Philosophy & Technology (2022) 35: 2
https://doi.org/10.1007/s13347-021-00495-y

RESEARCH ARTICLE

People Prefer Moral Discretion to Algorithms: Algorithm Aversion Beyond Intransparency

Johanna Jauernig[1] · **Matthias Uhl**[2] · **Gari Walkowitz**[3,4,5]

Received: 10 June 2021 / Accepted: 5 December 2021 / Published online: 26 January 2022

Abstract

We explore aversion to the use of algorithms in moral decision-making. So far, this aversion has been explained mainly by the fear of opaque decisions that are potentially biased. Using incentivized experiments, we study which role the desire for human discretion in moral decision-making plays. This seems justified in light of evidence suggesting that people might not doubt the quality of algorithmic decisions, but still reject them. In our first study, we found that people prefer humans with decision-making discretion to algorithms that rigidly apply exogenously given human-created fairness principles to specific cases. In the second study, we found that people do not prefer humans to algorithms because they appreciate flesh-and-blood decision-makers per se, but because they appreciate humans' freedom to transcend fairness principles at will. Our results contribute to a deeper understanding of algorithm aversion. They indicate that emphasizing the transparency of algorithms that clearly follow fairness principles might not be the only element for fostering societal algorithm acceptance and suggest reconsidering certain features of the decision-making process.

Keywords Algorithm aversion · Artificial intelligence · Moral discretion · Behavioral ethics

JEL Classification O33 · D91 · C91

✉ Gari Walkowitz
gari.walkowitz@tum.de

[1] Leibniz Institute of Agricultural Development in Transition Economies, Halle, Germany

[2] Technical University of Ingolstadt, Ingolstadt, Germany

[3] TUM School of Social Sciences and Technology, Technical University of Munich, Arcisstraße 21, 80333 München, Germany

[4] Research Group "Ethics of Digitization", Faculty of Informatics, Technische Hochschule Ingolstadt, Ingolstadt, Germany

[5] International Laboratory for Experimental and Behavioural Economics, National Research University Higher School of Economics, Moscow, Russia

1 Introduction

The use of decision-making algorithms promises societal benefits in a wide variety of applications. For many such applications, decisions have moral implications. Consider, for example, cases of lending and policing through algorithms. These tasks can be understood as a centralized agent's allocation of scarce resources (i.e., loans and police officers) among several groups with the goal of maximizing objectives (i.e., repayment and security). Fairness considerations based on the principle of equal opportunity require that equally creditworthy individuals or equally criminal individuals have the same chances of receiving a loan or of being arrested (Elzayn et al., 2019). However, it is especially in these morally sensitive domains that the use of algorithms faces societal resistance. Understanding algorithm aversion is therefore of utmost importance because it allows us to understand whether the resulting resistance to this technology can be addressed and, if so, at what level (Khasawneh, 2018). Necessary responses could be located on the level of governance where certain laws (e.g., liability law) would have to be adjusted or on the educational level where certain fears would have to be addressed through a demystification of algorithms and their actual functioning.

A prominent ethical objection to the use of algorithms concerns the opaqueness of machine learning (see, for instance, Mittelstadt et al., 2016; Lepri et al., 2018). Transparency is considered important to be able to contest the algorithm's implicit values on both epistemic and normative grounds (Binns, 2018). Epistemic arguments comprise questions regarding the performance of the algorithm like whether a model is generalizable or over-fitted, while normative arguments comprise, for instance, the inclusion of discrimination detection or fairness constraints (Binns, 2018). On normative grounds, opaqueness may be less of a problem if the algorithm's goal is to increase a clearly defined performance measure such as accuracy or speed in the absence of fairness concerns. If an algorithm-managed fund permanently outperforms the market, its opaqueness might be of less concern. If an algorithm's morally relevant case-by-case decisions cannot be accounted for by its programmers, how can we be sure that the machine follows an ethical rationale? How can we know that the algorithm does not base its decision to grant a loan or arrest someone on—say—racial characteristics? The problem becomes particularly apparent in the case of deep learning algorithms that follow allocation rules that they adjust endogenously based on incoming data.

The explainability of algorithmic decisions is urged for good reasons by ethicists of information technology (e.g., Mittelstadt, 2016; Wachter et al., 2017). Assuring this explainability might imply restricting the algorithm to follow predetermined and exogenous rules that it cannot adjust or accidentally change. However, it is less evident whether opaqueness is the sole reason for laypeople's algorithm aversion. It is well documented that people's attitude toward transparency is at least ambiguous. We all have a tendency to ignore relevant available information because processing that information is individually costly (Bettman et al., 1990) and might fundamentally challenge our self-image (Grossman & Van

Der Weele, 2017). People are especially likely to avoid potentially negative feedback regarding qualities that they care about, such as intelligence and beauty (Eil & Rao, 2011) or work performance (Moss et al., 2009). Thus, a decision to one's disadvantage can be attributed to the decision-maker's supposed (or actual) bias instead of one's own mediocrity, which may be a comforting conviction. Thus, the inclination to maintain a favorable self-image might feed into the resistance to algorithms, as they might also reveal unpleasant facts.

In this paper, we examine whether there is more to people's aversion to algorithms than the fear of opaque and biased decisions. This seems justified in light of evidence that suggests that people might not doubt the quality of algorithmic decisions, but still reject them. A representative large-scale survey shows that 73% of Germans do not want algorithms to make decisions without a human check.[1] A European survey contained similar results: 64% of respondents chose the statement "Algorithms might be objective, but I feel uneasy if computers make decisions about me. I prefer humans make those decisions" to the statement "I prefer that algorithms judge me instead of humans. They make more objective decisions that are the same for everyone."[2] The surveys also show that respondents provide many reasons for their reticent attitude toward algorithms. But these reasons seem to be less based on a reflected rationale but rather serve an underlying emotionally driven conviction. Thus, the negative attitude is a feeling of unease—following a hardwired heuristic—that is subsequently justified by numerous plausible arguments (Haidt, 2001; Sunstein, 2005). Put differently, the reticent attitude seems to be an initial emotional response that is only then followed by a conscious post hoc rationalization of the emotion.

It can be assumed that this skeptical attitude is as pronounced when decisions are explicitly ethical in nature. Experimental research shows that people perceive the same decision as less ethical and authentic when it is made by an algorithm instead of a human (Jago, 2019). The skepticism is also exemplified in the field of health care by the titles of recent books reviewed in *Nature*: *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (Topol, 2019) and *How Algorithms Could Bring Empathy Back to Medicine* (Insel, 2019). Topol (2019) claimed that as machines get smarter and take over more tasks, people must become even more human to compensate. From an analytical perspective, this raises the question of what exactly people miss in non-human decision-making.

The skepticism against algorithms may also be fueled by the fact that algorithms might perpetuate human prejudice. There is an extensive body of literature that elaborates on systematic und subconscious distortions in human decision-making such as in-group favoritism (e.g., Tajfel & Turner, 1986), self-serving bias (e.g., Babcock & Loewenstein, 1997), anchoring effects (e.g., Furnham & Boo, 2011), or overconfidence (e.g., Kahneman & Tversky, 1977). If, for example, algorithms allocate fewer

[1] https://algorithmenethik.de/wp-content/uploads/sites/10/2018/09/Was-die-Deutschen-%C3%BCber-Algorithmen-denken_ohneCover.pdf, retrieved on October 13, 2020.

[2] https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WhatEuropeKnowsAndThinkAboutAlgorithm.pdf, retrieved on October 13, 2020.

loans to certain groups or more police officers to certain areas due to racial characteristics, it is because their training data are derived from prejudiced human decisions. Indeed, recent research shows that algorithms are bound to perpetuate human biases because they are based on data generated by humans (Chander, 2016). A case in point is the Microsoft bot that was programmed to learn from Twitter users and mimic their conversation styles—and thus became a racist bully.[3]

To better understand the roots of algorithm aversion beyond opacity, we conducted two experimental studies. In Study 1, following our conjecture, we tested whether people empirically prefer humans, who are less restricted in their freedom to take morally charged decisions, over clearly rule-bound algorithms, even if veiled discrimination can be ruled out. Support for our conjecture would enable us to test whether people favor human decision-makers for principled or more instrumental reasons. Therefore, in Study 2, we tested whether people appreciate the mere presence of a human being in the decision-making process or the specifically human quality of moral autonomy (i.e., the ability to transcend rules in light of specific cases).

To investigate our research questions, we employed an economic experiment that used monetary incentives. This means that the choices that we elicited in our studies express participants' true preferences and are less likely to be biased, for instance, by expressions of social desirability (Grimm, 2010). Using an incentivized experiment is particularly advantageous in the context of algorithm aversion, as, in a hypothetical setting, people might merely choose humans over algorithms because they believe that this is what they are expected to choose as social beings. In our experiment, however, expressing a clear preference for one or the other decision-making entity (human or algorithm) comes with actual monetary costs and potential benefits. Furthermore, choosing a decision-making entity that one does not actually prefer might lead to an undesired monetary outcome. In line with the standards of economic research laboratories, no deception was applied to any of the participants: The course of the experiment was made transparent to all participants, and they knew about the no-deception policy beforehand. This is especially important in experiments with ethically relevant research questions, because the anticipation of being deceived might distort participants' decisions (e.g., Hertwig & Ortmann, 2001).

2 Experimental Setting

The experimental setting comprises a distribution conflict in which algorithms may apply human-created fairness principles to concrete cases. In the experiment, participants acted as "deciders" or "workers."[4] The core unit of the experiment consisted of a real-effort task in which participants in the role of workers generated a joint team budget in teams of two; this was later distributed between the two workers.

[3] https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html, retrieved on October 13, 2020.

[4] See Appendix A for an English translation of the experimental instructions.

In the real-effort task, each worker was individually confronted with a list of disarranged sliders on the computer screen (Gill and Prowse, 2012, see Fig. 1). Each slider was initially positioned between the positions "0" and "100," leaving out the middle position "50." Each worker had to position as many sliders as possible exactly to the 50 position within a given time frame of 8 min by clicking on the sliders with the mouse to drag and drop them. For each correctly positioned slider, a worker earned one Experimental Currency Unit (1 ECU = 10 Eurocent). After time was up, the joint team budget was determined as follows: One worker from the team was selected randomly, and this worker's earnings were doubled. The sum of the selected worker's doubled individual earnings and the (not doubled) individual earnings of the other (not selected) worker constituted the joint team budget for each worker team. For example, if the randomly selected worker correctly positioned 20 sliders and the other worker positioned five sliders correctly, the joint team budget was 20 ECU * 2 + 5 ECU = 45 ECU. Based on the workers' individual earnings (and the work efforts behind them) and the random factor in duplicating one worker's earnings, we expected a conflict to arise which triggered various individual moral attitudes on how to distribute the joint team budget.

The described real-effort task represented one stage of the experiment. In total, the experiment comprised four successive stages:

1. Determination of the distribution rule,
2. Regime choice,
3. Real-effort task (as described above), and

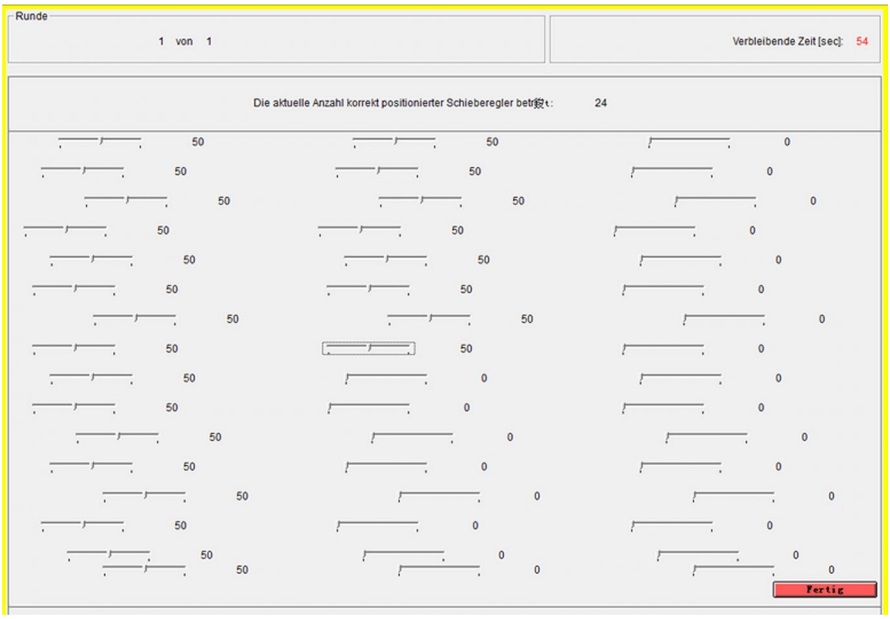

Fig. 1 Real-effort slider task

4. Implementation of distribution rule.

The course of the experiment is described in detail below.

2.1 Stage 1: Distribution rule determination

As mentioned above, participants were randomly assigned either the role of a worker or the role of a decider. The experiment started with the deciders, who determined a distribution rule according to which the joint team budget of each worker team (generated later during Stage 3) was to be distributed. For this, each decider was presented with the same set of six distribution rules. The distribution rules accounted for the individual earnings of each worker in a team, as well as for the duplication of the individual earnings of one randomly selected worker. The rules included "merit-based," "equal shares," and "winner-takes-all" principles (Messick, 1993), or a combination thereof.[5] For example, one rule posited that each worker in the team shall receive exactly the amount the worker earned, and the additional amount resulting from the duplication of the amount earned by one worker shall be divided equally between the two workers. Deciders had to choose one out of the six presented distribution rules. That way, we ensured that the distribution rule was ultimately determined in a process which engaged humans.

Each rule chosen by the deciders was placed in a virtual ballot box from which one of those rules was subsequently drawn by the computer. If several deciders opted for the same rule, the chances of this rule being drawn increased accordingly.[6] Later on, this rule could determine the distribution of the joint team budget among the workers on each team. The deciders were immediately informed about which rule from the virtual ballot box had been chosen. The workers were informed about how the distribution rule was determined. However, they learned neither the set nor the content of the six distribution rules. They learned which rule had been selected only at the end of the experiment. Note that we intentionally did not disclose the content of the rule to the workers. Workers could only choose the way in which the rule was applied, which is described in the subsequent section.

2.2 Stage 2: Regime choice

To investigate whether people disliked the algorithm's lack of moral discretion, we varied experimentally how rigidly the previously determined distribution rule was implemented (changeable vs. not changeable) and by whom (human vs. computer). Consequently, at Stage 2, the workers could choose among three regimes, which

[5] For the complete set of available distribution rules, see Appendix B.

[6] This voting mechanism traces back to the concept of "lottery voting." Scholars argue that using this procedure, minority interests are more fairly represented in the framing of legislation and every group in a community would get its fair stint of representation (e.g., Amar, 1984). We could have opted for another mechanism to determine the rule, e.g., a majority vote. Participants may perceive different mechanisms as differing in fairness. However, as noted earlier, for our question, the crucial aspect was the human involvement in the mechanism to determine the rule. This aspect was kept constant across all treatments.

Table 1 Overview of the regimes that workers could choose

Regime	Rule implementer	Rule changeable?	Cost for rule change (ECU)
(1) Discrete human	Human	Yes	0
(2) Rule-bound human	Human	Yes	100
(3) Rule-bound algorithm	Computer	No	N/A

The second column indicates who implements the distribution rule in each regime. The third column indicates whether the initially determined rule can be changed. The last column shows the associated cost for changing the initially determined distribution rule.

determined how rigidly and by whom the previously determined distribution rule was implemented after the workers generated the joint team budget (see Table 1 for an overview).

1) In the *discrete human* regime,[7] human deciders were assigned to a worker team and could—after learning the individual earnings of the workers on the assigned team and whose individual earnings were duplicated by chance—freely change the previously determined rule. The decider could either stick to the previously determined distribution rule or select one of the other five. Hence, the workers knew that in the *discrete human* regime, the deciders were not bound to the previously determined distribution rule. (Note, however, that the workers did not know the content of the determined rule.) The workers also knew that the deciders in the *discrete human* regime earned a fixed amount of 100 ECU whether they implemented the previously determined distribution rule or selected another.

2) Under the *rule-bound human* regime, deciders were also assigned to worker teams. As in the *discrete human* regime, they were not bound to the previously determined distribution rule and could change it after learning the individual earnings from the workers of their team and whose individual earnings where duplicated by chance. However, under the *rule-bound human* regime, the deciders' payment depended on whether they implemented the previously determined distribution rule or decided to change it. If they implemented the previously determined distribution rule, they earned 100 ECU. If they picked another rule, they received no payment (i.e., 0 ECU). Workers were informed that changing the previously determined distribution rule was costly to the deciders under the *rule-bound human* regime.

3) Under the *rule-bound algorithm* regime, no deciders were assigned to the worker teams. The previously determined distribution rule was applied automatically to the worker teams by the computer without human interference.

Generally, workers were not forced to choose any of these regimes. They could refrain from choosing a specific regime and express their indifference. If they chose a specific regime,

[7] To ease understanding, here, we refer to regimes by the more telling names "discrete human," "rule-bound human," and "rule-bound algorithm," respectively. In the experimental instructions, regimes were only distinguished by their number to prevent influencing participants through wording.

1 ECU was deducted from their experimental income. Again, please note that when choosing a regime, workers did not know which rule had been determined during Stage 1. This is a crucial design feature: Knowing that the rule is, say, merit based could open up strategic reasoning including one's own performance estimation as well as the chance that a human decider would switch to a more equity-based rule. We needed to rule out this kind of reasoning qua design to measure our main dependent variable without additional interfering variables. Therefore, workers only knew about the features of each regime: who will finally implement the distribution rule (a human or a computer), whether the implementer could change the previously determined distribution rule (yes or no), and what the monetary consequences were if the rule was changed (costly or not).

After workers chose their preferred regime during Stage 2, teams of two workers who chose the same regime were formed. Those workers who did not choose a regime during Stage 2 were either assigned to a regime for which there was a spare worker who had no team partner, or, if this demand was fulfilled, randomly assigned to a regime in pairs of two.[8]

Next, deciders were actually assigned to worker teams consisting of workers who either chose regime *discrete human* or *rule-bound human*. Workers who chose *rule-bound algorithm* were not assigned a decider because, in this regime, the previously determined distribution rule was implemented automatically by the computer without the possibility of human interference. If all workers chose *rule-bound algorithm* or did not prefer a regime and were therefore randomly assigned to the *rule-bound algorithm* regime, surplus deciders were not assigned to a worker team. They remained inactive for the rest of the experiment and received fixed earnings of 100 ECU.

2.3 Stage 3: Real-effort task and generation of joint team budget

At this stage, the workers generated the joint team budget in teams of two by positioning the sliders on their computer screen within the given period of 8 min. After the workers finished, the computer selected one worker randomly from each team, duplicated the worker's earnings, and calculated the generated joint team budget for each team. (See above for a detailed description of this stage.)

2.4 Stage 4: Change and implementation of selected distribution rule

After the joint team budgets were generated by the worker teams, the deciders in the *discrete human* and *rule-bound human* regimes were informed of the size of the team budget for their assigned worker team, the individual earnings of each worker on the team (corresponding to the number of correctly solved slider tasks), and whose earnings were randomly duplicated. The deciders in the *discrete human* and *rule-bound*

[8] Due to the applied matching procedure, it was possible (in the case of uneven numbers of workers who chose the same regime and in cases without workers who were indifferent) that one worker could not be assigned to their chosen regime. In this case, the worker was forced into another regime, and no fee was deducted from their earnings. Participants were informed about this very unlikely possibility up front.

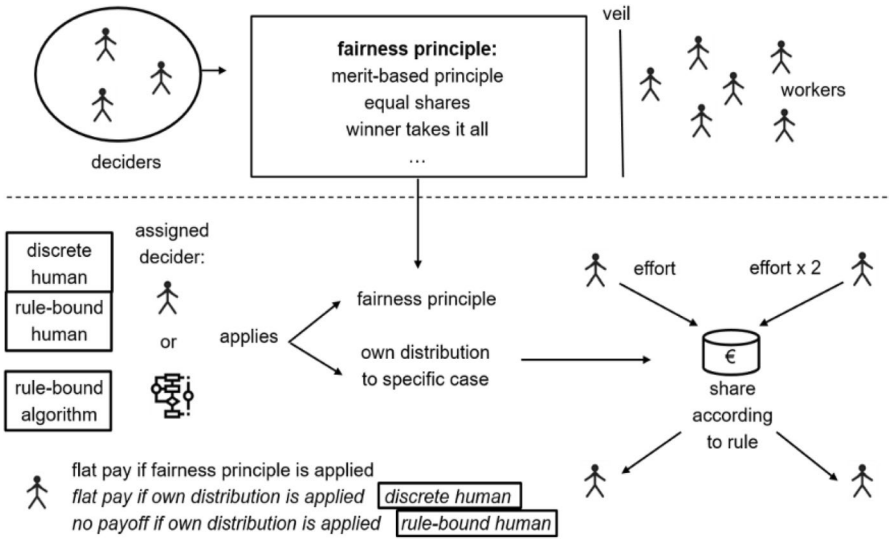

Fig. 2 Overview of experimental setting

human regimes could then implement a distribution rule. This rule could be the rule that was originally determined in the random draw from the virtual ballot box during Stage 1 (and displayed on their computer screens) or another rule from the set of distribution rules. If the deciders chose to change the previously determined rule, they incurred no cost under the *discrete human* regime and a cost of 100 ECU under the *rule-bound human* regime. Because there was no decider assigned to the worker teams under the *rule-bound algorithm* regime, the previously selected distribution rule was implemented automatically without human interference.

After the distribution rules were implemented, individual payoffs were calculated and paid out at the end of the experiment. Before the experiment ended, participants completed a set of post-experimental questions and were informed about their final earnings. Figure 2 provides an overview of our experimental setup.

3 Study 1: Testing for Algorithm Aversion Beyond Opaqueness

In our first study, we tested whether workers expressed algorithm aversion by preferring a regime involving human deciders who have the discretion to discard a fairness principle at will in light of an individual case. This implied that workers rejected a regime in which the distribution rule was unchangeable and automatically applied to an individual case with no possibility of human interference. If people's skepticism toward algorithms overlays potential reservations about human imponderables when veiled discrimination can be ruled out by design, people will prefer human deciders who are free to decide idiosyncratically over the rule-bound algorithms. We explicitly addressed this question to investigate whether the often-identified algorithm

aversion might be based on more fundamental grounds than the fear of opaque or potentially biased decisions.

3.1 Participants and Procedure

Study 1 was conducted in an economic laboratory of a large university with students majoring in various disciplines. Data were collected in January 2020, and participants were contacted through university mailing lists using the online recruitment system ORSEE (Greiner, 2015). We recruited 90 participants (50% female) who took part in three experimental sessions. Upon arrival at the laboratory, participants drew an individual code number and were seated individually in opaque cubicles. Participants then received written instructions for the experiment.[9] The instructions explicated the entire course of the experiment and were read aloud by the experimenters. Subsequently, participants answered a set of comprehension questions. Only after participants had successfully answered all items was the computerized experiment (programmed with z-Tree, Fischbacher, 2007) begun.[10] At the outset, one-third of the participants were randomly assigned the role of deciders, and two-thirds of the participants the role of workers.[11] The experiment lasted about 1 h. Subsequently, participants were compensated with a fixed amount of 4€ for showing up, along with the amount they earned during the experiment.

3.2 Measures

In Study 1, workers were instructed about the three regimes described in Sect. 2 and could then choose between the *discrete human* and *rule-bound* algorithm regimes. Workers could also express indifference and choose neither regime. Under the *discrete human* regime, deciders could change the previously determined distribution rule without cost (i.e., they had full discretion over the implementation of the distribution). Under the *rule-bound algorithm* regime, the previously determined distribution rule was unchangeable and was implemented automatically by the computer without human interference. In total, 60 participants acted as workers, and our main dependent variable was the workers' choice of regime. We also assessed the number of correctly solved tasks depending on workers' regime preferences. All decisions were incentivized monetarily.

3.3 Results

Regime choice Descriptive results can be inferred from the left panel in Fig. 3. In total, 73.33% (CI = 0.603, 0.839) of the workers chose one of the two available regimes. Only 26.67% (CI = 0.161, 0.397) expressed indifference. The fraction of

[9] See Appendix A for an English translation of the experimental instructions.

[10] Please refer to the complete list of comprehension questions in Appendix C.

[11] In the experimental instructions and on the computer screens, we used the neutral terms "participants with role A" and "participants with role B.".

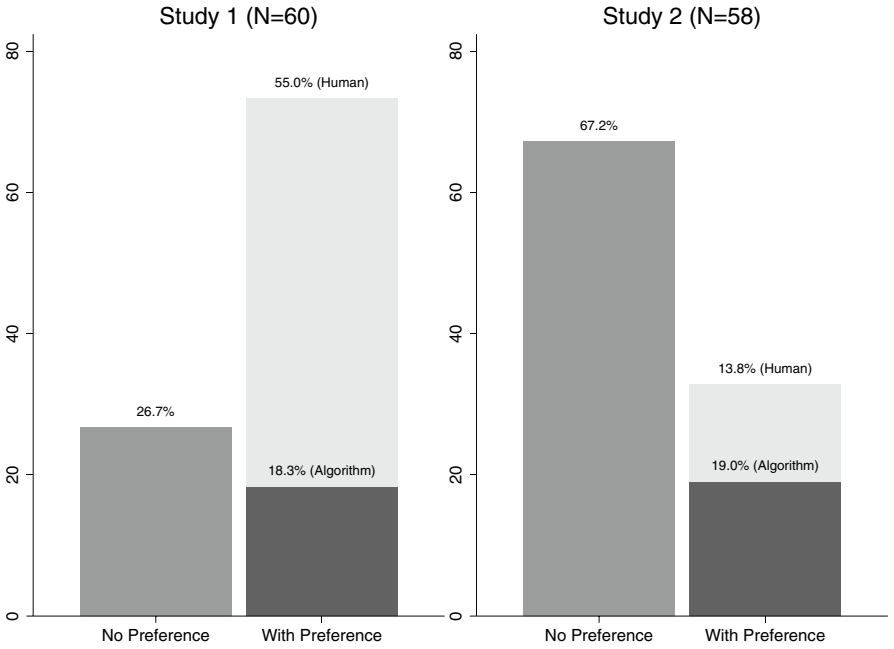

Fig. 3 Preferences for regime variants

workers who had a specific preference was significantly larger than the fraction who had no specific preference ($p < 0.001$, binomial probability test,[12] two-sided). Most workers who expressed a preference preferred the human decider with decision discretion (55%, CI = 0.416, 0.67) over the rule-bound algorithm (18.33%, CI = 0.095, 0.304). The number of workers who preferred the human decider was significantly larger relative to the number of workers who preferred the algorithm ($p = 0.001$, BPT).

Performance We find no evidence that the number of correctly solved tasks differed significantly depending on whether the participants preferred a regime (M = 21.41, CI = 16.746, 26.072) or not (M = 24.19, CI = 17.258, 31.117, $d = 0.188$; $p = 0.522$, Fisher–Pitman permutation test for two independent samples[13]). Performance also did not differ significantly between those participants who chose a human decider (who was not bound to the previously determined distribution rule) (M = 19.67, CI = 14.953, 24.381) or the algorithm (which implemented the rule automatically) (M = 26.64, CI = 13.102, 40.171) and those who had no preference for a specific regime variant (M = 24.19, CI = 17.258, 31.117; for all pair-wise comparisons, $d < 0.458$, and $p > 0.201$, FPT).

When we contrasted the performance of participants who were eventually confronted with the human decider—independently of whether they initially had a preference for the human decider or had been indifferent (M = 19.23, CI = 15.213,

[12] In the following denoted as *BPT*. All statistical tests are carried out two-sided if not stated otherwise.
[13] In the following, denoted as *FPT*.

23.238)—with the performance of those participants, who were eventually confronted with the algorithm—also independently of whether they initially had a preference for the algorithm or had been indifferent (M = 28, CI = 19.982, 36.018)—we found that the former group performed significantly worse compared to the latter ($d = 0.617$; $p = 0.031$, FPT).

3.4 Discussion

Study 1 identified an aversion to the algorithm that implemented a human-created exogenous rule. The workers clearly preferred a human decision-maker with discretion. Did this occur because people value the involvement of flesh-and-blood beings in ethical decision-making or because they value moral autonomy in view of individual cases? Put differently, is it the human nature of the decision-making entity itself or the human capacity to transcend rules and apply one's own ethical standards that causes participants to favor humans over algorithms? We wanted to answer this question in Study 2 to achieve a deeper understanding of the observed algorithm aversion.

The first explanation implies that people have an intrinsic aversion to the use of algorithms in taking morally charged decisions. Ethicists usually define intrinsic value as a thing's inherent value (Zimmerman, 2019). In this vein, previous studies indicate an aversion to algorithms that seems based on intuitive or emotional discomfort rather than on specific rational motives. This is corroborated by the fact that on surveys, respondents typically provide numerous justifications for their reticent attitude toward algorithms.[14] Bigman and Gray (2018) have documented in a series of studies that people consider moral decisions to belong to the domain of humans, not algorithms. They show that people are averse to machines making ethically relevant legal, medical, and military decisions and that this aversion is mediated by the perception that machines can neither fully think nor feel. Gogoll and Uhl (2018) ruled out the possibility that algorithm aversion was based on a misperception of machine errors or a distrust in the technology. They concluded that people might exhibit an aversion to algorithms making moral decisions per se. One potential explanation for preferring that humans be involved in decision-making merely because they are human is that this may allow people to blame someone if they are unsatisfied with the outcome of the decision-making process (Danaher, 2016) and are thus able to lodge a (formal) complaint against the individual decider.

The second explanation implies that people ascribe an instrumental value to humans—as opposed to machines—making moral decisions. Instrumental value

[14] https://algorithmenethik.de/wp-content/uploads/sites/10/2018/09/Was-die-Deutschen-%C3%BCber-Algorithmen-denken_ohneCover.pdf, retrieved on October 13, 2020.

is defined as the value that something has as a means to an end.[15] A fundamental instrumental difference between humans and algorithms is that the former usually have discretion in moral decision-making. Empathy sometimes lets us spontaneously transcend known rules in ethical decision-making in light of specific circumstances. The positive interpretation of human discretion is reflected in the saying "Let mercy take precedence over justice." An algorithm is incapable of feeling empathy and showing mercy in the traditional human sense.[16] However, it is also incapable of denying someone their rights due to personal dislike. Rules are reliable, whereas discretion creates room for interpretation that opens the door for biases such unintended discrimination. Although we believe that people's judgment is reliable, they often expose unintentional unethical behavior (e.g., Kim et al, 2015; Sezer et al., 2015), based on their perceived expertise or their personal relationships. Hence, the human mind still remains a black box (Chander, 2016) and is, therefore, also incomprehensible and potentially susceptible to distortions that might have significant social implications. If the aversion to algorithms was driven mainly by the fear of inexplicability, we might observe a similar skepticism toward human discretion and a preference for strictly rule-based behavior in the context of morally charged decisions. There exists evidence that people favor humans to take decisions that affect them over algorithms more, if the task at hand is perceived to be highly subjective (Castelo et al., 2019) or if they consider themselves as very special people (Longoni et al., 2019). It is, however, not clear whether this phenomenon is also relevant for taking decisions that are morally charged.

4 Study 2: Testing Intrinsic and Instrumental Explanations of Algorithm Aversion

In our second study, we tested whether people prefer the mere presence of a human being in the decision-making process to a rule-bound algorithm. We conjectured that if people appreciated the mere presence of human deciders, they should prefer a regime with a human decider to a regime with algorithm-based decisions, even in situations where deciders must pay a high price if they deviate from the previously determined distribution rule (which makes this action extremely unlikely). In contrast, if people appreciate human deciders' ability to transcend rules in light of individual work performances (and do not prefer human deciders per se), they should be indifferent toward the choice between a decider who, de facto, is bound to the previously determined distribution rule and an algorithm with the same

[15] Although Korsgaard (1983) argued that intrinsic value should be contrasted with "extrinsic value" (a thing's value in virtue of relational properties) and instrumental value should better be contrasted with "final value" (a thing's value as an end for its own sake), the opposition of intrinsic and instrumental value is still the traditional one.

[16] One might argue that showing mercy does not necessarily require feeling empathy. Thus, a deep learning algorithm might be able to mimic human mercy. Granting this mercy, however, would have to be random, because if it was systematic, it would be possible to encapsulate it in the ex ante rule. It seems questionable that this randomness would accommodate people's intuition of what spontaneous mercy is.

property. We explicitly tested these two potential explanations to gain a deeper understanding of the causes of the algorithm aversion identified in Study 1.

4.1 Participants and Procedure

We recruited 90 participants (55% female) who took part in four experimental sessions. Data were collected in February and March 2020. Study 2's procedures were similar to those of Study 1.

4.2 Measures

In Study 2, the workers could again choose between two of the three regimes described in Sect. 2, namely between the *rule-bound human* and *rule-bound algorithm* regimes. Workers could also express indifference and choose neither regime. Under the *rule-bound human* regime, deciders could change the previously determined distribution rule, incurring a cost (i.e., they had discretion over the implementation of the distribution rule but faced the loss of their payoff if they switched to another rule). Under the *rule-bound algorithm* regime, the previously determined distribution rule was unchangeable and was implemented automatically by the computer without human interference. In total, 58 participants acted as workers. Once again, our main dependent variable was the workers' regime choice. We also assessed the impact of participants' regime choice on the number of correctly solved tasks. All decisions were incentivized monetarily.

4.3 Results

Regime choice Descriptive results can be inferred from the right panel of Fig. 3. Most workers (67.24%, CI = 0.537, 0.790) chose neither regime. Only about one-third of the workers (32.76%, CI = 0.21, 0.463) either preferred the human rule-bound decider (13.79%, CI = 0.061, 0.254) or the algorithm (18.97%, CI = 0.099, 0.314). The proportion of workers who expressed no preference was significantly larger than the proportion of workers who had a specific preference ($p = 0.012$, BPT); it was also significantly larger than the proportion of workers who had no specific preference in Study 1 ($p < 0.001$, Chi square test). In addition, in contrast to Study 1, for workers who had a regime preference, there is no evidence that the preferences for the human rule-bound decider and the algorithm differed significantly from an equal distribution ($p = 0.647$, BPT).

Performance We found no evidence that the number of correctly solved tasks differed significantly depending on whether the participants preferred a regime (M = 25, CI = 17.538, 32.462) or not (M = 24.15, CI = 20.352, 27.956, $d = 0.065$; $p = 0.825$, FPT). Performance also did not differ significantly between those participants who chose a human rule-bound decider (M = 23.88, CI = 8.181, 39.569) or the rule-bound

algorithm (M = 25.82, CI = 16.731, 34.905) and those who did not prefer a specific regime (M = 24.15, CI = 20.352, 27.956; for all pairwise comparisons, $d < 0.137$ and $p > 0.704$, FPT). Similarly, when comparing the performance of participants who were *eventually* confronted with the human rule-bound decider independently of whether they initially had this preference (M = 26.032, CI = 21.001, 31.064) with the performance of those participants who were *eventually* confronted with the algorithm independently of whether they initially had a preference for the algorithm (M = 22.59, CI = 17.852, 27.334), we found no significant performance difference ($d = 0.266$; $p = 0.321$, FPT).[17]

4.4 Discussion

Study 2 revealed that participants' aversion to algorithms as identified in Study 1 was not based on a resistance to the artificial nature of the decision-maker per se. The results instead indicate that people ascribe an instrumental value to humans' making moral decisions as opposed to machines' doing so. Decision-makers in the *discrete human* regime in Study 1 had full discretion concerning which distribution rule to implement. Once this possibility was no longer available, the workers expressed no clear preference for the human decider. Workers' regime choices were not driven by their desire to have a human being apply the fairness principle to an individual case (e.g., to attribute responsibility and project blame onto this person). If humans are perceived as bureaucratic executors of predetermined fairness principles, they may be easily replaced by algorithms. In this sense, the aversion to algorithms is not intrinsic.

Although deciders are not the focus of analysis, it is informative to also report how they actually behaved in our experiment. This holds particularly in the *discrete human* regime of Study 1, where deciders were not bound to the previously determined distribution rule and could change it without cost after learning workers' earnings and whose earnings where duplicated by chance. As argued above, it is plausible to assume that decider behavior is associated with workers' expectations and that deciders reacted to the information on workers' individual efforts and duplication luck by spontaneously rewarding or punishing single workers through rule change. Two points stand out in this investigation. Firstly, only five out of 35 deciders (four out of 20 in the *discrete human* regime of Study 1 and one out of 15 in the *rule-bound human* regime of Study 2) made use of their right to change the previously determined distribution rule. Secondly, all these deciders re-enforced their initially chosen, but not previously applied, distribution rule. This was the case although the content of the initially chosen rules was different among these deciders. As some of the initially proposed distribution rules were then actually drawn at random, the number of those deciders who subsequently wanted to switch back to their originally proposed distribution rule might be even higher and our figures represent only a lower bound for this behavioral pattern.

[17] In the entire experiment, all participants who had a preference ended up with the regime they preferred.

5 General Discussion

In conjunction with the findings from Study 1, the results from Study 2 indicate that algorithm aversion is, at least partly, driven by instrumental deliberations: People do not dislike algorithms intrinsically, but cherish the discretionary scope of human deciders. The distinction between both kinds of algorithm aversion may seem rather academic at first, but it has important implications. If the rejection of algorithms in the context of morally charged decisions was based on an intrinsic aversion to the very artificiality of the decision-making entity, it might prove difficult to replace human decision-makers by algorithms as this artificiality is an essential feature of the algorithm. If, however, the rejection of algorithms is based on an instrumental aversion rather than the appearance of the algorithm, features of the decision-making process need to be changed to accommodate the algorithm's functioning with people's moral attitudes.

Krishnan (2020) expresses concern that to the extent that researchers working on interpretability emphasize its indispensability, they may fuel the public mistrust of algorithms. The results of our studies suggest that the reasons for the widely observed aversion against algorithms in the context of morally charged decisions may indeed be more multifaceted than implied by some parts of the ethical literature. Although many ethicists discuss the important problem of opaque algorithms, the traceability of moral decisions through transparency might not be the public's only concern. On the contrary, participants in our studies seemed to appreciate an element of discretion in moral decision-making, as it is peculiar to human beings. The implication of this discretion is the ability to override a fairness principle in light of a specific case that was ex-ante deemed appropriate across various cases. People's taste for "instantaneous fairness" might imply that opaque algorithms with humans involved in the decision-making process might be more readily accepted than transparent and rule-bound ones that do not involve humans, simply because these algorithms have no capacity to discard fairness principles based on spontaneous inspirations. Testing this conjecture might be an interesting path for future research.

It is noteworthy that the regimes under which workers in our experiments performed and their actual performances were interrelated. When people worked under the regime of the human decider with a discretionary scope, their performance was lower. This might have been caused by a lower ability to perform a given task, by a lower motivation to do so, or both. In any case, low performers seemed to expect the discrete human decider to let mercy take precedence over justice. Their behavior indicates that they expected deciders to deviate from the previously determined fairness principle in favor of equality of outcome through a spontaneous act of pity for the low performer in light of their poor performance. This result points to the fact that the degree of rule conformity implied by the decision-making regime should not be discussed irrespective of its consequences. The chosen regime interacts with the actions people take under the regime. People will likely adapt to the incentives they face. If borrowers rely on the spontaneous leniency of human lenders, this could have implications for their fiscal discipline. If the convicted hope for the spontaneous leniency of human judges, this could influence their parole

behavior. This indirect effect of discrete or rule-bound regimes of moral decision-making, which our findings indicate, warrants the attention of ethicists and future studies.

With regard to the regime's direct influence on performance, our study is a starting point for further investigations into the implications of algorithm supported decision-making in the context of morally charged decisions. Here, new experimental studies are needed to shed light on the question of whether the decision regime (human or algorithm-based) influences people's propensity to engage in undesirable behavior such as laziness, free-riding, or even cheating. This research will help to address the issues arising when algorithms are implemented in decision-making processes and ultimately ensure that the positive effects of algorithms such as efficiency and incorruptibility are realized while avoiding negative counter-effects. Numerous voices are calling for caution in the application of algorithms to domains with ethical implications such as lending, policing, or medicine, so research must accompany technological development with social and ethical analyses. This research agenda delivers on this claim.

Our studies are subject to several limitations, two of which we want to mention here. First, our results are based on a sample of students from a technical university. Assuming that individuals from this sample are technophiles who are more open-minded with respect to the use of algorithms in decision-making, some caution is required in generalizing our effects. It is conceivable that a more technophobic sample would express a preference for human decision-makers even if these are perceived as rule-bound bureaucrats without moral discretion. Second, we opted for the lottery voting mechanism to determine the distribution rule. Yet, it is an open question if other mechanisms, e.g., a majority vote, would have been perceived as more or less fair and whether this would have changed the preference for the algorithm. Empirical research should address this question to gain further insights into the circumstances under which algorithms are accepted or not. Third, our findings are derived in the context of income distribution in light of work performance where the information that is available to the human decision-maker and to the algorithm is limited. Other ethically relevant situations or multidimensional performance measures that include indications for efficiency or creativity might produce different results. For the purpose of identifying a reason for algorithm aversion beyond intransparency, it was crucial to apply a straightforward and rule-bound algorithm which was not based on machine learning. It should be emphasized, however, that the fact that participants preferred the human decider with moral discretion over the rule-bound algorithm does not imply that people do not appreciate transparency in algorithms. It might well be that while in our experiment transparency was not considered decisive in human decision-making in the context of morally charged decisions, it could still be perceived as an essential feature in algorithmic decision-making. In this case, we might very well expect algorithm aversion to grow even stronger once we equip the algorithm with machine learning skills. Future research can, therefore, also confront participants with more sophisticated—and thus potentially less understandable and more error-prone—algorithms and assess whether this conjecture holds true.

Appendix 1

Instructions for the experiment (translated from German)

You are now taking part in a decision experiment. Please read the instructions for the experiment carefully and completely. It may occur that possible questions you have while reading will be clarified after you have read the instructions completely. For the entire duration of the experiment, it is very important that you do not talk to other experiment participants. Also, your cell phones must be turned off and stowed away. Violations will result in the termination of the experiment without compensation to the participants. If there is something you do not understand, please refer to these instructions first. If you still have questions, please give us a hand signal. We will then come to your booth and answer your questions personally.

For your appearance, you will receive an allowance of **4,00 €.** During the course of the experiment, you can earn extra money. The amount of your earnings depends on your decisions or on the decisions of other participants. You will not learn the identity of the other participants at any time. Likewise, the other decision-makers will not learn your identity at any time.

All data and answers are evaluated anonymously.

When you have read the instructions, we ask you to answer some comprehension questions.

Description of the experiment

In this experiment, there are two different roles: **"Role A"** and **"Role B."** One of the two roles will be randomly assigned to you at the beginning of the experiment, with 10 participants receiving Role A and 20 participants receiving Role B. The experiment consists of three parts. **In Part 1**, a rule for distributing amounts of money between two people is determined. Then, participants in Role B make a choice (more on this later). **In Part 2**, the amounts of money are each earned by two individuals by solving tasks. **In Part 3**, the distribution of the total amounts earned are made. The following table shows when which participants make which decision.

Part of the experiment	Part 1	Choice by participant in Role B	Part 2	Part 3
Active participant	Role A		Role B	Role A/Computer
Description	Rule determination for the distribution of the generated amounts of money		Task solution and generation of the total amount	Distribution of the total amount generated

To understand the rule determination in Part 1, it is important to first understand Part 2 of the experiment. Therefore, Part 2 will be described first.

Part 2: Task solution

All participants in Role B are divided into **groups of two**. They solve tasks in which an adjusted slider on the computer screen, ranging from 0 to 100, must be moved with the mouse to exactly the middle position (value 50). If the slider is exactly in the middle, the task is solved. The participants have a total of **8 minutes** to solve as many tasks as possible. For each task solved, a participant earns **1 ECU (= 10 cents)**. After completing the tasks, **one of the two participants** in each group of two is randomly **selected** with equal probability. The amount earned by the selected participant **is doubled**. The participants in the group of two do not learn what amount the other participant has earned and whose earned amount has been doubled. The doubled amount of the randomly selected participant and the unchanged amount of the unselected participant then flow into a common pot for each group of two and represent the **total earned amount** of the group of two.

Part 1: Rule determination by participants in Role A

In Part 1 of the experiment, all participants in Role A first determine a **distribution rule**. This distribution rule determines how the total amount earned by the two participants in Role B in a group of two **should later be distributed** among these two participants. The amounts earned in each case and the doubling of the amount earned by the selected participant can be taken into account. To determine the distribution rule, the participant in Role A is presented with a choice of six possible distribution rules, from which he must select one. When all 10 participants in Role A have selected a distribution rule, one of these rules is **randomly selected**. The selected distribution rule is displayed to all other participants in Role A on their screens. Participants in Role B will not know which rule was so selected until the end of the experiment. The meaning of the selected rule for the individual participant in Role A depends on which of the following variants the participant in Role A is assigned to:

Variant 1. Participants in Role A assigned to Variant 1: The rule determined at the beginning of the experiment in Part 1 is **not binding**. All six rule options are **presented again** to the participant with Role A after he learns the total earned amounts of the group of two to which he has been assigned. He **must then select one of the six rules** and apply it to the group of two that has been assigned to him. According to this rule, the total amount earned is then actually distributed to the two participants in the group of two. The participant in Role A receives an **invariant lump sum payment of 100 ECU** himself. The rule selected at the beginning of the experiment in Part 1 remains on the screen.

Variant 2. Participants in Role A assigned to Variant 2: The **rule** determined at the beginning of the experiment in Part 1 is **not binding**. All six rule options are **presented again** to the participant in Role A after he learns the total earned amounts of the group of two to which he has been assigned. He must then **select one of the six rules** and apply it to the group of two that has been assigned to him. According to this rule, the total amount earned is then actually distributed to the two participants in the group of two. Whether the participant in Role A receives compensation or not **depends on whether he applies the rule determined in part 1**. The participants in Role A will each **receive a compensation** of ECU 100 **only if they have selected the rule determined** in Part 1. If they select any other rule, they receive nothing. The rule selected at the beginning of the experiment in Part 1 remains on the screen.

Variant 3. The groups of two from participants in Role B are **not assigned to any participant in Role A**. The total amount earned by the group of two is then **distributed according to the rule determined at the beginning of the experiment**. This means that the **rule** determined at the beginning of the experiment **is binding** and is **automatically applied** after the total amount is earned.

Choice of a variant by participants in Role B

After a distribution rule was determined in Part 1 of the experiment, all participants in Role B are now presented with **two of the three variants just described** (see above). The selected rule for distributing the total amount generated is not known to them. The participants can then **select one variant from the two variants for a fee of 1 ECU**. This 1 ECU was then deducted from the participants' final amount when it is paid out and the **participant is assigned to the desired variant**. Participants **can also choose no variant**. In this case, no fee is deducted, and they are randomly assigned to a variant. When all participants in Role B have made their choice, **pairs of two are randomly formed from participants in Role B** who have expressed the same preference for one of the two variants. Participants in Role B who have not expressed a preference are assigned to a variant so that they too are part of a pair of two. When pairs of two are formed, it may happen that at most one participant who has expressed a preference cannot be assigned to the desired variant. In this case, this participant is then assigned to the other variant (not selected by him). In this case, 1 ECU will not be deducted from the participants' final amount because his choice could not be taken into account. **At the end of the assignment all participants are informed in which variant they are.**

After all groups of two have been formed, the participants in Role A are each randomly assigned to a group of two. This also decides in which variant the participants in Role A will end up. Participants in Role A do **not** have a choice. During the assignment it can happen that not all participants in Role A can be assigned to a group of two. These participants then receive a fixed payment of 100 ECU.

Subsequently, the participants in Role B solve the tasks to generate the total amount (see part 2: Task solution).

Part 3: Distribution of the total amount earned

In Part 3, the **distribution of the total amount earned in the groups of two** is decided. In Variants 1 and 2, the participants in Role A learn what individual amount was earned by each of the two participants in the group of two and whose amount was randomly doubled. In addition, the rule determined in Part 1 remains visible on the screen of everyone in Role A. As a reminder, in **Variant 1**, the participant in Role A receives a **lump sum payment of 100 ECU** regardless of whether or not he selects the rule determined in Part 1. In **Variant 2**, the participant in Role A receives **a compensation of 100 ECU** only if he selects the rule determined in Part 1. In **Variant 3**, the total amount earned is **automatically distributed according to the rule** determined at the beginning of the experiment.

This part of the experiment is finished. Before the experiment starts, you can use comprehension questions to check whether you have understood the instructions.

Appendix 2

Distribution rules that could be chosen by deciders at stage 1 of the experiment

1. Each participant in the group of two should receive exactly the amount he or she has earned. The additional amount resulting from the doubling of one participant's earned amount shall remain entirely with the participant whose amount was randomly doubled.

2. Each participant in the group of two shall receive exactly the amount he or she has earned. The additional amount resulting from the doubling of the amount earned by one participant shall be divided equally between the two participants.

3. Each participant in the group of two should receive exactly the amount that he or she has earned. The additional amount resulting from the doubling of the amount earned by one participant should also be divided according to the respective work performance of the participants.

4. The participant who generates the higher amount will receive the generated amounts of both participants. In addition, he or she will receive the additional amount resulting from the doubling of the amount earned by one participant. The participant who generates the lower amount does not receive anything.

5. Both participants in the group of two should receive half of the jointly earned amount. The additional amount resulting from the doubling of the amount earned by one participant should be divided equally between the two participants.

6. Both participants in group of two should receive half of the jointly earned amount. The additional amount resulting from the doubling of a participant's earned amount should be divided according to the participants' respective work performance.

Appendix 3

Comprehension questions (original text and English translation)

Original Text	Translation and Correct Solution
Teilnehmer in beiden Rollen können gegen eine Gebühr von 1 ECU entscheiden, welcher Variante Sie zugeordnet werden wollen. Antworten Sie entweder mit "Richtig" oder "Falsch".	Participants in both roles can decide which variant they want to be assigned to for a fee of 1 ECU. Answer either "True" or "False". FALSE
Nur Teilnehmer in Rolle B können aus zwei Varianten eine auswählen. Antworten Sie entweder mit "Richtig" oder "Falsch".	Only participants in role B can select one of two variants. Answer either "True" or "False". TRUE
Die Teilnehmer in welcher Rolle lösen Aufgaben, um einen Geldbetrag zu erwirtschaften?	Participants in which role solve tasks to generate an amount of money? ROLE B
Nehmen Sie an, ein Teilnehmer in der Rolle B hat 30 ECU durch das Lösen der Aufgaben erwirtschaftet. Der zweite Teilnehmer in der Gruppe hat 65 ECU erwirtschaftet und dieser Teilnehmer wird zufällig ausgewählt und sein Betrag verdoppelt. Wie hoch ist dann der Betrag, der für die Zweiergruppe in den gemeinsamen Topf fließt (=erwirtschafteter Gesamtbetrag)?	Assume that one participant in role B has earned 30 ECU by solving the tasks. The second participant in the group has earned 65 ECU and this participant is randomly selected and his amount is doubled. What is then the amount that goes into the common pot for the group of two (= total amount earned)? $30 + 2 \times 65 = 160$
Die Teilnehmer in welcher Rolle entscheiden über die Aufteilung des erwirtschafteten Geldbetrages?	Participants in what role decide how to divide the amount of money generated? Role A
In Teil 1 des Experiments bestimmen die Teilnehmer in der Rolle A eine Regel, nach der der erwirtschaftete Geldbetrag aufgeteilt werden soll. Eine der vorgeschlagenen Regeln wird danach zufällig ausgewählt. Wie groß ist die Wahrscheinlichkeit, dass diese Regel tatsächlich die Aufteilung des erwirtschafteten Geldbetrages einer Zweiergruppe bestimmt, wenn Variante 3 zur Aufteilung des erwirtschafteten Geldbetrages für diese Zweiergruppe zum Zuge kommt?	In part 1 of the experiment, participants in Role A determine a rule according to which the amount of money earned should be divided. One of the proposed rules is then chosen at random. What is the probability that this rule actually determines the distribution of the amount of money earned by a group of two if variant 3 is used to distribute the amount of money earned for this group of two? 100%
Wenn Variante 1 zur Aufteilung des erwirtschafteten Geldbetrages für eine Zweiergruppe zum Zuge kommt, muss der Teilnehmer in der Rolle A die Regel auswählen, die in Teil 1 des Experiments bestimmt wurde. Antworten Sie entweder mit "Richtig" oder "Falsch".	If variant 1 is used to divide the amount of money earned for a group of two, the participant in Role A must select the rule that was determined in Part 1 of the experiment. Answer either "True" or "False". FALSE
Die Auszahlung der Teilnehmer in der Rolle A hängt von der Höhe der erwirtschafteten Geld beträge in ihrer Zweiergruppe ab. Antworten Sie entweder mit "Richtig" oder "Falsch".	The payout of the participants in Role A depends on the amount of money generated in their group of two. Answer either "True" or "False". FALSE

Original Text	Translation and Correct Solution
Wenn die Variante 2 zur Aufteilung des erwirtschafteten Geldbetrages für eine Zweiergruppe zum Zuge kommt, kann der Teilnehmer in der Rolle A in Teil 3 eine neue Regel auswählen. Antworten Sie entweder mit "Richtig" oder "Falsch".	If variant 2 for dividing the amount of money earned for a group of two comes to pass, the participant in Role A can select a new rule in part 3. Answer either "True" or "False". TRUE
Wenn die Variante 3 zur Aufteilung des erwirtschafteten Geldbetrages für eine Zweiergruppe zum Zuge kommt, kann ein Teilnehmer in der Rolle A in Teil 3 eine neue Regel auswählen. Antworten Sie entweder mit "Richtig" oder "Falsch".	If variant 3 for dividing the amount of money earned for a group of two comes into play, a participant in Role A can select a new rule in part 3. Answer either "True" or "False". FALSE
Nehmen Sie an, die Variante 2 zur Aufteilung des erwirtschafteten Gesamtbetrages für eine Zweiergruppe kommt zum Zuge. Welche Auszahlung erhält ein Teilnehmer in Rolle A, wenn er eine Regel auswählt, die von der in Teil 1 des Experiments bestimmten Regel, abweicht.	Assume that variant 2 for dividing the total amount earned for a group of two comes into play. What payoff does a participant in Role A receive if he chooses a rule that differs from the rule determined in part 1 of the experiment? 0 ECU

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Amar, A. R. (1984). Choosing representatives by lottery voting. *The Yale Law Journal, 93*(7), 1283–1308.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives, 11*(1), 109–126.

Bettman, J. R., Johnson, E. J., & Payne, J. W. (1990). A Componential Analysis of Cognitive Effort in Choice. *Organizational Behavior and Human Decision Processes, 45*(1), 111–139.

Bigman, Y. E., & Gray, K. (2018). People Are Averse to Machines Making Moral Decisions. *Cognition, 181*, 21–34.

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology, 31*, 543–556.

Carillo, J. D., & Mariotti, T. (2000). Strategic Ignorance as a Self-Disciplining Device. *The Review of Economic Studies, 67*(3), 529–544.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research, 56*(5), 809–825.

Chander, A. (2016). The racist algorithm. *Michigan Law Review., 115*, 1023.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309.

Eil, D., & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics, 3*(2), 114–138.

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., & Schutzman, Z. (2019). Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 170–179).

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics, 10*(2), 171–178.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*(1), 35–42.

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review, 102*(1), 469–503.

Glikson, E., & Woolley, A. W. (2020). Human trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.

Gogoll, J., & Uhl, M. (2018). Rage Against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics, 74*, 97–103.

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

Grimm, P. (2010). Social desirability bias. *Wiley international encyclopedia of marketing*

Grossman, Z., & Van Der Weele, J. J. (2017). Self-Image and Willful Ignorance in Social Decisions. *Journal of the European Economic Association, 15*(1), 173–217.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383–403.

Insel, T. R. (2019). How Algorithms Could Bring Empathy Back to Medicine. *Nature, 567*(7747), 172–174.

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries, 5*(1), 38–56.

Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures*. Decisions and Designs Inc Mclean Va.

Korsgaard, C. (1983). Two Distinctions in Goodness. *The Philosophical Review, 92*(2), 169–195.

Khasawneh, O. Y. (2018). Technophobia without boarders: The influence of technophobia and emotional intelligence on technology acceptance and the moderating influence of organizational climate. *Computers in Human Behavior, 88*, 210–218.

Kim, T. W., Monge, R., & Strudler, A. (2015). Bounded ethicality and the principle that "ought" implies "can." *Business Ethics Quarterly, 25*(3), 341–361.

Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology, 33*, 487–502.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology, 31*, 611–627.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research, 46*, 629–650.

Messick, D. M. (1993). Equality as a decision heuristic. In B. A. Mellers & J. Baron (Eds.), *Psychological Perspectives on Justice: Theory and Applications* (pp. 11–31). Cambridge University Press. https://doi.org/10.1017/CBO9780511552069.003

Mittelstadt, B. (2016). Automation, Algorithms, and Politics: Auditing for Transparency in Content Personalization Systems. *International Journal of Communication, 10*, 12.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society, 3*(2), 2053951716679679.

Moss, S. E., Sanchez, J. I., Brumbaugh, A. M., & Borkowski, N. (2009). The mediating role of feedback avoidance behavior in the LMX—performance relationship. *Group & Organization Management, 34*(6), 645–664.

Sezer, O., Gino, F., & Bazerman, M. H. (2015). Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology, 6*, 77–81.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences, 28*(4), 531–541.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson-Hall.

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124–1131.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, Explainable, and Accountable AI for Robotics. *Science Robotics, 2*(6), eaan6080.

Zimmerman, M. J. (2019). Intrinsic vs. Extrinsic Value. In Zalta, Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*. Available at https://plato.stanford.edu/entries/value-intrinsic-extrinsic/. Aceessed 29 Jun 2020.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*(4), 1060–1089.