker (2019). "Economics, fairness and algorithmic bias". In: Available at SSRN: https://ssrn.com/abstract=3361280.

Dawes, Robyn M, David Faust, and Paul E Meehl (1989). "Clinical versus actuarial judgment". In: *Science* 243.4899, pp. 1668–1674.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey (2015). "Algorithm aversion: People erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology: General* 144.1, p. 114.

Dijkstra, Jaap J, Wim BG Liebrand, and Ellen Timminga (1998). "Persuasiveness of expert systems". In: *Behaviour & Information Technology* 17.3, pp. 155–163.

Eastwood, Joseph, Brent Snook, and Kirk Luther (2012). "What people want from their professionals: Attitudes toward decision-making strategies". In: *Journal of Behavioral Decision Making* 25.5, pp. 458–468.

Epley, Nicholas and David Dunning (2000). "Feeling" holier than thou": are self-serving assessments produced by errors in self-or social prediction?" In: *Journal of personality and social psychology* 79.6, p. 861.

Fehr, Ernst and Klaus M Schmidt (1999). "A theory of fairness, competition, and cooperation". In: *The quarterly journal of economics* 114.3, pp. 817–868.

Fisher, Anne (2019). *An Algorithm May Decide Your Next Pay Raise*. URL: `https://fortune.com/2019/07/14/artificial-intelligence-workplace-ibm-annual-review/` (visited on 12/09/2021).

Gallier, Carlo (2020). "Democracy and compliance in public goods games". In: *European Economic Review* 121, p. 103346.

Gogoll, Jan and Matthias Uhl (2018). "Rage against the machine: automation in the moral domain". In: *Journal of Behavioral and Experimental Economics* 74, pp. 97–103.

Grensing-Pophal, Lin (2021). *Algorithm Helps Companies with Back-to-Work Decisions*. URL: `https://hrdailyadvisor.blr.com/2021/01/22/algorithm-helps-companies-with-back-to-work-decisions/` (visited on 12/09/2021).

Hayes, Taylor R, Alexander A Petrov, and Per B Sederberg (2015). "Do we really become smarter when our fluid-intelligence test scores improve?" In: *Intelligence* 48, pp. 1–14.

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

Hertz, Nicholas and Eva Wiese (2019). "Good advice is beyond all price, but what if it comes from a machine?" In: *Journal of Experimental Psychology: Applied*.

Hidalgo, César A, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin (2021). *How humans judge machines*. MIT Press.

Hülle, Sebastian, Stefan Liebig, and Meike Janina May (2018). "Measuring attitudes toward distributive justice: The basic social justice orientations scale". In: *Social Indicators Research* 136.2, pp. 663–692.

Humm, Bernhard G et al. (2021). "Machine intelligence today: applications, methodology, and technology". In: *Informatik Spektrum* 44.2, pp. 104–114.

Konow, James (2003). "Which is the fairest one of all? A positive analysis of justice theories". In: *Journal of economic literature* 41.4, pp. 1188–1239.

Lee, Min Kyung (2018). "Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management". In: *Big Data & Society* 5.1, p. 2053951718756684.

Leyer, Michael and Sabrina Schneider (2019). "Me, You or AI? How Do We Feel About Delegation". In: *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Springer. ISBN: 978-1-7336325-0-8. URL: `https://aisel.aisnet.org/ecis2019_rp/36`.

Li, Danielle, Lindsey R Raymond, and Peter Bergman (2020). *Hiring as exploration*. Tech. rep. National Bureau of Economic Research.

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

Longoni, Chiara, Andrea Bonezzi, and Carey K. Morewedge (2019). "Resistance to medical artificial intelligence". In: *Journal of Consumer Research* 46.4, pp. 629–650.

Luhan, Wolfgang J, Odile Poulsen, and Michael WM Roos (2019). "Money or morality: fairness ideals in unstructured bargaining". In: *Social Choice and Welfare* 53.4, pp. 655–675.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook (2001). "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27.1, pp. 415–444.

Mellizo, Philip, Jeffrey Carpenter, and Peter Hans Matthews (2014). "Workplace democracy in the lab". In: *Industrial Relations Journal* 45.4, pp. 313–328.

Monin, Benoît and Anna Merritt (2012). "Moral hypocrisy, moral inconsistency, and the struggle for moral integrity." In.

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

Neyer, Franz J, Juliane Felber, and Claudia Gebhardt (2012). "Entwicklung und validierung einer kurzskala zur erfassung von technikbereitschaft". In: *Diagnostica*.

Parker, Sharon K and Gudela Grote (2020). "Automation, algorithms, and beyond: Why work design matters more than ever in a digital world". In: *Applied Psychology*.

Rawls, John (1971). *A theory of Justice/Revised Edition*.

Riberolles, Hervé de (2021). *Modernisation of incentive compensation: from simple commission rates to the most sophisticated calculation algorithms*. URL: https://www.primeum.com/en/blog/modernisation-of-incentive-compensation-from-simple-commission-rates-to-the-most-sophisticated-calculation-algorithms (visited on 12/09/2021).

Sanfey, Alan G., James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen (2003). "The neural basis of economic decision – making in the ultimatum game". In: *Science* 300.5626, pp. 1755–1758.

Sausgruber, Rupert, Axel Sonntag, and Jean-Robert Tyran (2021). "Disincentives from redistribution: Evidence on a dividend of democracy". In: *European Economic Review*, p. 103749.

Shank, Daniel B. (2012). "Perceived justice and reactions to coercive computers". In: *Sociological Forum*. Vol. 27. 2. Wiley Online Library, pp. 372–391.

Solow, Robert M (1957). "Technical change and the aggregate production function". In: *The review of Economics and Statistics*, pp. 312–320.

Stiroh, Kevin J (2001). "What drives productivity growth?" In: *Economic Policy Review* 7.1.

Strobel, Christina (2019). "The hidden costs of automation". In.

Sznycer, Daniel, Maria Florencia Lopez Seal, Aaron Sell, Julian Lim, Roni Porat, Shaul Shalvi, Eran Halperin, Leda Cosmides, and John Tooby (2017). "Support for redistribution is shaped by compassion, envy, and self-interest, but not a taste for fairness". In: *Proceedings of the National Academy of Sciences* 114.31, pp. 8420–8425. ISSN: 0027-8424. DOI: 10.1073/pnas.1703801114. eprint: https://www.pnas.org/content/114/31/8420.full.pdf. URL: https://www.pnas.org/content/114/31/8420.

Tajfel, Henri (1970). "Experiments in intergroup discrimination". In: *Scientific american* 223.5, pp. 96–103.

Tucker, Catherine E. and Shuyi Yu (2019). "Does it lead to more equal treatment? an empirical study of the effect of smartphone use on customer complaint resolution". In.

Van Esch, Patrick, J Stewart Black, and Joseph Ferolie (2019). "Marketing AI recruitment: The next phase in job application and selection". In: *Computers in Human Behavior* 90, pp. 215–222.

Wakslak, Cheryl J, John T Jost, Tom R Tyler, and Emmeline S Chen (2007). "Moral outrage mediates the dampening effect of system justification on support for redistributive social policies". In: *Psychological science* 18.3, pp. 267–274.

Warschauer, Mark (2003). "Demystifying the digital divide". In: *Scientific American* 289.2, pp. 42–47.

Waytz, Adam and Michael I Norton (2014). "Botsourcing and outsourcing: robot, British, Chinese, and German workers are for thinking — not feeling — jobs." In: *Emotion* 14.2, p. 434.

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

# A Appendix

## A.1 Additional Analysis: Algorithmic performance

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 colours (ttest p=0.000 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 made a choice of decision maker, Potentially, the difference in decisions could have affected the satisfaction and fairness scores, but we control for the type of the 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: $\chi^2(4, N=190)= 3.05$, $p = 0.5$; Talent Tokens: $\chi^2(4, N=205)=4.05$, $p = 0.399$; Effort Tokens: $\chi^2(4, N=217) = 3.6$, $p=0.461$).

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

## A.2 Additional tables

Table 5: Determinants of satisfaction and fairness ratings. Fixed effects panel regression

| | (1) | (2) |
|---|---|---|
| VARIABLES | Satisfaction | Fairness |
| DV:Unfair | -0.781*** | -1.376*** |
| | (0.207) | (0.246) |
| DV:DM Human | 0.248** | 0.241* |
| | (0.107) | (0.123) |
| DV:Lost Tokens | -0.519** | -0.899*** |
| | (0.210) | (0.250) |
| DV:Preferred DM | -0.0220 | 0.0412 |
| | (0.0962) | (0.124) |
| No Discrimination | 0.00301 | 0.00406 |
| | (0.164) | (0.175) |
| Neg. Discrimination | -0.166 | -0.0552 |
| | (0.134) | (0.163) |
| Luck: Own-Partner (After) | 0.00571*** | 0.00463** |
| | (0.00185) | (0.00214) |
| Talent: Own-Partner (After) | 0.00313** | 0.000247 |
| | (0.00128) | (0.00140) |
| Effort: Own-Partner (After) | 0.00371*** | 0.00147* |
| | (0.000821) | (0.000885) |
| Luck: Before-After | 0.00806*** | -0.00155 |
| | (0.00247) | (0.00313) |
| Talent: Before-After | 0.0143*** | 0.00266 |
| | (0.00347) | (0.00417) |
| Effort: Before-After | 0.0116*** | 0.00367 |
| | (0.00233) | (0.00329) |
| Luck: Hyp-Actual | -0.00495* | -0.00310 |
| | (0.00264) | (0.00271) |
| Talent: Hyp-Actual | -0.00223 | -0.000756 |
| | (0.00264) | (0.00273) |
| Effort: Hyp-Actual | -0.00108 | 0.000245 |
| | (0.00147) | (0.00150) |
| Constant | 1.196*** | 1.139*** |
| | (0.161) | (0.194) |
| Observations | 1,152 | 1,152 |
| R-squared | 0.366 | 0.187 |
| Number of id | 192 | 192 |

Observations from Tr1 and Tr2. Standard errors in parentheses

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

Table 6: Adjustment of satisfaction and fairness rankings after learning the exchange rate. OLS regression.

| VARIABLES | (1)<br>Final Satisfaction | (2)<br>Final Fairness |
|---|---|---|
| Satisfaction | 0.624*** | |
| | (0.0558) | |
| Fairness | | 0.779*** |
| | | (0.0362) |
| DV:Unfair | -0.448*** | -0.434 |
| | (0.162) | (0.271) |
| DV:DM Human | 0.0573 | 0.0274 |
| | (0.136) | (0.131) |
| Final Payoff | 0.00283*** | 0.00117*** |
| | (0.000491) | (0.000411) |
| Payoff Hyp Decision | -0.000927*** | -0.000442 |
| | (0.000282) | (0.000290) |
| Payoff Before Redistr | -0.000903** | -0.000398 |
| | (0.000364) | (0.000341) |
| Trust | 0.180* | 0.156* |
| | (0.0976) | (0.0818) |
| Age | -0.00379 | 0.00621 |
| | (0.0168) | (0.0162) |
| DV:Female | -0.273* | -0.309** |
| | (0.139) | (0.131) |
| No Discrimination | 0.297 | 0.0360 |
| | (0.259) | (0.246) |
| Neg. Discrimination | -0.0738 | -0.201 |
| | (0.212) | (0.190) |
| DV:Choice | -0.0382 | -0.0235 |
| | (0.137) | (0.136) |
| DV:Info | 0.281 | 0.0182 |
| | (0.216) | (0.212) |
| Constant | -0.494 | -0.110 |
| | (0.532) | (0.488) |
| Observations | 334 | 334 |
| R-squared | 0.670 | 0.691 |

Robust standard errors in parentheses

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

Table 7: Continuation: Determinants of satisfaction and fairness ratings. Pooled OLS regression.

| VARIABLES | (1) Satisfaction | (2) Fairness |
|---|---|---|
| DV:Unfair | -0.857*** | -1.489*** |
| | (0.161) | (0.166) |
| DV:DM Human | 0.137 | 0.0954 |
| | (0.0983) | (0.104) |
| DV:Unfair#DM Human | 0.359 | 0.409 |
| | (0.275) | (0.327) |
| DV:Lost Tokens | -0.503*** | -0.699*** |
| | (0.166) | (0.186) |
| DV:Info | -0.290* | -0.295** |
| | (0.149) | (0.150) |
| No Discrimination | -0.134 | -0.0309 |
| | (0.151) | (0.171) |
| Neg. Discrimination | -0.309** | -0.136 |
| | (0.133) | (0.154) |
| Luck: Own-Partner (After) | 0.00607*** | 0.00447*** |
| | (0.00143) | (0.00160) |
| Talent: Own-Partner (After) | 0.00506*** | 0.00250** |
| | (0.000913) | (0.000978) |
| Effort: Own-Partner (After) | 0.00520*** | 0.00335*** |
| | (0.000567) | (0.000610) |
| Luck: Before-After | 0.00604*** | -0.00174 |
| | (0.00199) | (0.00234) |
| Talent: Before-After | 0.00949*** | 0.000306 |
| | (0.00229) | (0.00280) |
| Effort: Before-After | 0.00783*** | 0.00220 |
| | (0.00169) | (0.00211) |
| Luck: Hyp-Actual | -0.00519** | -0.00303 |
| | (0.00213) | (0.00235) |
| Talent: Hyp-Actual | -0.00633*** | -0.00386* |
| | (0.00192) | (0.00202) |
| Effort: Hyp-Actual | -0.00174 | 0.000181 |
| | (0.00117) | (0.00114) |
| DV:Female | -0.00773 | 0.0918 |
| | (0.112) | (0.120) |
| Age | -0.0378*** | -0.0480*** |
| | (0.0129) | (0.0141) |
| Trust | 0.183** | 0.250*** |
| | (0.0764) | (0.0835) |
| Constant | 2.502*** | 2.611*** |
| | (0.386) | (0.415) |
| Observations | 2,004 | 2,004 |
| R-squared | 0.334 | 0.185 |

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

# B Experimental Instructions for the treatment Choice Info

The experiment was conducted in German. Original instructions are available upon request. As the experiment was conducted online, the instructions were displayed to participants screen by screen. They could take as much time as needed to read them. Highlights and formatting is preserved. Subsection names indicate if the instructions were shown to all participants or only some types.

## B.1 Displayed to all participants

**Welcome to the experiment!**

You participate in a paid economic experiment. How much you will earn depends on your decisions and partly on the decisions of other participants. Therefore, it is important that you read the following explanations carefully. Your decisions in the experiment and your answers in the subsequent questionnaire are **anonymous**. Neither the experiment leaders nor the other participants know which decisions *you* made and how much *you* earned.

**General rules of conduct.**

This experiment lasts about 35 minutes. You will be paid for your participation and we ask for your full attention during the experiment. Please do not listen to music or engage in conversation with others. Please switch your phone to silent or airplane mode. Please do not read or respond to emails, or interact on social media. Your payout will be displayed at the end of the experiment. You must complete the experiment and all questionnaires to receive the payment.

**End of screen**

**Which picture do you prefer?**

The experiment will start shortly. Please choose the picture that you like best. There is no right or wrong answer, just choose according to your taste. The chosen picture will be used to form groups during the experiment.

Please choose the picture that you like best.

**End of the screen**

Figure 2: By Wassily Kandinsky

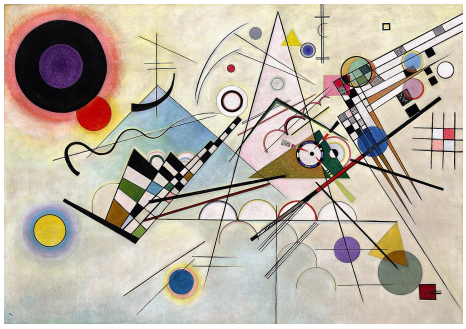


Figure 3: By Paul Klee

**General procedure**

This experiment consists of **two stages** and a final **questionnaire**. After completing the experiment, you will be redirected to a separate secure page to enter your bank details to transfer payment.

All participants in this experiment will be randomly divided into two types, **type P** and **type D**.

Type P and D have different tasks, which are explained in detail below.

**End of the screen.**

You are a **type X** participant. *You will receive further explanation at the beginning of each stage.*

**End of the screen.**

## B.2 Displayed to Type P only

**Stage 1: Type P**

In this stage you will get 3 separate tasks. In each of them you can earn money.

What you earn in these tasks will be carried over to stage 2 and determine your total payment.

During the experiment we use "points" instead of Euros. In each task you can earn points of different color (green, yellow, blue). At the end of the experiment, points of each color will exchanged for Euros with a different exchange rate. The exchange rate of each point is between 1 and 6 cents. It will be announced at the end of the experiment.

*Type D participants do not take part in this stage.*

In the following, each task will first be explained and then you can start working on it to earn points and money. After each task you will see the description of the next task. As soon as all participants have completed all tasks, you will see instructions for stage 2.

*End of the screen*

**Task 1**

In this task you do not have to do anything. The program will toss a virtual coin. If it land **Heads**, you will receive **100 green points**, if **Tales** you receive **0 green points**.

Button: Toss the coin.

**End of screen, followed by the screen with the outcome of the coin toss.**

**Task 2:**

In this task, you will see rows of "0" and "1" on the screen, as **in the example below.**

1 1 1 0 1 0 1 1 0 0
1 1 0 1 1 1 1 0 0 1
1 0 1 1 1 1 0 1 1 1
0 0 0 0 1 1 1 0 0 1
1 1 0 0 1 1 0 1 1 1
1 0 1 1 1 1 1 0 1 1

Please count how many **zeros are displayed** and enter the number in the field below and confirm the submission by pressing the button. You have 15 seconds for each screen.

To familiarize yourselves with the task and the time constraints, you will see an example screen before proceeding with the task. For solving the example screen you will not receive any points. After solving the example screen you will receive feedback and further instructions.

Button: Begin with the example

**End of screen, followed by the example task**

**Task 2: Further Instructions.**

You have solved the example task (in-)correctly.

In this task you can solve two screens like that. After you submitted a first answer, a second screen will appear and you can again count the number of zeros shown on the screen and enter the result.

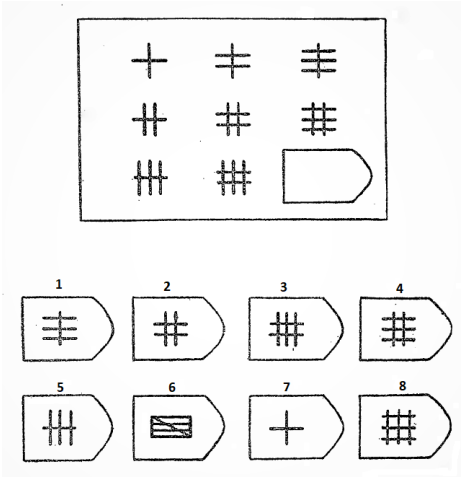
For each **correct** input you receive **100 yellow points**. For **incorrect** entries they receive **0 yellow points**.

*Previous research has shown that attention and diligence are the most important factors in this task.*

Button: Start the task

**End of screen, followed by the two sequential screens with the task and a feedback screen that displayed how many screens they solved correctly and how many points received**

**Task 3**
You will see a picture with 8 elements, as in the example below.




The elements in rows and columns follow a pattern. Your task is to find the element missing in the lower right corner. One of the elements 1 - 8 at the bottom of the picture is the correct element. Please enter the number of the correct missing element. In the example above the correct answer would be 8.

You have 30 seconds to answer.

*Previous research has shown that success in this task depends on talent.*

For a **correct** answer you get **100 blue points**. For an **incorrect** answer they get **0 blue points**.

Button: Start the task

**End of the screen. Followed by the task and feedback screen. A separate screen with the results of Stage 1 that displayed chosen picture and earned points of each colour was shown. **

**Stage 2 Type P**

You will now will be paired with another participant of type P. The groups are formed randomly. You will how many points from stage 1 the paired player earned and what picture did the paired participant choose at the beginning of the experiment.

The points you and the other participant earned will now be redistributed.

The new distribution of points can be determined either by a **type D** participant or by **a decision algorithm**. The points can remain with the respective participant or be redistributed. The decision will be made in points, since the value of the points will be revealed only at the end of the experiment.

Both group members can individually decide whether the algorithm or a (human) participant of type D should determine the distribution.

> **Type D** receives a flat payment, regardless of the distribution decision. We ask type D to choose a ***fair*** distribution. This participant also sees the two sets of points that members of the pair earned in **stage 1** and what picture both participants had chosen. You will also know which picture the type D participant had chosen. Type D will decide how the total number of points should be distributed between both Type P participants.
>
> **The algorithm** determines a fair distribution based on data from a survey with several hundred participants. The participants of the survey were informed about the three tasks they 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.

First, we ask both group members to choose **type D** or **the algorithm**. If both members make the same choice, the preferred decision maker (human/algorithm) is assigned. If the two group members choose different decision makers, it is randomly determined whether type D or the algorithm determines the distribution.

While the distribution of points is being determined, we ask you to indicate what you think is a fair distribution of points. We ask for your opinion, there is no right or wrong. What you state will not be shown to any of the other participants, type P or D. What you indicate here has no influence on your payout.

**Stage 2 is played through a total of six times.**

In each round of stage 2 you will form a pair with another participant of type P and each time you will choose whether type D or the algorithm will determine the distribution. A new distribution will be determined each time, always starting from your initial earnings in stage 1.

**After all six rounds are completed, the distribution of points in each round will be shown.** You will see both the initial distribution of points and the new distribution. We then ask you to indicate for each round to what extent you are satisfied with the distribution.

At the end of the experiment, **one** of these six rounds will be selected and you will be paid according to the points you have after the distribution in this round. Each round is equally likely to be payed out.

Once you have read and understood all the information on the screen, please press continue.

Button: Continue

**End of the screen. As participants had to wait for other players to finish reading the instructions, the instructions were repeated on the waiting screen.**

**Round 1**

You are in the group with the player who earn the following points.

| Players | Image | Green points | Yellow points | Blue points |
|---|---|---|---|---|
| You (Player A) | Klee | ... | ... | ... |
| Another player (Player B) | Kandinksy | ... | ... | ... |

The participant D chose **Klee/Kandinsky**.

Reminder:

- **Green points** were earned by the coin toss.
- **Yellow points** were earned by counting zeros.
- **Blue points** were earned by searching a missing puzzle piece.

How should the distribution of points be determined?

- The participant D should determine the distribution of points for the group.
- The algorithm should determine the distribution of points for the group.

**End of the screen**

**Round 1: Decision Maker**

You chose [**participant D (Picture)/algorithm**] to be the decision maker.

Your partner chose [**participant D (Picture)/algorithm**] to be the decision maker.

Therefore [**participant D (Picture)/algorithm**] will determine the distribution of points in the group.

Button: Continue

**End of the screen**

**Hypothetical Distribution**
If you were to make a distribution decision for your group, what distribution of points would be fair?

| Players | Picture | Green points | Yellow points | Blue points |
|---|---|---|---|---|
| You (Player A) | Klee | ... | ... | ... |
| Another player (Player B) | Kandinksy | ... | ... | ... |

**Available distribution options for green, yellow and blue points separately.**

**End of the screen. Matching into the groups, choice of the decision maker and the hypothetical decision by the player is repeated 6 times.**

**Results of Round 1: Decision Maker [participant D (Picture)/algorithm]**

In Round 1, [**participant D (Picture)/algorithm**] chose the following distribution for your group:
How happy are you with the decision of [participant D (Picture)/algorithm]? **7 point Likert scale: Very happy-very unhappy**

How fair do you find the decision of [participant D (Picture)/algorithm]? **7 point Likert scale: Very fair-very unfair**

| **Players** | **Picture** | **Green points** | | **Yellow points** | | **Blue points** | |
|---|---|---|---|---|---|---|---|
| | | Initial Points | Redistr. Points | Initial Points | Redistr. Points | Initial Points | Redistr. Points |
| You (Player A) | Klee | ... | ... | ... | ... | ... | ... |
| Another player (Player B) | Kandinksy | ... | ... | ... | ... | ... | ... |

Button: Continue

**End of the screen. Repeated for each round separately**

**Final results**

Round X was chosen for payment.

The value of the points for all participants is the following[17]:

- 1 green point is 5 cents
- 1 yellow point is 4 cents
- 1 blue point is 4 cents

Button: Continue

**End of the screen.**

**Final results**

In Round X, [**participant D (Picture)/algorithm**] chose the following distribution for your group:

| **Players** | **Image** | **Green points** | | **Yellow points** | | **Blue points** | |
|---|---|---|---|---|---|---|---|
| | | Initial Points | Redistr. Points | Initial Points | Redistr. Points | Initial Points | Redistr. Points |
| You (Player A) | Klee | ... | ... | ... | ... | ... | ... |
| Another player (Player B) | Kandinksy | ... | ... | ... | ... | ... | ... |

Given the value of points, your payment is calculated (in Euro) as follows:

| **Players** | **Image** | **Green points** | **Yellow points** | **Blue points** | **Total** |
|---|---|---|---|---|---|
| You (Player A) | Klee | ... | ... | ... | ... |
| Another player (Player B) | Kandinksy | ... | ... | ... | ... |

How happy are you with the decision of [participant D (Picture)/algorithm]? **7 point Likert scale: Very happy-very unhappy**

How fair do you find the decision of [participant D (Picture)/algorithm]? **7 point Likert scale: Very fair-very unfair**

Button: Continue to Questionnaire

[17]The value of the tokens was adjusted after the first session due to longer than expected average completion time. In the first session, 20 participants received 4 cents per green point, 2 cents for yellow point and 3 cents per blue point respectively. Initial and adjusted exchange rates are within the range that was announced to the participants at the beginning of the experiment.

**End of the screen.**

## B.3 Displayed to Type D only:

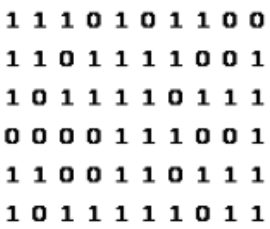
In this stage, you, as type D, do not have to make any decisions. You will decide in stage 2 how to redistribute the earnings of the other participants from this stage, an explanation of how type P participants earn money in this stage is provided.

Type P participants are given 3 separate tasks in which they can earn money. During the experiment, "points" are used instead of Euros. In each task, one earns points of one color (green, yellow, blue). How much each point is worth will be announced at the end of the experiment.

**Task 1:** In this task the participants do not have to do anything. The program will toss a virtual coin. If it lands **Heads**, they will receive **100 green points**, if **Tales** they will receive **0 green points**.

**Task 2:**

In this task, participants will see rows of "0" and "1" on the screen, as in the example below.




Participants count **how many zeros are displayed** and enter the number in the field below. They have **15 seconds** for each screen. After they enter the number, a new screen appears and they can again enter the number of zeros shown on the screen.

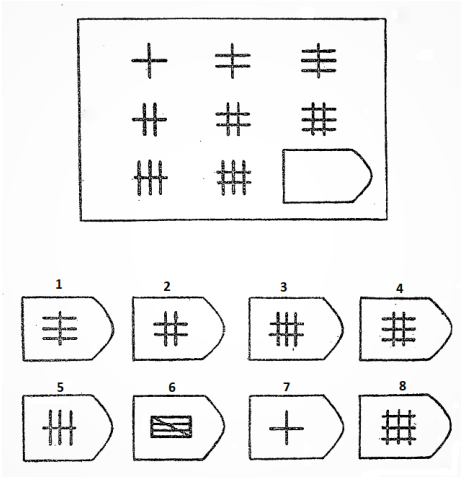
For each **correct** answer they receive **100 yellow points**. For **incorrect** answers they receive **0 yellow points**.

*Previous research has shown that attention and diligence are the most important factors to succeed in this task.*

To familiarize themselves with the task and the time constraints, participants solve a sample screen before proceeding with the task.

**Task 3**

Participants will see a picture with 8 elements, as in the example below.




The elements in rows and columns follow a pattern. The task of the participants is to find the element missing in the lower right corner. One of the elements 1 - 8 at the bottom of the picture is the correct element. Participants of type P have 30 seconds to submit an answer. In the example, the correct answer would be 8.

For a **correct** answer they get **100 blue points**. For an **incorrect** answer they get **0 blue points**.

*Previous research has shown that success in this task depends on talent.*

Button: Proceed to Stage 2

**End of the screen.**

**Stage 2 Type D**

After the participants of type P have earned points in stage 1, they are randomly matched into pairs. As player D, they are shown the group members' earnings in points and you are now asked to determine a fair distribution of the points. You will see the income from each task in the form of a table as shown below.

Example:

| **Players** | **Picture** | **Green points** | **Yellow points** | **Blue points** |
|---|---|---|---|---|
| You (Player A) | Klee | 0 | 100 | 100 |
| Another player (Player B) | Kandinksy | 100 | 200 | 0 |

You will know the income of the participants from stage 1 (as in the table above) and what picture the participant had chosen at the beginning of the experiment.

For each color (green, yellow, blue) you should separately specify how many points should each member get. One point of each color is exchanged for Euros at a different exchange rate. The exchange rate of each point is between 1 and 6 cents and will be revealed only at the end of the experiment.

You can redistribute the points as you like, there is no right and wrong. ***Please choose a distribution that you think is fair.***

You will see several groups one after the other (maximum 5 but it could happen that you are not assigned any group) and you are supposed to determine a fair distribution for each of these groups. The players in these groups are always called A and B, but they are always different people in each group. The distribution you choose will determine the payment the type P participants. You yourself will receive **a fixed payment of 10 euros**, regardless of the distribution you choose.[18]

**Stage 2 is played through a total of six times**. In each round, you choose distributions for up to five groups.

At the end of the experiment, **one** of these six replicates is selected and the type P participants are paid according to their score distribution.

Each round is equally likely to be relevant for payment.

Button: Continue

**End of the screen.**

**Round 1: Please make a decision**

Please make a decision for the following group:

| **Players** | **Image** | **Green points** | **Yellow points** | **Blue points** |
|---|---|---|---|---|
| You (Player A) | Klee | ... | ... | ... |
| Another player (Player B) | Kandinksy | ... | ... | ... |

Should the green points that were earned by tossing the coin be redistributed? **Possible redistribution options depending on the points of the pair**

Should the yellow points that were earned by counting zeros be redistributed? **Possible redistribution options depending on the points of the pair**

Should the blue points that were earned by finding a missing puzzle piece be redistributed? **Possible redistribution options depending on the points of the pair**

Button: Continue

**End of the screen. Repeated multiple times. The exact number of groups varies depending on the chosen nature of the decision maker **

**Final results**

Round X was chosen to be payed out.
The value of the points for all participants is the following:

- 1 green point is 5 cents
- 1 yellow point is 4 cents
- 1 blue point is 4 cents

Button: Continue to Questionnaire

[18]Payment to the Type D player was also adjusted after the first session. Two type D participants in the first session received 7 Euro flat payment.