CRED
Centre de recherche
en économie
et droit

CRED WORKING PAPER n^o 2026-05

Judged by robots: Preferences and perceived fairness of algorithmic vs. human punishments

September, 2026

IRENE LOCCI* SÉBASTIEN MASSONI†

*Université Paris-Panthéon-Assas, CRED, France.
†Université de Lorraine, Université de Strasbourg, CNRS, BETA, Nancy, France

Judged by robots: Preferences and perceived fairness of algorithmic vs. human punishments*

Irene Locci[†1] and Sébastien Massoni[2]

[1]*Université Paris-Pantheon-Assas, CRED, Paris, France*
[2]*Université de Lorraine, Université de Strasbourg, CNRS, BETA, Nancy, France*

April, 2024

Abstract

Automated decision-making is increasingly prevalent, prompting discussions about AI replacing judges in court. This paper explores how machine-made sentencing decisions are perceived through an experimental study using a public good game with punishment. The study examines preferences for human versus automated punishers and the perceived fairness of penalties. Results indicate that rule violators prefer algorithmic punishment when penalty severity is uncertain and violations are significant. While human judges are typically reluctant to delegate, they are more likely to do this when they do not have discretion over the sanction level. Fairness perceptions are similar for both humans and algorithms, except when human judges choose a less severe penalty, which enhances perceived fairness.

Keywords: Algorithms ·Human-automation interaction ·Automated decision-making ·Preferences ·Perceived fairness ·Experiment

JEL-Codes: C91, O33, K10

1 Introduction

Algorithms are largely widespread in our society and impact people's lives in various ways. They can provide tailored advertisements to consumers, assist health practitioners with diagnoses, screen and select applicants during hiring processes, provide directions, and even drive cars. The role of technology in the practice of law has also evolved significantly in recent decades. Initially, computers were mainly employed to store documents, word process, and retrieve information from large datasets (Nissan 2017). Nowadays, Artificial Intelligence (AI) is widely exploited by legal practitioners. For example, attorneys use review software

*We thank Maxim Frolov and the Fédération S2CH - CNRS -PSE- UP1 - UPN for their technical support for the experiment. The authors are grateful to Nicolas Jacquemet, Marie Obidzinski, Charlotte Saucet, Roberto Galbiati and three anonymous referees for their insightful comments. This work was supported by the program FUTURE LEADER of Lorraine Université d'Excellence within the program Investissements Avenir (ANR-15-IDEX-04-LUE) operated by the French National Research Agency for S. Massoni.

[†]corresponding author: irene.locci99@gmail.com

to assess the relevance of documents for specific cases, police apply machine learning to predict future crime locations, and judges employ AI risk assessment techniques to make bail decisions (Surden 2019; Bagaric et al. 2019). Moreover, new technologies are being developed worldwide to replace human judges. For instance, Estonia announced in 2019 the project of creating an AI capable of deciding small cases with limited or no human intervention. In Beijing, an online internet court provides litigation services for internet disputes (Themeli and Philipsen 2021). Many Legal Tech companies have also emerged with this increasing demand, providing software tools such as Juristat, Lex Predict, or Lex Machina. In response to this growing trend, the European Commission for the Efficiency of Justice published the "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment" to establish principles and guidelines for the adoption of AI (Efficiency of Justice (CEPEJ) 2018).

Legal scholars are joining the discussion on the implementation of such powerful technologies. Advocates of AI in law highlight numerous benefits that technology can bring, such as correcting the cognitive biases of humans: judges are affected by fatigue, hunger, mood, and weather, which can impact their decision-making (Danziger et al. 2011; Eren and Mocan 2018; Heyes and Saberian 2019; Sourdin and Cornes 2018); many magistrates do not reveal their relationships with litigants or attorneys, and different judges may rule in opposite ways on the same matter (Kugler 2018; Ramji-Nogales et al. 2007). Other benefits of the introduction of recent technologies in law include reducing the administrative burden of cases, as shown by England and Wales in 2022 (Buckland 2023), lowering procedural time, and improving access to justice. However, the research community also shares many concerns. The introduction of automated decision-making in the context of law could lead to codified justice, changing established practices, values, and ways of thinking (Re and Solow-Niederman 2019). Additionally, algorithms are created by humans and may carry on their biases, as AI is mostly trained on previous cases, thus perpetuating historical discriminations, such as racism or sexism (Ash 2018; Malek 2022; Bagaric et al. 2019). The most discussed issues with the adoption of AI technologies are transparency and interpretability, as most software is developed by private firms protected by trade secret laws, which prevents the defendants from challenging the evidence against them (Matulionyte and Hanif 2021; Završnik 2020). This raises issues about the right to a fair trial and due process (Matulionyte and Hanif 2021; Završnik 2020).

Among the many legal debates and questions on the subject, many researchers argue about the public acceptance of machines ruling in court, given that the feelings of justice, legitimacy, usefulness, and fairness are crucial mechanisms for compliance with the law.

Recent literature distinguishes two main categories of attitudes toward technology decision-making: algorithm aversion and algorithm appreciation. Algorithm aversion is defined as reluctance to trust decisions and output coming from algorithms compared to the ones from people (Dietvorst et al. 2015). Algorithm appreciation, on the other side, is the tendency to rely more on advice coming from algorithms than individuals (Logg et al. 2019). The literature on this topic is filled with conflicting results. Filiz et al. (2022) confirmed the findings by Dietvorst et al. (2015) on algorithm aversion, showing that automated agents are not entrusted for financial decisions, while Holzmeister et al. (2023) found that people prefer to rely on algorithms for financial investments.[1]

The varying results in the literature indicate that there are different views also on the delegation of decisions affecting a third party. Gogoll and Uhl (2018) found that participants

1. See the recent reviews of Burton et al. (2020) and Mahmud et al. (2022) on the topic.

were strongly averse to delegating to AI when the outcome affected another person. However, Maasland and Weißmüller (2022) showed that people tend to delegate unpleasant tasks in the human resource domain, such as dismissals, to automated agents.

Some studies attempted to reconcile the conflicting findings by analyzing factors that may influence this dichotomy. The first is the type of task: people are more willing to rely on algorithms in computational tasks, but not in moral and social ones (Bigman and Gray 2018; Yalcin et al. 2023). The amount of information on the decision process is also crucial for the perception of the system: Angerschmid et al. (2022) found that explanations about automated processes increased users' trust in AI and perceived fairness. Another crucial aspect seems to be the outcome of the decision: if it is favorable towards the defendant, the fairness perception increases (Wang et al. 2020).

The present article aims to test attitudes towards algorithms in the judicial context of sentencing. It experimentally investigates whether the dichotomy between algorithm aversion and appreciation is a general tendency or depends on the situational characteristics listed above: the task type, the amount of information given to participants, and the outcome of the decision (severity of the punishment). Varying these features, we analyze two aspects of automated decision-making: preferences and fairness. Three questions are assessed: Would people prefer to be judged by another person or by a machine? Would judges prefer to make the punishment decision themselves or to delegate the decisions to automated agents? And how would individuals rate the fairness of decisions?

To answer these questions, we employ a public good game with punishment (Fehr and Fischbacher 2004). After the contribution to the public good, violators can choose if they prefer to be punished by an algorithm or by a human judge, and judges decide whether to delegate the decision to an algorithm or choose the sanction themselves. The punished players then rate the fairness of the punishment received. The characteristics and choice sets of automated and human punishers are made as similar as possible by design to study if the preferences are general or context-specific. Different between-subjects treatments are created to study the impact of the situational characteristics. In the first treatment both punishing agents, human and algorithm, are given a precise table to follow for the number of points to subtract for the punishment. In the second treatment, players are informed that both the algorithm and the human have a 50% chance of imposing a mild sanction and a 50% chance of imposing a severe one. The third treatment mirrors the black box of algorithms currently used in the legal domain: violators are not explained the functioning of the algorithm and are not informed about the choices the human punisher has.

The main results of this experiment show that participants prefer to be punished by an automated agent when their contribution is low and in situations of uncertainty. However, they are indifferent when they know precise rules are imposed on human and algorithmic judges. Judges are generally averse to delegate to algorithms, and the strength of this aversion depends on the task type and their risk tendencies. They delegate more when the task is automatic, meaning they have no discretion on the number of points to subtract from violators. Additionally, risk-averse judges delegate less. The perception of fairness by punished violators is connected to the judge type, the amount of information about the sanctioning rules, and the severity of the punishment. The highest fairness ratings were found when the human judge chose the mild sanction and participants knew that there were two possible levels of punishment.

The present study contributes to the literature on automated decision-making. This topic has recently attracted much interest due to the fast development of technology. Our

contribution is threefold.

First, our paper adds to the investigation of the conflicting preferences between algorithm aversion and appreciation, which have been observed in past results. In this regard, we examine the preferences of people who are affected by the decisions, as well as those making such decisions. Concerning the individuals who are affected, recent studies have identified several factors that may trigger these opposing preferences. Our treatment manipulation specifically focuses on the task type (Yalcin et al. 2023), codified as the discretion given to the judges, the amount of information provided (Hamm et al. 2023), and the outcome of the decision (Wang et al. 2020). Concerning the delegating agents, we contribute to the literature on decisions concerning third parties (Gogoll and Uhl 2018) and the task type, specifically, whether it is more or less automatic.

Second, our study analyzes the perceived fairness of decisions. The perception of fairness is a crucial concept in understanding whether the public would accept the idea of robot judges, as it has a direct impact on people's compliance with the law (Tyler 1997). Previous experiments on perceived fairness mainly involved third-party judgment and not the perception of sanctioned individuals. Here, we explore how punished violators perceive the outcome of automated and human decisions.

Third, we contribute to the study of sanctioning by automated agents by using a new methodology. While previous research on this topic has focused on human resources decisions (Maasland and Weißmüller 2022), healthcare (Longoni et al. 2019), and law (Parlangeli et al. 2023), there is a lack of experimental evidence in the legal strand, as most studies rely on surveys or scenario-based questionnaires (for example, Araujo et al. (2020) and Watamura et al. (2023)). To our knowledge, there exist no monetary incentivized experiments investigating preferences over sanctions by automated agents. By using this type of design, participants' choices can express their true preferences, which are considered more credible than self-report measures (Jauernig et al. 2022).

This article is organized as follows: Section 2 explains the details about the design of the experiment and the behavioral predictions. Section 3 displays the results. Section 4 discusses the findings and concludes.

2 Method and design

2.1 Set-up

The experiment was conducted in June 2023 using oTtree (Chen et al. 2016) and hosted online on the server of S2CH (HUMANUM). This was a between-subjects design, with each individual participating in only one of the three sessions. Subjects were recruited through the Laboratoire d'Economie Experimentale de Paris (LEEL) via ORSEE (Greiner et al. 2004). The participation fee for the study was €5. The game consisted of 20 rounds, with only one round randomly selected for payment. The conversion rate was 1 point = €0.037. On average participants were paid €10.72 (sd = 2.79) for a 90-minute experiment. Since the game was played in groups of five people, any subjects exceeding a multiple of five were directed to the payment page to receive the show-up fee. 125 participants were eligible to proceed to the game.

2.2 Experimental design

This study implements a linear public good game with punishment (Fehr and Fischbacher 2004).[2] Participants are divided into groups of five players: four contributors and one judge.

They interact over 20 periods in a pseudo-stranger design with random matching, meaning that groups are different in each round. This makes reputation creation impossible (Keser and Van Winden 2000). Each round is composed of four stages.

Contribution - This stage involves only the four contributors. At the beginning of each period, they are assigned an endowment of 200 points. Participants simultaneously, independently, and anonymously determine how much of their endowment to assign to the public good.

The sum of the contributions is then doubled and divided among the four players so that the marginal per capita return is 0.5. Hence, the payoff of player i at the end of this stage is given by: $\pi_i = 200 - c_i + 0.5 \sum_{i=1}^{4} c_i$ where π_i represents the payoff of player i and c_i the contribution of player i. The parameters chosen ensures that players face the social dilemma where a full contribution is socially efficient while a zero contribution is individually optimal.

Violator's punishment choice - After the contribution stage, all the participants who do not contribute the full endowment are considered as violators and have to be punished. This feature is due to the need to have a large amount of data related to punishment. Violators have to decide whether they want to be punished by an algorithm or they prefer the punishment decision to be made by another person (Figure 1).

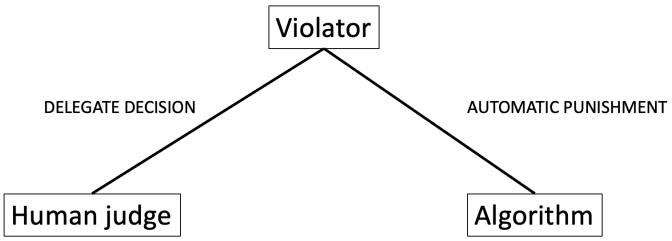


Figure 1: Sequence of the game from the point of view of the contributors.

Punishment - Until now, the judge is passively watching the game as an external observer. The choice of having a centralized third party *super partes* punisher was made to mirror the computer, in order to have the most similarities and less confounding factors about the preferences. That is also the reasoning behind the punishment being costless for external observers. Their payoffs were flat ones of 7.40 € (200 points) and not affected by the group contributions to ensure the absence of any emotional involvement.

At the punishment stage, judges intervene only if someone chooses to be punished by them. In this condition, judges can choose whether to decide the punishment themselves or to delegate the decision to an algorithm. Figure 2 represents the sequential nature of the game from the judge's point of view.

2. The experimental instructions, translated from French, can be found in Appendix B.

The wording used for the instructions given to participants is precisely chosen to allow the violators to perceive the human and computers to be equal. They read: "After the contribution, participants can choose whether to be punished automatically by the computer or delegate the decision to an external observer". This language choice allows to keep an information asymmetry necessary to study in the same setting on one side, the preference of punishers to delegate or not, on the other side, the preference of the players between humans and algorithms without the confounding factor of them knowing that the judge has also the possibility to delegate.

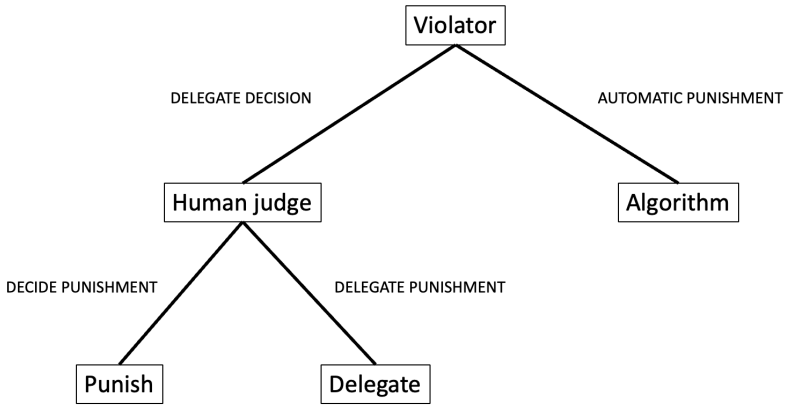


Figure 2: Sequence of the game from the point of view of the external observer judges.

The payoff of contributors after this stage is given by: $\pi_i = 200 - c_i + 0.5 \sum_{i=1}^{4} c_i - f_i$ where f_i represents the monetary punishment.

Fairness rating - Contributors are revealed their punishment level and have to rate how fair they consider it on a Likert scale from 1 (completely unfair) to 7 (completely fair).

Survey - At the end of the last round, participants have to complete a survey about demographics and attitudes toward artificial intelligence. The demographics may impact on the preferences regarding AI as proven among others by Gnambs and Appel (2019). For the attitudes, we used the General Attitudes towards Artificial Intelligence Scale (GAAIS) provided by Schepman and Rodway (2023) translated in French. It is composed of 20 items, 12 in the Positive and 8 in the Negative GAAIS. The items are scored with a Likert scale from 1 (strongly disagree) through 5 (strongly agree). The scale has good internal consistency and discriminant validity (Schepman and Rodway 2023).

2.3 Treatments

Three between-subject treatments were implemented to study how the rules used for the punishment impacted on the dependent variables. Each treatment had two levels of severity for the punishment.

The severity of the punishment- The same two severity tables are used for the punishment in all the treatments to enhance the comparison. They only vary in the punishment level for each amount contributed; we created a mild and a severe table. The implementation

of two pre-defined tables for punishment is due to the need to allow the same level of factual fairness between algorithm and human. Every level of contribution had a certain level of punishment, following the rule of proportionality: higher amount of subtracted points for higher distance from the full contribution of 200 points. The principle of proportionality has proven to be a good indicator of fairness and it has been used to study the difference in the perception of fairness between a robot's and a human's decision before (see Kim and Phillips 2021). The levels are presented in Table 1. The sanctions were set at a high level to reflect the seriousness of breaking the law, as typically the economically rational choice is to free-ride in a public good game. Note that a full contribution is now the optimal choice at the individual level for standard 1 and a standard 2.[3]

Levels of punishment		
Contribution	Mild	Severe
0	195	250
10	185	240
20	175	230
30	165	220
40	155	210
50	145	200
60	135	190
70	125	180
80	115	170
90	105	160
100	95	150
110	85	140
120	75	130
130	65	120
140	55	110
150	45	100
160	35	90
170	25	80
180	15	70
190	5	60
200	0	0

Table 1: Tables used for the levels of punishment.

Treatment 1, Guidelines - The first treatment is designed as a control. Both the algorithm and the human are forced to follow one of the tables for the punishment, which are revealed to the contributors before the beginning of each round. In this treatment, participants are given ten random rounds of the mild sanctions and the other ten rounds, the severe ones. Whether the first round uses the mild or severe punishment is randomized to avoid any order effect. Then, subjects will alternate between mild and severe punishments at each round.

Hence, subjects are fully informed before each round about the rules and amounts implemented, and there is no uncertainty about the number of points subtracted for each level of contribution.

3. Payoffs are equivalent in standard 1 between a full contribution of 200 and a contribution of 190.

Human punishers are only asked to choose between punishing themselves or delegating the task. The output of this choice does not affect the amounts subtracted from the violators. Therefore, punishers have no discretion or intentionality regarding the severity of the sanction.

Treatment 2, Probabilities - The second treatment includes some levels of uncertainty and intentionality. The participants are fully informed about the functioning of the algorithm: it punishes with 50 % probability using the mild table and with 50 % probability using the severe table. The human, if designated as a punisher from the violators, has the choice of whether to use the first or the second table or to delegate.

The standards are implemented at a group level. That is, the human has just one choice of levels of punishment for the whole group, and the computer randomly picks the table once and applies it to all the violators of the group.

Treatment 3, Black Box - The last treatment does not include any information to the contributors about the rules used for the punishment. They can only make a choice between being punished by the computer or delegating the decision to the external observer. This condition has the maximum level of uncertainty and intentionality of the human punisher in the eyes of the violators. However, the functioning is the same as in the second treatment to enhance the comparison with the previous one. That is, the algorithms works in the same way and observers have the same choices when given the possibility to punish (delegate, mild sanction, severe sanction).

2.4 Behavioral predictions

This main research question of this study concerns people's attitudes and preferences toward an automated judge. Three sub-questions are analyzed: whether people prefer to be punished by a human or an algorithm; whether the judges would be willing to delegate their decisions to an algorithm; and how fair violators consider the decisions' outputs. We rely on previous literature to make a set of behavioral predictions regarding these three questions.

The preference for the punishing agent is measured by the binary choice violators make after the contribution between the algorithm or human. At this stage of the game subjects are only influenced by the treatment characteristics, meaning the amount of information they have on the rules for the punishment.

The first set of predictions regarding this variable is:

Prediction 1a *When violators know that both humans and algorithms are imposed a strict guideline to follow for the punishment (Treatment 1), violators are indifferent.*

This particular condition of Treatment 1 mirrors an automated and mechanical task, as both agents have specific tables to follow. People are more willing to rely on algorithms and display less algorithm aversion in computational tasks (Bigman and Gray 2018; Lee 2018).

Prediction 1b *When violators know that the human punishers can choose the severity of the punishment (Treatment 2), violators prefer humans to algorithms.*

This prediction is based on the study by Jauernig et al. (2022), that shows that human discretion is preferred to automation even when algorithms' rules are created by humans.

Finally, the last prediction on violator's preferences concerns the amount of information given to subjects. Recent studies found that people are averse to relying on automated agents as they do not understand their functioning (Yeomans et al. 2019).

Prediction 1c *When violators do not know about the rules for the punishment (Treatment 3), they prefer human punishers to algorithms. This effect is more significant than in*

Treatment 2.

The same considerations about algorithm aversion apply to the judges when they are deciding whether to delegate or not. Gogoll and Uhl (2018), in particular, found that participants were strongly averse to delegating to AI when the decision affected a third party. However, Lee (2018) showed how algorithm aversion decreases with automated tasks. In Treatment 1, the judges' choice is binary between "punish" or "delegate," so it is an automated task, while in Treatments 2 and 3, they get to decide the level of severity.

The prediction regarding their behaviors is the following:

Prediction 2 *Human punishers are generally averse to delegating the decision to the algorithm. However, they delegate more when the task is automatic and they cannot choose the severity of the punishment (Treatment 1).*

The last set of predictions concerns the perceived fairness of punishment. Fairness ratings are assessed after the subjects know the amount subtracted from them for the sanction. They are thus linked to three factors: the amount of information (coming from the specific treatment), the agent making the choice (which violators was selected in the previous step), and the severity of the punishment (which they were informed of just before the fairness rating). The predictions about this assessment are:

Prediction 3a: *When the rules for the punishment are pre-defined (Treatment 1), there is no difference in fairness ratings between humans and algorithms or between more severe or less severe sanctions.*

Prediction 3b: *When a specific rule for punishment is not imposed (Treatments 2 and 3), participants rate human punishers as fairer than algorithms. This effect is larger in Treatment 3.*

This predictions are based on findings about task type: individuals find that machines or other humans in mechanical tasks are equally fair, but people judge humans as fairer in social tasks (Lee 2018). The last part of Prediction 3b is based, as Prediction 1c, on the aversion to automated agents when their functioning is not understood (Yeomans et al. 2019)

Predicition 3c: *When a specific rule for punishment is not imposed (Treatments 2 and 3) participants rate mild punishments as fairer.*

Wang et al. (2020) showed how fairness is related to the outcome of the decision: if it is in favor of the person rating, fairness increases.

3 Results

3.1 Descriptive statistics

Of the 125 participants, 3 were absent for more than 10 rounds. Their data were eliminated, alongside the rounds in which they interacted.

This study's analysis concerns three variables: the violator's choice of punishing agent, the judge's choice of whether to delegate the punishing decision, and the fairness ratings by violators of the punishment received. Table 2 presents the descriptive statistics for these variables of interest.

The percentage of violators, i.e. who did not contribute their full endowment is as follows: 49.6% in Treatment 1, 49.8% in Treatment 2 and 39.8% in Treatment 3. Among these violators, their choice for the punishing agent was presented to participants as binary (being punished automatically by an algorithm or delegating the decision to a human judge), but they also had the chance to wait for the timer and leave the page blank. Only 3 out of the

821 choices of violators were left blank (0.4%). These entries were eliminated, and thus, the variable was treated as binary. In treatment 1, the human was chosen 51.4% of the time; in treatment 2, 41.3% of the time, and 44.1% in treatment 3.

The judge's choice was binary in the first treatment (punish or delegate) and included three items in the other two treatments (mild sanction, severe sanction, delegate). This design was important for analyzing the perceived fairness of violators but not for studying delegation. Thus, this variable was treated as binary for the analysis of delegation preferences. Judges delegated with a frequency of 32.6% in Treatment 1, and this percentage dropped to 7.8% in Treatment 2 and 11.7% in Treatment 3.

The fairness ratings are measured with a 7-item Likert scale. The means of fairness for the punishment coming from a human were 4.12, 4.21, and 3.95, respectively, for Treatment 1, Treatment 2, and Treatment 3. The algorithms' fairness means were 3.92 in Treatment 1, 3.83 in Treatment 2, and 3.45 in Treatment 3.

Descriptive statistics			
	Treatment 1	Treatment 2	Treatment 3
Violators(%)	49.6 N = 560	49.7 N = 580	39.8 N = 640
Choice of human as punisher (%)	51.4 N = 265	41.3 N = 284	44.1 N = 252
Delegation by judge (%)	32.6 N = 89	7.8 N = 90	11.7 N = 77
Mean fairness rating			
Human	4.12 (sd = 2.8)	4.21 (sd = 2.8)	3.95 (sd = 2.8)
Algorithm	3.92 (sd = 2.97) N = 265	3.83 (sd = 2.78) N = 284	3.45 (sd = 2.63) N = 252

Table 2: Descriptive statistics of the variables of interest.

Given our number of observations, we rely on two-sided non-parametric tests for statistical comparisons. Binary variables (choice of punisher and judge's decision on delegation) are compared using a Binomial test while fairness ratings are compared using a Wilcoxon rank sum test. Treatments effects are based on Kruskal Wallis test followed by a Dunn test for post hoc analysis with Bonferroni correction for multiple comparisons. The regressions are OLS and Probit ones with standard errors clustered at the individual level.

3.2 Choices of violators

This section reports the analysis performed on the first variable: the choice made by the violator concerning the punishing agent type. The results are displayed in Figure 3 and are related to set of Predictions 1.

Prediction 1a was confirmed as there is no difference in Treatment 1 in the percentages of players choosing the algorithm or the human punisher (48.6% vs. 51.4%, p = 0.6747). Prediction 1b and 1c were not validated as we observe in both Treatment 2 and 3, that participants chose the algorithm more than the human judge (58.7% vs. 41.3%, p = 0.0037

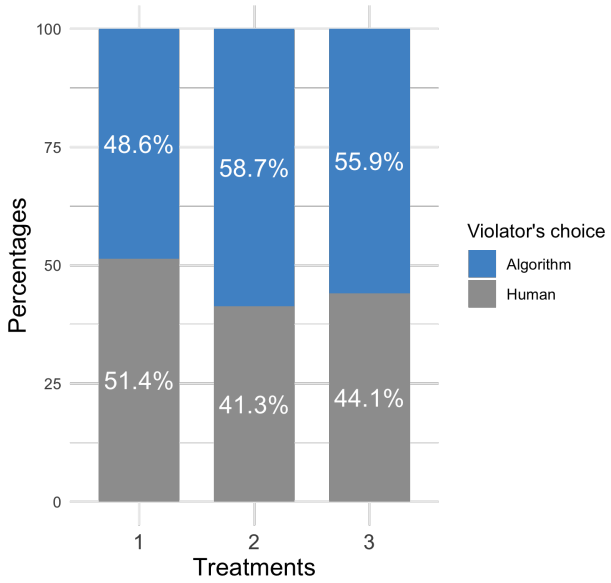


Figure 3: Composition of violators' choices for the punishing agent between human judge or algorithm.

in Treatment 2; 55.9% vs. 44.1%, p = 0.0686 in Treatment 3). The second part of Prediction 1c was also not confirmed, as there is no treatment effect comparing Treatment 2 and Treatment 3 (mean difference = 2.8%, p = 0.7639).

Our research question does not concern the amount contributed to the public good, but we also analyzed this variable.[4] We observed that the percentage of violators was significantly lower in Treatment 3 compared to Treatments 1 and 2 (T3 vs T1: 0.398 vs. 0.496, z = 3.396, p = 0.0010; T3 vs T2: 0.398 vs 0.498, z = 3.432, p = 0.0009).

A Probit regression for the choice of the violator showed a significant correlation with the player's contribution (dx/dy = 0.002, p < 0.01) and level of studies (dx/dy = -0.028, p < 0.05). It is presented in the Table 6.

3.3 Delegation by judges

For the analysis of the choice made by the external observers, we compared the proportion of time judges decide to delegate to the algorithm in each treatment. The results are displayed in Figure 4 and are related to Prediction 2.

The first part of Prediction 2 was confirmed as judges in all treatments present an aversion for delegating to the algorithms. In Treatment 1, delegation was significantly lower than punishing (32.6 % vs. 67.4 %, p = 0.0013). Treatment 2 showed an aversion to delegation (7.8% vs 92.2%, p < 0.0001). That is also confirmed in Treatment 3 (11.7% vs 88.3%, p < 0.0001)

For the second part, we compared the delegation proportions across treatments. The treatments effects were statistically significant across the three modalities (chi-square =

4. See Appendix A for supplementary results on the amount contributed and the percentage of violators according to the treatments.

Table 3: Marginal effects from a Probit regression on the choice of the violators between being punished by the human judge or by the algorithm.

	(1) Choice of punishing agent dy/dx (se)
Treatment 2	-.0744
	(.0763)
Treatment 3	.0080
	(.0776)
Attitudes towards AI	-.0296
	(.0511)
Round	-.0046
	(.0034)
Age	-.0011
	(.0026)
Female	-.0193
	(.0625)
Risk seeking	-.0186
	(.0137)
Level of studies	-.0281**
	(.0142)
Contribution	.0016***
	(.0004)
Observations	804
log(likelihood)	-525.8458
Wald $\chi^2(8)$	30.57
Prob > χ^2	0.0004
Pseudo-R^2	0.0512

Standard errors in parentheses clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

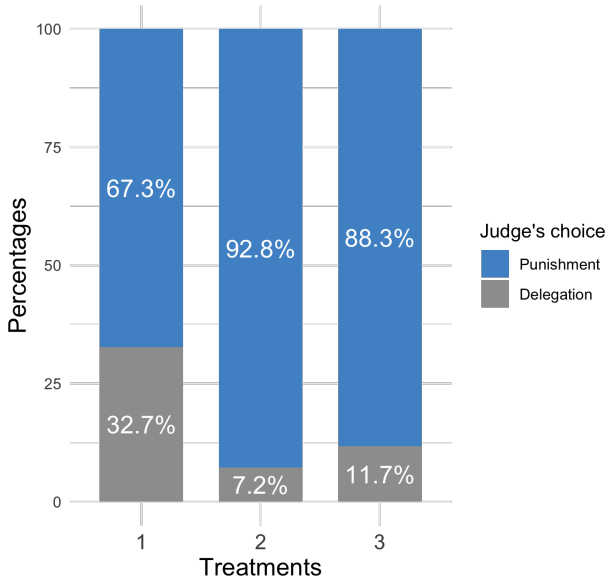


Figure 4: Composition of judges' choices between delegating or choosing the sanction themselves.

9.370, df = 2, p = 0.0092). Post-hoc pairwise comparisons showed that the proportion of delegations in Treatment 1 was significantly higher compared to the two other treatments (T1 vs T2: +24.8%, p < 0.0001; T1 vs T3: +20.9%, p = 0.0006). We did not observe a significant difference between Treatments 2 and 3 (T2 vs T3: -3.9%, p = 0.7634). This confirmed Prediction 2 as a whole.

We performed a Probit regression with standard errors clustered at the individual level to investigate further the factors influencing judges' decisions to delegate. The marginal effects are reported in Table 4 and confirm the treatment effects and the decreasing delegation (Treatment 2: dx/dy = -0.340, p < 0.001; Treatment 3: dx/dy = -0.272, p = 0.05). Additionally, the delegation was negatively related to the level of studies (dx/dy = -0.080, p = 0.035), and positively related to the player's risk seeking tendency (dx/dy = 0.055, p = 0.003), which was measured with the question about willingness to take risk in general by Dohmen et al. (2011). We found no effect of the group level of contribution to the public good (dx/dy = 0.0002, p = 0.207).

Table 4: Marginal effects from a Probit regression of delegation choices by judges.

	(1) Delegation by judges dy/dx (se)
Treatment 2	-0.3400***
	(.0686)
Treatment 3	-0.2723***
	(.0.0974)
Attitudes towards AI	.0823
	(.0765)
Round	.0002
	(.0035)
Age	.0071
	(.0064)
Female	.0588
	(.0900)
Risk seeking	.0549***
	(.0187)
Level of studies	-.0800***
	(.0380)
Total contribution of the group	.0002
	(.0002)
Observations	256
log(likelihood)	-93.8343
Wald $\chi^2(8)$	45.80
Prob > χ^2	< 0.0001
Pseudo-R^2	0.2116

Standard errors in parentheses clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.4 Fairness ratings

The last section analyzes the perceived fairness of the punishment by the violators. Results are related to the set of Predictions 3.

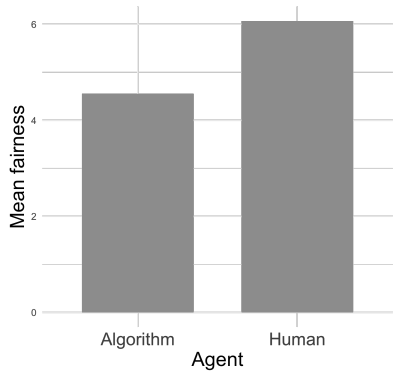
Prediction 3a concerning the lack of difference in fairness rating in Treatment 1 was confirmed. For its first part, the difference in ratings between individuals punished by humans and algorithms was not statistically significant (3.92 vs. 4.12, z = -0.538, p = 0.5905). For its second part, we did not observe a statistically significant difference in fairness ratings between severe and mild punishment (3.75 vs. 4.22, z = -1.382, p = 0.1670).

We could not confirm Prediction 3b about humans being perceived as fairer in Treatments 2 and 3. Judgments of fairness for human punishments were not different from those for algorithmic punishments (T2: 4.21 vs. 3.83, z = 1.072, p = 0.2839; T3: 3.95 vs. 3.45, z = 1.347, p = 0.1508).

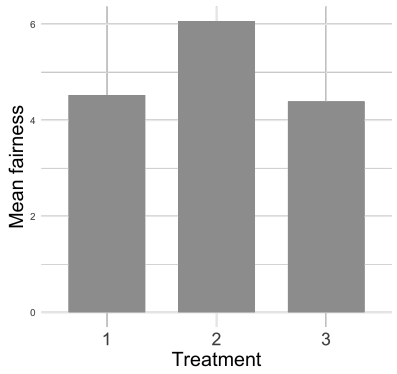
Lastly, Prediction 3c about the severity of the sanction in Treatments 2 and 3 was confirmed. Mild sanctions were perceived as fairer than severe ones in both treatments (T2: 5.04 vs. 2.72, z = 6.974, p < 0.0001; T3: 4.40 vs. 2.72, z = 4.572, p < 0.0001).

Since our predictions do not concern variables across treatments, we do not perform Kruskal-Wallis tests neither a regression for fairness.

The second and third treatments did not yield results in the analysis of differences in fairness between humans and computers. We complemented this result with a detailed comparison by grouping the participants in treatments combining severity levels and punishing agent types. This procedure created 12 groups (3 treatment $\times$ 2 severity levels $\times$ 2 punishing agents), among which we report here the most relevant pairs to study.[5] In Treatment 2, participants rated humans significantly fairer than algorithms when the mild sanction was chosen (6.07 vs. 4.40, z = -3.76, p = 0.0002). Furthermore fairness ratings about human judgments when the mild sanction is chosen in the Treatment 2 were significantly higher than those in Treatment 1 and 3 (T2 vs T1: 6.07 vs. 4.53, z = 3.463, p = 0.0008; T2 vs T3: 6.07 vs. 4.55, z = 3.4008, p = 0.0010). This comparison is shown in Figure 5b.



(a) Mean fairness for algorithms vs human judges with the mild sanction in Treatment 2.



(b) Mean fairness for human judges choosing the mild sanction across treatments

Figure 5: Comparisons of mean fairness when the mild sanction is chosen.

Table 5 summarizes our main results and their comparisons to our behavioral predictions.

5. All the other analyzed pairs showed no difference. The related values can be found in Table 8 and in Table 9 in Appendix A.

	Predicted	Observed
Preferences		
Treatment 1	Indifference	Indifference
Treatment 2	Human	Algorithm
Treatment 3	Human	Algorithm
Delegation		
Within treatments	No delegation	No delegation
Between treatments	Treatment 1	Treatment 1
Fairness		
Treatment 1	No difference	No difference
Treatment 2	Human	No difference
Treatment 3	Human	No difference
Between punishment	Mild	Mild

Table 5: Predicted and observed results on our three main variables of interest (choice of punishing agent, delegation by judge, fairness ratings).

4 Discussion

As technology advances, algorithms are increasingly replacing humans. This study explores, with an economically incentivized experiment, people's attitudes towards automated decision-makers, specifically in the legal domain. We employed a public good game where non-contribution is strongly penalized. In this setup, free-riding is less rational than in typical public good games, which emulates the implementation of a law.

The study found that participants were indifferent between human judges and algorithms when they knew that both had to follow a fixed standard for deducting points. This confirms our first prediction and aligns with the findings of Lee (2018), who observed that individuals do not have a preference for either machines or humans in managerial decisions if the task is mechanical.

Based on this research, we also predicted a preference for human punishers in the other treatments since we expected people to prefer human discretion (Jauernig et al. 2022). However, in these situations they preferred to be penalized by an algorithm rather than a human. This paper is not the first to fail to confirm algorithm aversion. Holzmeister et al. (2023) also found that people are willing to rely on automated agents over human advisors for financial portfolio decisions. A recent study conducted by Hou and Jung (2021) found that the way human or AI decision-makers are described can influence people's perception of them. The framing of agents in various ways can yield different outcomes in the same situation, suggesting that the dichotomy between algorithm appreciation and aversion may be a phenomenon guided by heuristics. Specifically, they found competence to be an influencing factor. In court, real judges are trained and experienced. On the other hand, our study used a neutral description of the situation and the agents, which likely reduced the perceived differences between algorithms and humans, leading to a preference for the computer's objectivity in the second treatment. In fact, Yalcin et al. (2023) showed that algorithms are perceived as more objective and the algorithm aversion decreases when the situation is perceived as technical or uncomplicated (compared to ones involving emotional complexities). Another explanation for this algorithm preference may be the stigma associated with breaking the rules. Previous

studies in the healthcare domain showed that individuals feel less stigma when interacting with automated agents (An and Boman 2021; Miles et al. 2021). In our experiment, the punishment by humans may be perceived as higher than merely the monetary sanction, being accompanied by the moral judgment of the person implementing the punishment.

The findings of Treatment 3 are noteworthy as we anticipated more algorithm aversion based on the research indicating that understanding the processes is crucial for perceiving automated agents. This treatment emulated a black box mechanism, where participants were unaware of the rules for punishment and were only informed that they could be sanctioned by either a human or a computer. The fact that participants still preferred algorithms in the third treatment contradicts the literature emphasizing the significance of explanation (Alam and Mueller 2021; Hamm et al. 2023; Araujo et al. 2020; Angerschmid et al. 2022).

We did not hypothesize a connection between the choice of punishing agent made by violators and their contributions. However, we observed that those who made lower contributions (and therefore violated more) tended to choose the algorithm judge more often. This finding suggests once again that people perceive algorithms as more objective than humans. Participants likely believed that human judges would base their punishment level on the contribution, whereas the algorithm always chooses with a 50% chance, regardless of the amount contributed.

We also studied the preferences of the external observer judges. Our findings revealed that these judges generally dislike delegation, which is consistent with the results of Gogoll and Uhl (2018). However, in Treatment 1, where the rules were pre-established and objective, judges delegated significantly more. In this treatment, the judges were only required to click either "punish" or "delegate" without having any discretion on the amount of punishment that was subtracted from violators. This treatment was designed to simulate automatic and repetitive tasks. Our study supports the previous literature, which suggests that aversion to delegation decreases when the task is mechanical and automatic (Castelo et al. 2019) and that aversion to delegating tasks to algorithms is due to the perceived loss of control (Dietvorst et al. 2018). In Treatment 1, judges had no control whatsoever on the level of sanctions compared to the discretion allowed in Treatments 2 and 3. Treatments 2 and 3 involved uncertainty about the punishment that the algorithm could assign, so external observers did not delegate in these situations and took the matter into their own hands. Indeed, our probit regression found a positive correlation between delegation and risk-seeking. Additionally, there was no difference between Treatment 2 and 3 in the delegation proportions. These conditions were designed to study the impact of the amount of information on violators but were identical to external observers. Thus, their similar behaviors in these two treatments are consistent with the aforementioned explanations. The fact that delegation is not related to the group level of contribution confirms our hypothesis that it relates to the characteristics of the judge and the discretion allowed in the decision.

Regarding preferences, we cannot conclude in favor of either algorithm appreciation or aversion. Violators preferred the objectivity of algorithms for their sanctions, while judges displayed aversion to delegation to algorithms. We must acknowledge that our design did not include the option for judges to delegate the decision to a human. Our results may only reflect the general aversion to delegation found in the literature (Bobadilla-Suarez et al. 2017), without implicating algorithm aversion.

The paper also studies the perceived fairness of decision outputs. The first analysis is on the difference between outputs of human and automated decisions. The study found no differences in perceived fairness in all treatments, confirming Prediction 3a but not

Prediction 3b. This is not surprising for Treatment 1, as participants knew from the beginning the parameters used for their sanction and that the human punishers had no discretion in the choice. The results of the other treatments, which included more uncertainty but not emotional aspect, are consistent with the findings of Lee (2018): people find decisions made by humans or by algorithms equally fair in non-emotional tasks. The study also predicted a difference in fairness based on the sanctioning level, as the psychological literature suggests that outcomes favorability impacts perception in the AI domain (Wang et al. 2020). This was confirmed as the main driver of perceived fairness when there was uncertainty about the specific number of points to be subtracted. However, this significance disappears when the rule is pre-defined, which is consistent with the first part of Prediction 3a and with Prediction 3c.

Interesting results about perceived fairness are shown when combining treatment, agent type, and punishment severity. Even though algorithms were preferred, especially by participants with lower contributions, Treatment 2 shows a peak in the fairness ratings when comparing the mild sanction between the human and the algorithm. In this condition, players were aware that there was a general 50% chance of receiving a more or less severe punishment. However, when choosing the human judge, they knew the punishment received was a consequence of the willing choice of a human, not just a simple probability distribution. The fairness results from the second treatment align with the literature about individuals tending to experience more intense emotions when the decision outcome comes from humans rather than from an automated system (Leyer and Schneider 2019) and moral attributions are less strong for algorithms (Gamez et al. 2020). Computers are perceived to have less intentionality (Candrian and Scherer 2022) and to be rule followers, to just "do what they are supposed to do" (Binns et al. 2018). When individuals do not have discretion in their decision-making and are required to apply a rule, they are perceived as fair and objective as the computer. However, this changes when humans have the opportunity to choose between two levels of punishment. They are expected to be harsher for higher violations, but when the outcome is favorable towards the punished individual (mild sanction), humans are perceived as fair and benevolent. This mechanism is confirmed by the difference with the same condition in Treatments 1, when the sanctioning level was imposed, and Treatment 3, when they were not informed about the probabilities.

Despite the contributions of this study to the algorithmic perception literature, this research project is subject to some limitations.

First, this paper focused on understanding the attitudes and preferences of punished individuals towards humans and computers. However, we did not assess how the other participants, such as full contributors, would perceive the outcomes. It is possible that violators' considerations of fairness could be biased since they are the ones targeted by the punishment. This means that fairness ratings could be used to express disapproval for the money subtracted from them, rather than being an expression of fairness as a concept involving parity of treatment. Additionally, we did not assess the participants' familiarity with automated systems, which could influence their perceptions.

Second, the computer's implementation of the two punishment rules was based on a binomial distribution, which had a 50% chance of occurring. This feature made it easy to compare the computer's decision-making with human decision-making. However, in future studies, we suggest creating an algorithm trained on previous human decisions in similar circumstances. This approach could help us better understand people's perceptions of technology perpetuating human biases.

Last, the study has neglected a significant trend in the literature on artificial intelligence for the sake of simplicity. For example, Bigman and Gray (2018) portrayed different attitudes towards advice coming from computers. Further research should investigate the attitudes towards using algorithms as advisors for the punishment, instead of fully delegating the decision to them. AI could be used only as an indicator of what an average punisher would do based on the data, leaving the ultimate ruling to the human. This is relevant as the widespread opinion is for AI to only assist judges, instead of completely substituting them (Buocz 2018; Xu 2022).

This study's purpose is to experimentally explore how people perceive a robot judge and expand on previous research. It provides additional evidence that algorithm aversion and appreciation are not general attitudes towards automated processes, but rather, they are influenced by contextual factors.

The study suggests that people's perceptions of humans and computers as judges vary depending on the level of information provided about the decision-making process and rules. When people are aware of just one rule to follow, they view algorithms and humans equally. However, when there are multiple options, algorithms apply standards and are preferred as punishing agents. Meanwhile, humans are believed to have discretion in their choices. They are believed to punish harder for higher violations, so they are considered fairer when they choose a less severe punishment.

The future of the legal system is rapidly approaching with the introduction of automated decision-making. The speed at which technology is developing makes it crucial to increase our knowledge on this subject. As shown, discussion on the matter has raised several ethical concerns that need to be addressed. These include human rights, biases and fairness, explainability, and public perceptions. By understanding the challenges involved, we can ensure that automated decision-making is implemented fairly and safely in the legal domain.

References

Alam, Lamia, and Shane Mueller. 2021. "Examining the effect of explanation on satisfaction and trust in AI diagnostic systems." *BMC medical informatics and decision making* 21 (1): 178.

An, Lam, and Laura Boman. 2021. "The Robot Won't Judge Me: How AI Healthcare Benefits the Stigmatized: An Abstract." In *Academy of Marketing Science Annual Conference-World Marketing Congress,* 363–364. Springer.

Angerschmid, Alessa, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. "Fairness and explanation in AI-informed decision making." *Machine Learning and Knowledge Extraction* 4 (2): 556–579.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. "In AI we trust? Perceptions about automated decision-making by artificial intelligence." *AI & society* 35:611–623.

Ash, Elliott. 2018. "Judge, Jury, and EXEcute file: The Brave New World of Legal Automation."

Bagaric, Mirko, Dan Hunter, and Nigel Stobbs. 2019. "Erasing the bias against using artificial intelligence to predict future criminality: algorithms are color blind and never tire." *U. Cin. L. Rev.* 88:1037.

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

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions." In *Proceedings of the 2018 Chi conference on human factors in computing systems,* 1–14.

Bobadilla-Suarez, Sebastian, Cass R Sunstein, and Tali Sharot. 2017. "The intrinsic value of choice: The propensity to under-delegate in the face of potential gains and losses." *Journal of risk and uncertainty* 54:187–202.

Buckland, Robert. 2023. "AI, Judges and Judgment: Setting the Scene." *M-RCBG Associate Working Paper Series.*

Buocz, Thomas Julius. 2018. "„Artificial Intelligence in Court." *Legitimacy Problems of AI Assistance in the Judiciary". Retskraft–Copenhagen Journal of Legal Studies* 2 (1): 41–59.

Burton, Jason W, Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A systematic review of algorithm aversion in augmented decision making." *Journal of behavioral decision making* 33 (2): 220–239.

Candrian, Cindy, and Anne Scherer. 2022. "Rise of the machines: Delegating decisions to autonomous AI." *Computers in Human Behavior* 134:107308.

Castelo, Noah, Maarten W Bos, and Donald R Lehmann. 2019. "Task-dependent algorithm aversion." *Journal of Marketing Research* 56 (5): 809–825.

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

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* 144 (1): 114.

———. 2018. "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them." *Management science* 64 (3): 1155–1170.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the european economic association* 9 (3): 522–550.

Efficiency of Justice (CEPEJ), European Commission for the. 2018. "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment." *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment.*

Eren, Ozkan, and Naci Mocan. 2018. "Emotional judges and unlucky juveniles." *American Economic Journal: Applied Economics* 10 (3): 171–205.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and human behavior* 25 (2): 63–87.

Filiz, Ibrahim, Jan René Judek, Marco Lorenz, and Markus Spiwoks. 2022. "Algorithm aversion as an obstacle in the establishment of robo advisors." *Journal of Risk and Financial Management* 15 (8): 353.

Gamez, Patrick, Daniel B Shank, Carson Arnold, and Mallory North. 2020. "Artificial virtue: The machine question and perceptions of moral character in artificial moral agents." *AI & SOCIETY* 35:795–809.

Gnambs, Timo, and Markus Appel. 2019. "Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe." *Computers in Human Behavior* 93:53–61.

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

Greiner, Ben, et al. 2004. *The online recruitment system orsee 2.0-a guide for the organization of experiments in economics.* Technical report.

Hamm, Pascal, Michael Klesel, Patricia Coberger, and H Felix Wittmann. 2023. "Explanation matters: An experimental study on explainable AI." *Electronic Markets* 33 (1): 1–21.

Heyes, Anthony, and Soodeh Saberian. 2019. "Temperature and decisions: evidence from 207,000 court cases." *American Economic Journal: Applied Economics* 11 (2): 238–265.

Holzmeister, Felix, Martin Holmén, Michael Kirchler, Matthias Stefan, and Erik Wengström. 2023. "Delegation decisions in finance." *Management Science* 69 (8): 4828–4844.

Hou, Yoyo Tsung-Yu, and Malte F Jung. 2021. "Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 1–25.

Jauernig, Johanna, Matthias Uhl, and Gari Walkowitz. 2022. "People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency." *Philosophy & Technology* 35 (1): 2.

Keser, Claudia, and Frans Van Winden. 2000. "Conditional cooperation and voluntary contributions to public goods." *scandinavian Journal of Economics* 102 (1): 23–39.

Kim, Boyoung, and Elizabeth Phillips. 2021. "Humans' Assessment of Robots as Moral Regulators: Importance of Perceived Fairness and Legitimacy." *arXiv preprint arXiv:2110.04729*.

Kugler, Logan. 2018. "AI judges and juries." *Communications of the ACM* 61 (12): 19–21.

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

Leyer, Michael, and Sabrina Schneider. 2019. "Me, you or AI? How do we feel about delegation."

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

Longoni, Chiara, Andrea Bonezzi, and Carey K Morewedge. 2019. "Resistance to medical artificial intelligence." *Journal of Consumer Research* 46 (4): 629–650.

Maasland, Christian, and Kristina S Weißmüller. 2022. "Blame the machine? Insights from an experiment on algorithm aversion and blame avoidance in computer-aided human resource management." *Frontiers in psychology* 13:779028.

Mahmud, Hasan, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. "What influences algorithmic decision-making? A systematic literature review on algorithm aversion." *Technological Forecasting and Social Change* 175:121390.

Malek, Md Abdul. 2022. "Criminal courts' artificial intelligence: the way it reinforces bias and discrimination." *AI and Ethics* 2 (1): 233–245.

Matulionyte, Rita, and Ambreen Hanif. 2021. "A call for more explainable AI in law enforcement." In *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW),* 75–80. IEEE.

Miles, Oliver, Robert West, and Tom Nadarzynski. 2021. "Health chatbots acceptability moderated by perceived stigma and severity: a cross-sectional survey." *Digital health* 7:20552076211063012.

Nissan, Ephraim. 2017. "Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement." *Ai & Society* 32:441–464.

Parlangeli, Oronzo, Francesco Currò, Paola Palmitesta, and Stefano Guidi. 2023. "Asymmetries in the Moral Judgements for Human Decision-Makers and Artificial Intelligence Systems (AI) Delegated to Make Legal Decisions." In *Proceedings of the European Conference on Cognitive Ergonomics 2023,* 1–4.

Ramji-Nogales, Jaya, Andrew I Schoenholtz, and Philip G Schrag. 2007. "Refugee roulette: Disparities in asylum adjudication." *Stan. L. Rev.* 60:295.

Re, Richard M, and Alicia Solow-Niederman. 2019. "Developing artificially intelligent justice." *Stan. Tech. L. Rev.* 22:242.

Schepman, Astrid, and Paul Rodway. 2023. "The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust." *International Journal of Human–Computer Interaction* 39 (13): 2724–2741.

Sourdin, Tania, and Richard Cornes. 2018. "Do judges need to be human? The implications of technology for responsive judging." *The Responsive Judge: International Perspectives,* 87–119.

Surden, Harry. 2019. "Artificial intelligence and law: An overview." *Georgia State University Law Review* 35:19–22.

Themeli, Erlis, and Stefan Philipsen. 2021. "AI as the Court: Assessing AI Deployment in Civil Cases." *E. Themeli and S. Philipsen, AI as the Court: Assessing AI Deployment in Civil Cases, in K. Benyekhlef (ed), AI and Law. A Critical Overview, Éditions Thémis* 2021:213–232.

Tyler, Tom R. 1997. "The psychology of legitimacy: A relational perspective on voluntary deference to authorities." *Personality and social psychology review* 1 (4): 323–345.

Wang, Ruotong, F Maxwell Harper, and Haiyi Zhu. 2020. "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–14.

Watamura, Eiichiro, Tomohiro Ioku, Tomoya Mukai, and Michio Yamamoto. 2023. "Empathetic Robot Judge, We Trust You." *International Journal of Human–Computer Interaction,* 1–10.

Xu, Zichun. 2022. "Human Judges in the era of artificial intelligence: challenges and opportunities." *Applied Artificial Intelligence* 36 (1): 2013652.

Yalcin, Gizem, Erlis Themeli, Evert Stamhuis, Stefan Philipsen, and Stefano Puntoni. 2023. "Perceptions of justice by algorithms." *Artificial Intelligence and Law* 31 (2): 269–292.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. "Making sense of recommendations." *Journal of Behavioral Decision Making* 32 (4): 403–414.

Završnik, Aleš. 2020. “Criminal justice, artificial intelligence systems, and human rights.” In *ERA forum,* 20:567–583. 4. Springer.

Appendix A

Contributions and number of violators

Percentage of violators

Subsection 3.2 presents the mean comparisons between treatments. Table 6 shows the marginal effects of a Probit regression on being a violator.

Table 6: Marginal effects from a Probit regression of full contribution violations.

	(1) Violation dy/dx (se)
Treatment 1	0.1101
	(.0717)
Treatment 2	0.1787**
	(.0726)
Attitudes towards AI	-.1058*
	(.0594)
Round	-.0131***
	(.0028)
Age	-.0077*
	(.0032)
Female	.1545*
	(.0599)
Risk seeking	.0487***
	(.0123)
Level of studies	.0267
	(.0174)
Observations	1761
log(likelihood)	-1101.3929
Wald $\chi^2(8)$	52.51
Prob > χ^2	< 0.0001
Pseudo-R^2	0.0931

Standard errors in parentheses clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We can observe that Treatment 3 is still giving a lower amount of violation than Treatment 2 (dy/dx = 0.1787, p = 0.014) but this effect does not hold anymore compared to Treatment 1 (dy/dx = 0.1101, p = 0.125). We can emphasize different interesting effects on the probability of not fully contribute: it decreases over time during the experiment (dy/dx = -0.0131, p < 0.0001) and individual characteristics matter (positively for being risk seeking and being a female, negatively for begin older and having a positive attitude toward A.I.).

Amounts contributed

In terms of the amount contributed, there is no statistically significant difference between treatments (T1 = 166.05, T2 = 159.07, T3 = 159.53, with all p-values higher than 0.10). Table 7 shows the OLS regression of these amounts on the treatments and our control variables.

Table 7: OLS regression of the amounts contributed to the public good

	(1)
	player.contribution
	b/se
Treatment 2	-13.9886*
	(7.1562)
Treatment 3	-6.9854
	(7.7199)
Attitudes towards AI	7.7685
	(6.3866)
Round	1.1500***
	(.3115)
Age	.4705
	(.3285)
Female	-11.4674*
	(6.6734)
Risk seeking	-5.8013***
	(1.8295)
Level of studies	-3.0778*
	(1.7091)
Constant	166.8885***
	(29.2351)
Observations	1761
log(likelihood)	-9.56e+03
F-test	4.3412
R^2	.0789

Standard errors in parentheses clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Consistent with the probability of being a violator, we find that the level of contribution is negatively affected by the fact of being risk seeking and being a female. It also increases with time over the experiment.

Fairness ratings

Between human judge and algorithm

Treatment	Severity	z	$P > \|z\|$
1	mild	-1.164	0.2444
	severe	0.397	0.6912
2	mild	-3.756	0.0002
	severe	1.455	0.1457
3	mild	-1.128	0.2592
	severe	-0.982	0.3262

Table 8: Wilcoxon rank sum tests of fairness ratings between human and algorithm based on treatment and severity of sanction.

Conditions across treatments

Agent type	Severity	Treatments	z	$P > \|z\|$
Human	mild	1-2	-3.463	0.0008
		1-3	0.042	1.000
		2-3	3.401	0.0010
Human	severe	1-2	2.536	0.0168
		1-3	1.022	0.4605
		2-3	-1.342	0.2692
Computer	mild	1-2	-1.053	0.4386
		1-3	-0.221	1.000
		2-3	0.815	0.6228
Computer	severe	1-2	1.448	0.2214
		1-3	2.499	0.0187
		2-3	1.206	0.3419

Table 9: Dunn test for pairwise comparisons of fairness across conditions based on agent type, severity and treatments with Bonferroni correction following a Kruskal-Wallis test.

Appendix B

Instructions

You are now taking part in an economic experiment.

Your payoff will depend on your decision and the decision of the other people in your group. It is therefore important that you take your time to understand the instructions.

The instructions which we have distributed to you are for your private information. Please do not communicate with the other participants during the experiment.

Should you have any questions please ask us.

During the experiment we shall not speak of Euros, but of points. Your entire earnings will be calculated in points. At the end of the experiment one random round will be selected and the points you have earned in that round will be converted to Euros at the rate of 1 point = 0.037 €.

The experiment is anonymous and lasts 20 rounds. At the beginning of the experiment the participants will be randomly divided into groups of five. You will therefore always be in a group with 4 other participants that you will not know the identity of. In every round the composition of the group will be different from the previous round. You may play or not play with people you have already played with. For each round, there will be two stages:

Stage 1

4 people are called contributors: they will have 200 points. In the following, we shall refer to this amount as the "endowment". Your task is to decide how to use your endowment.

In particular, you have to decide how many of the 200 points you want to contribute to a project (from 0 to 200) and how many of them to keep for yourself. The consequences of your decision are explained in detail below.

Once all the players have decided their contribution to the project you will be informed about the group's total contribution, your income from the project and your payoff in this stage.

Your payoff from the first stage in each period is calculated using the following formula. If you have any difficulties, do not hesitate to ask us.

*Income at the end of the stage = Endowment of points – Your contribution to the Project + 0.5 * Total contribution to the Project*

This formula shows that your income at the end of the period consists of two parts:

1) The points which you have kept for yourself (endowment – contribution),

2) The income from the project, which equals to 50% of the group's total contribution.

The income of each group member from the project is calculated in the same way. This means that each group member receives the same income from the project.

Suppose the sum of the contributions of all group members is 600 points. In this case, each member of the group receives an income from the project of: 0.5 * 600 = 300 points.

If the total contribution to the project is 90 points, then each member of the group receives an income of: 0.5 * 90 = 45 points from the project, regardless of how much they individually contributed to the project.

You always have the option of keeping the points for yourself or contributing them to the project. Each point that you keep raises your end of period income by 1 point.

Supposing you contributed this point to the project instead, then the total contribution to the project would rise by 1 point. Your income from the project would thus rise by 0,5 * 1 = 0,5 points. However, the income of the other group members would also rise by 0,5 points each, so that the total income of the group from the project would be 2 points. Your contribution to the project therefore also raises the income of the other group members.

On the other hand, you also earn an income for each point contributed by the other members to the project. In particular, for each point contributed by any member you earn 0,5 points.

In addition to the 200 points per period, each participant receives a one-off lump sum payment of €5 at the beginning of this part of the experiment. Note that this lump sum payment should not be used to calculate the "End of period income". It will only be added to your total income from all the periods at the very end.

Stage 2

At the second stage of each period, you will be informed about how much each group member contributed individually to the project at the first stage. All the participants who did not contribute 200 points will have to be punished.

After the contribution, the participants can choose whether to be punished automatically by the computer or delegate the decision to an external observer.

One person in the group will always be the external observer. He will never contribute in stage 1, but will always have to punish in stage 2.

Example: Suppose you contribute less than 200 points, then your total income after the two stages will be:

Total income at the end of the period = Income from the 1st stage – points of the punishment

Rules for the punishment

[TREATMENT 1]

For the punishment the computer and the external observer will follow the rules given by a table. This table has a certain number of points that will be subtracted for every level of contribution. An example of the table may be the one in Table 10.

For example, if you contribute 100 and chose to be punished by the computer, the computer will automatically subtract 95 points from you as a punishment.

For example, if you contribute 100 and chose to delegate the decision to the external observer, the observer will subtract 95 points from you as a punishment.

[TREATMENT 2]

For the punishment the computer and the external observer will follow the rules given by the following table. This table has two rules that represent two possible numbers of points that will be subtracted for every level of contribution. The computer will have 50% chance of using the first punishment rule and 50% chance of using the second punishment rule. The external observer will have the choice between using the first punishment rule or the second punishment rule. The rules are displayed in Table 11.

For example, if you contribute 100 and chose to be punished by the computer, the computer will subtract 95 points with 50% probability and 150 points with 50% probability from you as a punishment.

For example, if you contribute 100 and chose to delegate the decision to the external observer, the observer will choose whether to subtract 95 points or 150 points with from you as a punishment.

[TREATMENT 3]

Contribution	Rule
0	195
10	185
20	175
30	165
40	155
50	145
60	135
70	125
80	115
90	105
100	95
110	85
120	75
130	65
140	55
150	45
160	35
170	25
180	15
190	5
200	0

Table 10: Example of the table for the punishment in Treatment 1

For example, if you contribute 100 and chose to be punished by the computer, the computer will subtract some points from you as a punishment.

For example, if you contribute 100 and chose to delegate the decision to the external observer, the observer will subtract some points from you as a punishment.

Fairness rating

ALGORITHM JUDGE

You contributed x. You decided to be punished by the computer. You have been punished y. How fair do you think the punishment was?

HUMAN JUDGE

You contributed x. You decided to be punished by the external observer. You have been punished y.

How fair do you think the punishment was?

Contribution	Rule 1	Rule 2
0	195	250
10	185	240
20	175	230
30	165	220
40	155	210
50	145	200
60	135	190
70	125	180
80	115	170
90	105	160
100	95	150
110	85	140
120	75	130
130	65	120
140	55	110
150	45	100
160	35	90
170	25	80
180	15	70
190	5	60
200	0	0

Table 11: Rules for the punishment displayed in Treatment 2