Perceptions of Justice: Assessing the Perceived Effectiveness of Punishments by Artificial Intelligence versus Human Judges

Gilles Grolleau, Murat C Mungan, Naoufel Mzoughi

Perceptions of Justice: Assessing the Perceived Effectiveness of Punishments by Artificial Intelligence versus Human Judges

Gilles Grolleau

ESSCA School of Management, Lyon, France

gilles.grolleau@essca.fr

Murat C. Mungan

Texas A&M University – School of Law, USA

mungan@law.tamu.edu

Naoufel Mzoughi (Corresponding author)

INRAE, Ecodéveloppement, Avignon, France

naoufel.mzoughi@inrae.fr

Abstract: Using an original experimental survey, we analyze how people perceive punishments generated by artificial intelligence (AI) compared to the same punishments generated by a human judge. We use two vignettes pertaining to two different albeit relatively common illegal behaviors, namely not picking up one's dog waste on public roads and setting fire in dry areas. In general, participants perceived AI judgements as having a larger deterrence effect compared to the those rendered by a judge. However, when we analyzed each scenario separately, we found that the differential effect of AI is only significant in the first scenario. We discuss the implications of these findings.

Keywords: Artificial intelligence; AI, judges; punishments; unethical acts; wrongdoings.

1. Introduction

Artificial Intelligence (AI) systems have been used increasingly at various parts and stages of the legal system, including assisting judges in making sentencing decisions. Examples include robot judges handling small claims in Estonia, use of robot mediators in Canada, AI judges in Chinese courts and AI court sentencing tests in Malaysia (Chandran, 2022). Sometimes, these AI-powered systems even have great influence on the sentences imposed on offenders (Tashea, 2017; Ryberg, 2024). Despite this, little is known about whether AI-powered systems are perceived as being more or less deterrent than human judges when they are empowered with enforcing the same laws.

We address this important research question to elucidate whether an identical punishment is perceived as a greater or lesser deterrent, i.e., leading to higher or lower rates of offending, when they are enforced through an AI device rather than a human judge. While there are several contributions emphasizing the risks of AI in judicial systems (e.g., biases, lack of transparency), the perceived deterrence of AI-based enforcement remains ill-understood compared to a human judge.

In addition to discussing the possible channels through which AI-judgements may affect perceived deterrence, we design a survey experiment where all described features remain identical, and only the sentence determination is manipulated (between AI and a human judge). To the best of our knowledge, we are the first to explore how AI-based sentences impact perceived deterrence. A better understanding of this issue is crucial, given that a major goal of punishment and sentencing is to deter would-be offenders.

It is worth noting that our analysis focuses on perceived deterrence (i.e., how much subjects think policies reduce offenses overall) as opposed to the deterrence of the experimental subjects (i.e., how likely are respondents to commit the offense under the different policies in question). These two measures differ from each other, and are both important from a policy perspective. While the deterrence of the experimental subjects may provide information about

the cost-effectiveness of the policies in reducing the offense rate among the subjects, the perceived deterrence effects may inform policy makers about the deterrence of the general public. Thus, the former measure may be a better proxy for actual deterrence when the subject pool is representative of the social group considered, while the latter measure may be a better proxy if the subject's expectations are accurate, on average, regarding the aggregate deterrence effects of the policy. Moreover, perceived deterrence effects can play a pivotal role in the process through which third-parties form expectations regarding the character traits of individuals with (and without) an offense history.[1] Therefore, perceived deterrence plays a key-role in the process through which third-parties stigmatize (known-)offenders, and therefore is relevant for designing policies like 'ban-the-box' (e.g., Mungan, 2018; Sabia et al., 2021) and expungements (e.g., Mungan, 2017; Prescott & Starr, 2019), which focus on how information regarding offenses are to be released to various parties.

As we explain below (see section 3), our methodological approach is better suited to measure the perceived deterrence effects of AI versus human judges. Due to this reason, we limit our analysis to perceived deterrence effects, and do not analyze impacts on the deterrence of the subjects of our experiments. We also explain how our analysis can be extended to study both deterrence effects (see sections 4 and 5).

The remainder of this paper is structured as follows. In Section 2, we review the existing literature and formulate our main hypothesis. Section 3 describes the experimental design. Section 4 presents and explains the main results. In section 5 we provide concluding remarks.

[1] In many analyses of the stigmatizing effects of criminal records, the difference between the average traits of people who have criminal records and who do not have criminal records plays a central role (see, e.g., Rasmussen, 1996; Fluet & Mungan, 2022, 2024). The same principle applies in other settings, where people's choices are between an act that is pro-social and one that is not, more generally, instead of a choice between committing crime and refraining from committing crime (see, e.g., Bénabou & Tirole, 2006).

2. Literature overview and hypotheses

The prior literature has noted several reasons for why AI-generated sanctions may be perceived as a greater or lesser deterrent than similar sanctions imposed by a human judge. Here, we discuss some of the most intuitive reasons proposed in the literature (e.g., Bagaric & Wolf, 2017; Stobbs et al., 2017; Sourdin, 2018; Bagaric et al., 2022; Xu, 2022).

A usual argument in favor of AI use in the judicial system is the objectivity and consistency pertaining to AI (Bagaric & Wolf, 2017, see also Stobbs et al., 2017). These features may eliminate or mitigate the influence of undesirable factors (e.g., oratory tricks, emotions) in decision making. Sourdin (2018, pp. 1128–1129) argues that "[human] judging can be influenced by a range of factors that arguably would not be present where AI is involved (…) [A]djudicative decision-making can be influenced by a range of factors that can influence substantive justice. These include a range of impacts on the decision-maker that include: when and what a person has eaten; the time of day; how many other decisions a person has made that day (decision fatigue); personal values; unconscious assumptions; reliance on intuition; the attractiveness of the individuals involved; emotion."

In a similar vein, Bagaric et al. (2022, p. 146) state that "another advantage of computerized decision-making is that it necessarily makes decision-making more consistent and predictable – assuming that relevant integers are transparent. Hutton has noted that "one of the main aims of using computer technology to support sentencing has been to make the sentencing process more formal and more rational," and thereby to "reduce disparities" by ensuring that sentencing decisions are consistent with one another. Computerized decision-making has the potential to achieve consistency between sentences imposed on offenders for similar crimes. Feelings, emotions and subjective preferences cannot influence computerized decision-making. And, as Susskind observes, "computer systems will not suffer from 'off-days' that so often inhibit the performance of human beings." Indeed, lacking human irrationality,

there is no reason for computers to deviate from a consistent approach to decision-making."[2] (see also Sanghvi et al., 2022)

According to this view, the justice system may be perceived as being fairer and more predictable, and this can make the consequences of actions clearer. According to simple models of law enforcement, clearer mappings between actions and whether punishment is imposed have the potential of increasing deterrence (see Lando & Mungan, 2018 and the sources reviewed therein). On the other hand, inconsistencies merely in the form of mean-preserving dispersions of sanctions have the potential of increasing the deterrence of risk-averse individuals (see, e.g., Polinsky & Shavell, 1999; Mungan & Klick 2015).

This greater objectivity can also reduce favoritism and discrimination (Bagaric et al., 2019; Kleinberg et al., 2022). As we noted above, AI algorithms are often considered as being impartial and objective: they evaluate cases based on predetermined criteria rather than personal opinions or external factors. Thus, Bagaric et al. (2019, p. 1077) posit that "programs and algorithms need to be designed so that they do not include any integers that contain implicit bias. Once the programs and algorithms have been developed, there would be no scope for extraneous, racial considerations to have an impact on computerized sentencing decisions. As long as the data and the algorithm are transparent, then we can ensure greater consistency and fairness in judicial decision-making and can eradicate discrimination." As noted in Rizer and Watney (2018), this impartiality can lead to a perception of justice being served without favoritism or discrimination.[3] Intuitively, this can cause an increase in perceived deterrence. However, as noted in other contexts, a system that eliminates statistical discrimination may

[2] The scholarship that Bagaric et al are referring to here are Hutton (1995) and Susskind (2000).

[3] We note, however, that recent studies question whether algorithms may lead to discrimination, if the data used to train them contained biases (see, e.g., Malek, 2022).

have heterogeneous deterrence effects on those who were previously favored versus discriminated against (see Mungan, 2018).

As a cautionary note, we would like to stress that AI-related promises of consistency and fairness may remain unfulfilled if the system perpetuates biases and discrimination present in the data used to train it, primarily past human sentences (van Wingerden & Plesničar, 2022). For serious crimes, replicating biases from flawed human sentences could undermine the perceived fairness and deterrent value of AI sentencing in the eyes of the public.

Thus, although the precise impact of AI judgment on deterrence is yet to be clarified, they may reduce deterrence, if they are perceived as being unable to convey the moral condemnation and message behind a sentence, or if they are perceived as compounding existing errors or biases. Similarly, if the cases on which they are trained include some discrimination bias or errors, they are likely to replicate them. So, in difficult-to-judge cases, AI algorithms may inadvertently decrease deterrence.

In addition, AI algorithms may have transparency benefits if they clearly explain the reasons behind their decisions (see Bagaric & Wolf, 2017; Stobbs et al., 2017). This can help people better understand why they received a certain sanction, thus allowing the sanction to seem fairer. When individuals perceive the punishment as fair and justified, they may view it as a greater deterrent of future wrongdoings (see, e.g., Tyler, 1988, 2008; Paternoster et al., 1997; Yasrebi-De Kom et al., 2022). For instance, Yasrebi-De Kom et al. (2022, p. 200) found that "the deterrent effect of sanction severity on misconduct was dependent on procedural justice. Increased sanction severity only deterred from subsequent misconduct when treatment was perceived as procedurally neutral to just."

An additional argument is related to the efficiency of AI. Indeed, AI algorithms can analyze vast amounts of data and make decisions quickly, potentially leading to more efficient and timely administration of justice (Bagaric & Wolf, 2017; Stobbs et al., 2017; Rizer & Watney, 2018). Bagaric and Wolf (2017, p. 655) argue that "certainly, computers could make

sentencing decisions more efficiently and swiftly than judges because they process relevant information instantaneously." The swift and decisive nature of AI-generated punishments may enhance their deterrent effect by demonstrating that consequences will follow actions promptly. Interestingly, this AI potential of timely sentencing echoes one of the findings of the literature on swiftness and punishment (Pratt & Turanovic, 2018, p. 187) that "indicates that the "celerity effect" of deterrence tends to decay when the punishment is delayed at all after the offending behavior—even if by a matter of minutes or even seconds."

While AI-generated punishments may offer certain advantages in terms of consistency, impartiality, transparency, efficiency, and objectivity, their effectiveness as deterrents ultimately depends on how they are perceived and interpreted by individuals within the justice system and society as a whole. We do not assert that AI-generated punishments will automatically deliver all these benefits, but they have the potential to achieve this, provided that they are well designed, trained and implemented (Bagaric et al., 2022). More concretely, fair AI systems will require various conditions such as training on unbiased data, algorithmic transparency, human oversight and implementation of corrective measures. To test some of the existing claims regarding the potential effects of AI on perceived deterrence, we formulate our main hypothesis as follows: *Enforcement by an AI system is perceived as being more deterrent than the enforcement of the same law by a judge*.

3. Experimental design

3.1. Participants

209 individuals participated in our study (58% female; M_{age}=32 years old). They were contacted by students attending a course in environmental economics, and consisted of other fellow students in their university, friends, family members, etc. Individuals were invited by an email asking them to click on a link redirecting them to the survey. Participants were informed that

their participation was voluntary and without a monetary incentive.[4] Despite some criticisms (e.g., sampling bias, lack of generalizability), such convenience samples were shown to provide appropriately reliable data (e.g., Boeri & Lamonica, 2015; Mullinix et al., 2015; Underhill, 2019; Krupnikov et al., 2021; see also Grolleau et al., 2022), especially when the researcher is launching a new research agenda. Although we cannot fully rule-out some common biases related to using non-representative samples, we contend that our findings may serve as good starting point to analyze the perceived deterrence effects of AI systems.

3.2. Procedure and design

We considered two scenarios describing two illegal acts. The first scenario described an individual who was cited for refusing to pick up the waste of his/her dog from a public road. The second scenario described an individual who was cited for setting fire to dry grass. In order to test our hypothesis, we used four treatments in a 2 (dog waste or setting fire) x 2 (conviction generated by a judge vs. AI) between-subjects design. Each participant was randomly assigned to one of the four treatments. The full experimental survey is available in Appendix A. The vignettes used are presented below (changes across treatments are emphasized):

> Dog waste vignettes: For health reasons, dog waste is prohibited on sidewalks and public roads. For refusing to pick up his/her dog's waste, the judicial institution, through ***a judge who [an artificial intelligence (AI) program which]*** analyzed the offense committed,

[4] Classical criticisms about non-incentivized participants include (*i*) a possible lack of motivation, (*ii*) a hypothetical bias with participants behaving differently from real-world situations, and (*iii*) the risk of social desirability bias, where, participants do not reveal their true preferences. Nevertheless, while these concerns seem intuitive, they are not systematically supported by the available literature (see, e.g., Camerer & Hogarth, 1999; Rubinstein, 2001; El Harbi et al., 2015; Mentzakis & Sadeh, 2021).

sanctioned the guilty person to carry out 2 hours of community service (cleaning of public spaces damaged by dog waste and other dirt).

Setting fire vignettes: Last summer, a person was convicted by a court for setting fire to dry grass in the middle of summer, using a lighter. Because of this infraction, 10 hectares of a century-old forest went up in smoke. The court, ***through a judge who [artificial intelligence (AI) program which]*** analyzed this offense sentenced this person to 1 year in prison.

After reading the vignette ascribed to them, participants were asked to assess whether the considered punishment discourages individuals from leaving their pet's waste on public roads [scenario 1] or setting fire to dry grass [scenario 2]" on a 7-point Likert scale (1: Not discourages at all; 7: Highly discourages). We also collected data about some control variables. Using similar scales, participants were invited to indicate the fairness of the punishment, seriousness of the committed act, and whether people generally seek to justify themselves by finding excuses to their acts. Finally, individuals were asked to indicate some socio-demographic characteristics.

Before presenting and discussing the results, it is worth explaining why we questioned participants about their beliefs about how much the available enforcement discourages others from committing offenses, as opposed to how much it would deter them. This alternative question would have provided a source of variation only through participants who would be marginal offenders, i.e., those who would act different with and without AI. Otherwise, they would be either deterred or not deterred in both settings. This could greatly limit the number of meaningful observations, which motivated our choice. Nevertheless, it would be interesting to investigate effects on own-deterrence as opposed to perceived deterrence in follow up studies.

Another useful extension might be to screen participants on the basis of their own practices with regards to the examined scenarios.

4. Results and discussion

4.1. Main findings

We first present the mean responses by treatment of our main variable, denoted deterrence, corresponding to individuals' likelihood to perceive the punishment as likely to reduce the undesired behavior (Figure 1).

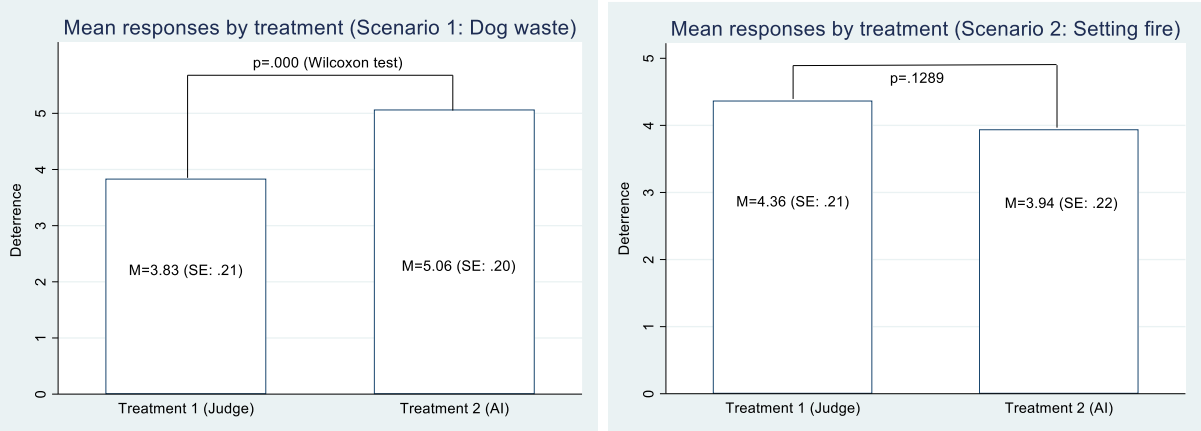

Figure 1. Mean deterrence by treatment for the whole sample and by scenario (SE stands for Standard Error)

In the first scenario (dog waste), deterrence was perceived significantly higher when the punishment is generated by AI, compared to a judge. Nevertheless, contrary to our prediction, this effect does not seem to be driven by fairness related considerations as measured in our survey. The latter is not significantly different across treatments (p-value=0.79). Moreover, no significant difference was found in the second scenario (setting fire).

Next, we analyze the effect of AI on deterrence, regardless of the scenario, that is, by pooling all observations, in a simple regression controlling for the fairness of the punishment, the seriousness of the act, whether people seek to justify themselves by finding excuses to their

acts, and individuals' characteristics. The variables used in estimation and some descriptive statistic are provided in Table 1, which also shows that our groups are not balanced in terms of age and gender. However, as we explain below, this does not appear to drive our results. Estimation results are presented in Table 2 (sample statistics and estimation results for all treatments are provided in Appendixes B and C, respectively).

Table 1. Mean values for the variables used in estimations

Variables		Treatments with a judge (N=114)	Treatment with AI (N=95)
Deterrence		4.140	4.463
Fair		5.158	5.010
Seriousness		5.517	5.474
Justification		5.017	4.842
Age (Continuous)		30.017	35.305
Gender (=1 if Female)		.482	.705
Education	Cat. 1 (*Ref*)	.035	.053
	Cat. 2	.438	.274
	Cat. 3	.438	.568
Income	Cat. 1 (*Ref*)	.228	.368
	Cat. 2	.158	.221
	Cat. 3	.412	.179
	Cat. 4	.202	.231

Education categories: baccalaureate or less, 1-3 years of university studies, and 4 years or more of university studies. Income categories: ≤ €800/month, €801-€1300/month, €1301-€2300/month, and ≥ €2301/month. ***, ** and * refer to significance at the levels of 1%, 5% and 10% respectively.

Table 2. Linear regression estimates of the effect of AI on deterrence (the two scenarios together)

Variables		Model 1 (without interaction)	Model 2 (with interaction)
Treatment	Judge (*Ref*)	.	.
	AI	.475** (.237)	.763 (1.258)
Fair		.322*** (.065)	. 316*** (.087)
Seriousness		-.008 (.085)	.059 (.129)
Justification		-.088 (.063)	.055 (.092)
Age (Continuous)		-.016* (.009)	-.015 (.009)
Gender (=1 if Female)		-.131 (.230)	-.118 (.236)
Education	Cat. 1 (*Ref*)	.	.
	Cat. 2	.270 (.353)	.277 (.357)
	Cat. 3	-.046 (.337)	-.037 (.341)
Income	Cat. 1 (*Ref*)	.	.
	Cat. 2	.394 (.331)	.383 (.336)
	Cat. 3	-.107 (.314)	-.109 (.318)
	Cat. 4	.826** (.394)	.805** (.400)
Treatment##Fair		.	.012 (.129)
Treatment##Seriousness		.	-.120 (.171)
Treatment##Justification		.	.061 (.128)
Constant		2.343*** (.747)	2.132** (1.013)
Observations		209	209
F		3.81***	3.00***
R2		.1754	.1782

Education categories: baccalaureate or less, 1-3 years of university studies, and 4 years or more of university studies. Income categories: ≤ €800/month, €801-€1300/month, €1301-€2300/month, and ≥ €2301/month. ***, ** and * refer to significance at the levels of 1%, 5% and 10% respectively. Standard Errors between brackets.

First, Table 2 shows that AI increases the perceived deterrence of punishments (H1 is supported). It is, however, likely that this positive effect is mainly driven by the high effect in scenario 1. Given that our groups are somewhat unbalanced, we tested the effect of AI in interaction with age and gender (see Appendix D). We found that AI is significant in both cases while the interaction terms are not, and thus the positive effect of AI does not appear to depend on participants' characteristics. Interestingly, the variable "Fair" is intuitively significant. In other words, participants who perceive the considered punishment as fair are more likely to perceive the punishment as deterrent. It is worth to noting that when testing the interaction between fairness and age (Appendix E), the results reveal a significant interaction, suggesting that the positive effect of fairness depends on participants' age.

Second, when considering interaction effects (Model 2), the variable AI turns out not to be significant. In addition, the three considered interaction terms (i.e., fairness, seriousness, and excuses) are not significant, which suggests that the effect of AI on deterrence does not depend

on fairness, the seriousness of the act, or individuals' likelihood to find excuses for their unethical acts, despite the fact that setting fire is perceived by our participants as significantly more serious (M=5.89) than not picking up pet's waste (M=5: p-value=0.000).

4.2. Discussion

Although one might a priori suppose that the seriousness of the violation and fairness concerns could appropriately explain the differences found across the considered scenarios, our empirical findings cast doubt on these explanations. While we cannot completely rule out these factors, particularly since our measurement strategy may be too crude to distinguish sub-dimensions that could influence our results, we propose alternative explanations that do not primarily rely on these considerations.

An important feature of the results above relates to the differential effect of AI across scenarios. A potential explanation is that the two scenarios have different moral overtones. Dog waste violations are often considered as common offenses without substantial moral implications. Like littering, they are often not committed with the purpose of causing harm and frequently driven by convenience considerations or contextual cues (e.g., existing dog waste). Unlike dog waste violations, arson is a purposeful crime carrying significant moral weight and societal condemnation. It frequently involves an intent to cause harm, and results in much more stringent punishment.

Human judges may be perceived as more legitimate moral authorities than AI systems when it comes to communicating the moral message behind a sentence, which is crucial for crimes with a strong moral dimension like arson (van Wingerden & Plesničar, M., 2022; see also Granulo et al., 2021). People may view sentences from human judges as having a stronger deterrent effect because they come from a moral agent capable of effectively conveying the rationale and societal condemnation behind the punishment. In contrast, AI systems lack the ability to engage with offenders, victims and society as moral agents. This could diminish the

perceived legitimacy and deterrent impact of AI-issued sentences, especially for situations that have a strong moral dimension.

In addition, AI is well-known to produce highly consistent sentences. On the one hand, for minor and quasi-standardized violations like dog waste violations without strong moral overtones, consistent AI sentencing could potentially be perceived as an effective deterrent by clearly establishing the consequences and norms around proper disposal of dog waste. The consistency and timeliness of AI may be viewed as an appropriate feature to deter future violations, and people may not suspect or be concerned with specific considerations such as appreciating underlying motives. On the other hand, arson often involves many elements that are situation- and individual-specific (e.g., location, weather, individual versus group crime, degree of intentionality, previous interactions with the victim(s), mental disorders, oral hearing and specific question and answer game) for which AI tools seem less well-equipped than human judges. For instance, the motives for arson are often more complex, which may eliminate the aforementioned benefits of AI judgment in the context of dog waste offenses. Kocsis (2002) has proposed six motives for arsons, namely profit, animosity, vandalism, crime concealment, political objectives and psychopathological factors. Appreciating these motives in a specific case is likely more difficult and may lead to a perception that it requires human judgment.

Moreover, the number of marginal offenders, as perceived by participants, may differ across offenses. For instance, if offenders believe that the distribution of people with regards to tendencies to set fires is bimodal with a very small intermediate group, they may believe that the mode of judgment is unlikely to have a deterrence effect, because almost everyone is inframarginal. If so, a small change in enforcement methods would not alter the behavior of many of people, and participants may have implicitly adopted this view in reporting their perceptions.

Last but not least, an additional point is related to the population likely to commit the dog waste violation *versus* arson. While the former may correspond to ordinary citizens, the

latter may be perceived as consisting of outliers, and may trigger the need of human judgment to fully appreciate some details, including possibly ones that were not surveyed or do not correspond to prior cases on which the AI was trained. For instance, in some cases, emotions can be considered as important factors in the legal process (Pillsbury, 1988). While the emotional dimension may seem irrelevant or too minor to significantly impact a dog waste case, appreciating it in an arson case can be crucial, giving human judges a comparative advantage. In these cases, human judges may be perceived as being more likely to appreciate the full range of parameters compared to AI.

In short, the reasons why AI would have increased deterrence effects in some cases and not in others remains an open issue. While we have suggested explanations such as moral overtones or the idea that humans may hold more moral authority, we lack explicit survey data to test these hypotheses. As a result, our explanations remain speculative and should be approached with caution. Moreover, other factors, such as the perception of accuracy, are also likely to play a significant role in deterrence. Whether people believe AI or human judges make more accurate judgments is critical. The deterrent effect is likely to be stronger if people think the legal system consistently convicts the guilty and acquits the innocent. In contrast, if they perceive frequent errors, leading to wrongful convictions, the deterrent effect may weaken. People may see the law as unjust or feel less bound by it due to the risk of being wrongly convicted, regardless of their behavior. Interestingly, Xu (2022) noted that "In many judicial practices, artificial intelligence has already demonstrated accuracy beyond that of human judges. But the public's demand for the accuracy of judicial artificial intelligence is higher than that of human judges." This underscores the importance of assessing public perceptions of AI versus human judges' accuracy in legal judgments in future research.

Although these explanations are intuitive, whether and to what extent they are driving the results we have presented is unclear. We list them with the hopes of guiding the design of

future studies that can be used to test them. Given the various limitations of our experiment, it should be viewed as a first step to stimulate further thinking about the issue, highlighting the need for rigorous research on this emerging topic.

5. Conclusion

The use of AI in sentencing decisions matters. Our findings are tentative but suggest that AI can impact the perception of deterrence. When (some) sentences are delivered by an AI device, the deterrence of misconduct can be affected. Nevertheless, not all AI impacts on perceived deterrence are equal. The misconduct nature can be perceived as more or less amenable to AI-based sentencing. A natural extension is to examine how the deterrence of AI sentences is influenced when subjects are informed about how the AI has been trained to reach decisions with an emphasis on procedural justice and fairness (see, e.g., Verboon & van Dijke, 2011 and Maguire et al., 2017).

Another fruitful extension that can be conducted in further experiments is to consider people's perceptions about marginal offenders: do they perceive some crimes or offenses to have bimodal distributions, and therefore a small measure of marginal offenders? Moreover, the vignettes presented in our survey only retrospectively talk about someone being caught (presumably someone who is accurately convicted as guilty). A challenging issue to test in future studies is whether deterrence would be increased or decreased if participants are informed of the actual probability that a given offense would result in conviction.

Our study has several limitations such as using a convenience sample with non-incentivized participants. While useful in exploratory research, going further by using a more representative sample or using incentive compatible methods constitute promising extensions. In future work, one can design and run an experiment similar to those which elicit social norms (Krupka & Weber, 2013). For instance, one could ask people to guess how the other participants will rank the deterrence effects and provide a monetary compensation that is decreasing in the

distance between their guess and the average response of the remaining participants. One can also use a larger sample to study the possible effect of age, gender or other demographic variables. Similarly, our survey experiment only used two typical scenarios but it makes sense to consider other ones that could correspond to various levels of seriousness. Another natural extension will be to design an experiment closer to real world conditions beyond a paper-based evaluation. Indeed, most human judges will hear evidence presentations orally which might have an impact on procedural justice perceptions (and compliance with the outcome).

Another limitation of our study is that our experimental treatments only assume true positives (a guilty person is sanctioned/convicted), but the issue of how people perceive human judges *versus* AI in relation to the rate of false positives and true/false negatives remains. To the best of our knowledge, there are no academic contributions that directly compare people's perceptions of whether AI or human judges are better at determining guilt.[5] Investigating this important issue by using experimental vignettes constitutes a promising extension.

The use of AI in legal decision-making remains a subject of debate, with various concerns about biases. While AI systems can provide data-driven sentence recommendations, the ultimate responsibility for sentencing decisions typically remains with human judges, who

[5] However, some relevant insights can be emphasized. First, there is a perceived "human-AI fairness gap" where people generally view human judges as fairer than AI judges (Chen et al., 2022), notably in situations requiring moral decision-making like determining guilt. Indeed, human judges can account for nuanced factors. Despite a promise of consistency, self-learning AI can become a "black box" lacking and can perpetuate biases present in the data it was trained on from past human judges' decisions on guilt. By contrast, human judges provide reasoning for their verdicts, though factors influencing them unconsciously remain opaque. Beyond determining guilt, an important aspect of a judge's role is communicating the moral message and societal condemnation behind a guilty verdict to the offender and public. Human judges can engage as moral agents in this process more effectively than current AI systems. In summary, the literature suggests that people currently perceive human judges as more legitimate and capable moral authorities for determining guilt, despite inconsistencies.

consider a variety of factors beyond what AI systems can assess. A challenging issue to consider in future research is the optimal combination of AI and human judges in legal decision making. Another avenue for future research is to focus on the mechanisms underlying when, why, and how AI influences deterrence. Although we have commented on some dimensions that may be relevant to this question, identifying specific mechanisms remains a challenge for future research.

References

Bagaric, M., & Wolf, G. (2017). Sentencing by computer: Enhancing sentencing transparency and predictability and (possibly) bridging the gap between sentencing knowledge and practice. *George Mason Law Review*, 25(3), 653–710.

Bagaric, M., Hunter, D., & Stobbs, N. (2019). Erasing the bias against using artificial intelligence to predict future criminality: Algorithms are color blind and never tire. *University of Cincinnati Law Review*, 88(4), 1037–1081.

Bagaric, M., Svilar, J., Bull, M., Hunter, D., & Stobbs, N. (2022). The solution to the pervasive bias and discrimination in the criminal justice system: Transparent and fair artificial intelligence. *American Criminal Law Review*, 59(1), 95–148.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678.

Boeri, M., & Lamonica, A. K. (2015). Sampling designs and issues in qualitative criminology. In, *The Routledge Handbook of Qualitative Criminology* (Edited by Heith Copes & J. Mitchell Miller), Chapter 9: 125–143, Routledge.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7–42.

Chandran, R. (2022). As Malaysia tests AI court sentencing, some lawyers fear for justice. *Reuters*, April, 12, https://www.reuters.com/article/idUSL8N2HD3V7/

Chen, B. M., Stremitzer, A., & Tobia, K. (2022). Having your day in robot court. *Harvard Journal of Law & Technology*, 36(1), 127–169.

El Harbi, S., Bekir, I., Grolleau, G., & Sutan, A. (2015). Efficiency, equality, positionality: What do people maximize? Experimental vs. hypothetical evidence from Tunisia. *Journal of Economic Psychology*, 47, 77–84.

Fluet, C., & Mungan, M. C. (2022). Laws and norms with (un) observable actions. *European Economic Review*, 145, 104129.

Fluet, C., & Mungan, M. C. (2024). Informational properties of liability regimes. *Journal of Legal Studies*, Forthcoming.

Granulo, A., Fuchs, C., & Puntoni, S. (2021). Preference for human (vs. robotic) labor is stronger in symbolic consumption contexts. *Journal of Consumer Psychology*, 31(1), 72–80.

Grolleau, G., Mungan, M. C., & Mzoughi, N. (2022). Seemingly irrelevant information? The impact of legal team size on third party perceptions. *International Review of Law and Economics*, 71, 106068.

Hutton, N. (1995). Sentencing, rationality, and computer technology. *Journal of Law and Society*, 22, 549.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10, 113–174.

Kocsis, R. (2002). Arson: Exploring motives and possible solutions. *Trends & Issues in Crime & Criminal Justice*, Australian Institute of Criminology, 236, 1–6.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495–524.

Krupnikov, Y., Nam, H. H., Style, H., Druckman, J. N., & Green, D. P. (2021). Convenience samples in political science experiments. In, *Advances in Experimental Political Science*

(Edited by James N. Druckman & Donald P. Green), Chapter 9: 165–183, *Cambridge University Press*.

Lando, H., & Mungan, M. C. (2018). The effect of type-1 error on deterrence. *International Review of Law and Economics*, 53, 1–8.

Maguire, E. R., Lowrey, B. V., & Johnson, D. (2017). Evaluating the relative impact of positive and negative encounters with police: A randomized experiment. *Journal of Experimental Criminology*, 13, 367–391.

Malek, M. A. (2022). Criminal courts' artificial intelligence: The way it reinforces bias and discrimination. *AI and Ethics*, 2(1), 233–245.

Mentzakis, E., & Sadeh, J. (2021). Experimental evidence on the effect of incentives and domain in risk aversion and discounting tasks. *Journal of Risk and Uncertainty*, 62, 203–224.

Mungan, M. C. (2017). Reducing crime through expungements. *Journal of Economic Behavior and Organization*, 137, 398–409.

Mungan, M. C. (2018). Statistical (and racial) discrimination, "ban the box", and crime rates. *American Law and Economics Review*, 20(2), 512–535.

Mungan, M. C., & Klick, J. (2015). Identifying criminals' risk preferences. *Indiana Law Journal*, 91, 791.

Paternoster, R., Brame, R., Bachman, R., Sherman, L. W., Law, S., Review, S., & Sherman, L. W. (1997). Do fair procedures matter? The effect of procedural justice on spouse assault. *Law & Society Review*, 31(1), 163–204.

Pillsbury, S. H. (1988). Emotional justice: Moralizing the passions of criminal punishment. *Cornell Law Review*, 74(4), 655–710.

Polinsky, A. M., & Shavell, S. (1999). On the disutility and discounting of imprisonment and the theory of deterrence. *Journal of Legal Studies*, 28(1), 1–16.

Pratt, T. C., & Turanovic, J. J. (2018). Celerity and deterrence. *In Deterrence, Choice, and Crime*, Volume 23 (pp. 187–210). Routledge.

Prescott, J. J., & Starr, S. B. (2019). Expungement of criminal convictions: An empirical study. *Harvard Law Review*, 133, 2460.

Rasmusen, E. (1996). Stigma and self-fulfilling expectations of criminality. *Journal of Law and Economics*, 39(2), 519–543.

Rizer, A., & Watney, C. (2018). Artificial intelligence can make our jail system more efficient, equitable, and just. *Texas Review of Law and Politics*, 23, 181.

Rubinstein, A. (2001). A theorist's view of experiments. *European Economic Review*, 45(4–6), 615–628.

Ryberg, J. (2024). Criminal justice and artificial intelligence: How should we assess the performance of sentencing algorithms? *Philosophy & Technology*, 37(1), 9.

Sabia, J. J., Nguyen, T. T., Mackay, T., & Dave, D. (2021). The unintended effects of ban-the-box laws on crime. *Journal of Law and Economics*, 64(4), 783–820.

Sanghvi, H., Ling, J. S. W., Tay, E. S., & Kuek, C. Y. (2022). Digitalisation of judiciary in Malaysia: Application of artificial intelligence in the sentencing process. In *International Conference on Law and Digitalization* (ICLD 2022) (pp. 91–97). Atlantis Press.

Sourdin, T. (2018). Judge v robot? Artificial intelligence and judicial decision-making. *University of New South Wales Law Journal*, 41(4), 1114–1133.

Stobbs, N., Hunter, D., & Bagaric, M. (2017). Can sentencing be enhanced by the use of artificial intelligence? *Criminal Law Journal*, 41(5), 261-277.

Susskind, R. (2000). *Transforming the Law: Essays on Technology, Justice and the Legal Marketplace*. Oxford University Press.

Tashea, J. 2017. Courts Are Using AI to Sentence Criminals. That Must Stop Now. Wired, April, 17. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/

Tyler, T. (2008). Psychology and institutional design. *Review of Law and Economics*, 4(3), 801–887.

Tyler, T. R. (1988). What is procedural justice-criteria used by citizens to assess the fairness of legal procedures. *Law and Society Review*, 22, 103–136.

Underhill, K. 2019. Price and Prejudice: An Empirical Test of Financial Incentives, Altruism, and Racial Bias. *Journal of Legal Studies*, 48, 245–274.

Verboon, P., & van Dijke, M. (2011). When do severe sanctions enhance compliance? The role of procedural fairness. *Journal of Economic Psychology*, 32(1), 120–130.

van Wingerden, S., & Plesničar, M. (2022). Artificial intelligence and sentencing: Humans against machines. In *Sentencing and Artificial Intelligence* (*Studies in Penal Theory and Philosophy*, Edited by J. Ryberg & J. Roberts), pp. 230–251.

Xu, Z. (2022). Human judges in the era of artificial intelligence: Challenges and opportunities. *Applied Artificial Intelligence*, 36(1), 2013652.

Yasrebi-De Kom, F. M., Dirkzwager, A. J., Van Der Laan, P. H., & Nieuwbeerta, P. (2022). The effect of sanction severity and its interaction with procedural justice. *Criminal Justice and Behavior*, 49(2), 200–219.

Appendix 1. Survey translation

Anonymous survey

In the following, we present a hypothetical scenario. We invite you to read it carefully and answer the questions. There are no true or false answers: we are only interested in your honest opinion.

Treatment 1: For health reasons, dog waste is prohibited on sidewalks and public roads. For refusing to pick up his/her dog's waste, the judicial institution, through a judge who analyzed the offense committed, sanctioned the guilty person to carry out 2 hours of community service (cleaning of public spaces damaged by dog waste and other dirt).

Treatment 2: For health reasons, dog waste is prohibited on sidewalks and public roads. For refusing to pick up his/her dog's waste, the judicial institution, through an artificial intelligence (AI) program which analyzed the offense committed, sanctioned the guilty person to carry out 2 hours of community service (cleaning of public spaces damaged by dog waste and other dirt).

Treatment 3: Last summer, a person was convicted by a court for setting fire to dry grass in the middle of summer, using a lighter. Because of this infraction, 10 hectares of a century-old forest went up in smoke. The court, through a judge who analyzed this offense, sentenced this person to 1 year in prison.

Treatment 4: Last summer, a person was convicted by a court for setting fire to dry grass in the middle of summer, using a lighter. Because of this infraction, 10 hectares of a century-old forest went up in smoke. The court, through artificial intelligence (AI) program which analyzed this offense, sentenced this person to 1 year in prison.

Please assess whether this sanction discourages individuals from leaving their pet's waste on public roads *(T3, T4: setting fire to dry grass in the middle of summer)*:

1 Not discouraging at all	2	3	4	5	6	7 Very discouraging

To what extent does this sanction seem (un)just and (un)fait to you?

1 Completely unjust and unfair	2	3	4	5	6	7 Completely just and fair

In my opinion, not picking up your pet's waste on sidewalks and public spaces *(T3, T4: deliberately setting fire to dry grass)* is an act:

1	2	3	4	5	6	7 Very serious

Not serious at all						

Generally speaking, people seek to justify themselves by finding excuses for their actions (e.g. other serious problems at the same time):

1 Fully disagree	2	3	4	5	6	7 Fully agree

Please complete the following information:

1. Age: _____years	4. Your monthly earnings: a) ≤ €800 ☐ b) Between €801 and €1300 ☐ c) Between €1301 and €2300 ☐ d) Between €2301 and €3155 ☐ e) > €3155 ☐
2. Education: French Bac or less ☐ Bac + __years ☐	
3. Gender: H. ☐ F. ☐	

Appendix B. Mean values for the variables used in estimations

		Scenario 1 (Dog waste)		Scenario 2 (Setting fire)	
Variables		Treatment 1 (Judge; N=49)	Treatment 2 (AI; N=44)	Treatment 3 (Judge; N=65)	Treatment 4 (AI; N=51)
Deterrence		3.837	5.068	4.369	3.941
Fair		5.571	5.614	4.846	4.490
Seriousness		5.082	4.909	5.846	5.961
Justification		4.898	4.409	5.108	5.216
Age (Continuous)		29.428	25.909	30.461	43.412
Gender (=1 if Female)		.571	.682	.415	.725
Education	Cat. 1 (*Ref*)	.041	.045	.061	.059
	Cat. 2	.571	.364	.338	.196
	Cat. 3	.388	.477	.477	.647
Income	Cat. 1 (*Ref*)	.306	.409	.169	.333
	Cat. 2	.204	.341	.123	.118
	Cat. 3	.449	.136	.385	.216
	Cat. 4	.041	.114	.323	.333

Education categories: baccalaureate or less, 1-3 years of university studies, and 4 years or more of university studies. Income categories: ≤ €800/month, €801-€1300/month, €1301-€2300/month, and ≥ €2301/month. ***, ** and * refer to significance at the levels of 1%, 5% and 10% respectively.

Appendix C. Linear regression estimates of the effect of AI on deterrence by treatment

		Model 1 (without interaction)		Model 2 (with interaction)	
Variables		Scenario 1 (Dog waste)	Scenario 2 (Setting fire)	Scenario 1 (Dog waste)	Scenario 2 (Setting fire)
Treatment	Judge (*Ref*)	.	.	.	.
	AI	1.131*** (.315)	-.275 (.394)	2.332 (1.616)	-2.053 (1.710)
Fair		.158 (.104)	.377*** (.089)	.144 (.107)	.380*** (.089)
Seriousness		.023 (.118)	.012 (.128)	.102 (.211)	-.005 (.173)
Justification		.251*** (.081)	-.048 (.094)	.327** (.126)	-.177 (.133)
Age (Continuous)		-.011 (.015)	-.001 (.013)	-.009 (.016)	-.003 (.013)
Gender (=1 if Female)		.093 (.305)	-.175 (.331)	-.046 (.312)	-.100 (.335)
Education	Cat. 1 (*Ref*)	.	.	.	.
	Cat. 2	-.358 (.565)	.769 (.480)	-.312 (.572)	.694 (.495)
	Cat. 3	-.859 (.536)	.650 (.449)	-.831 (.541)	.608 (.449)
Income	Cat. 1 (*Ref*)	.	.	.	.
	Cat. 2	.371 (.397)	-.422 (.572)	.405 (.403)	-.465 (.573)
	Cat. 3	-.345 (.414)	-.340 (.467)	-.374 (.420)	-.402 (.469)
	Cat. 4	.184 (.634)	.344 (.573)	.315 (.657)	.445 (.577)
Treatment##Seriousness		.	.	-.110 (.258)	.063 (.246)
Treatment##Justification		.	.	-.138 (.177)	.267 (.190)
Constant		1.364 (1.283)	3.177** (1.522)	1.731 (1.654)	3.141** (1.272)
Observations		93	116	93	116
F		3.74***	2.61***	3.17***	2.38***
R2		.3368	.2163	.3432	.2325

Education categories: baccalaureate or less, 1-3 years of university studies, and 4 years or more of university studies. Income categories: ≤ €800/month, €801-€1300/month, €1301-€2300/month, and ≥ €2301/month. ***, ** and * refer to significance at the levels of 1%, 5% and 10% respectively. The values between brackets correspond to Standard Errors.

Appendix D. Testing the effect of AI in interaction with age and gender (standard errors between brackets)

Variables		Interaction AI##Age	Interaction AI##Gender
Treatment	Judge (*Ref*)	.	.
	AI	1.236** (.563)	.616* (.363)
Fair		.317*** (.065)	.319*** (.065)
Seriousness		.004 (.085)	-.008 (.085)
Justification		.095 (.063)	.083 (.063)
Age (Continuous)		.000 (.014)	-.015* (.009)
Gender (=1 if Female)		-.088 (.231)	-.034 (.298)
Education	Cat. 1 (*Ref*)	.	.
	Cat. 2	.218 (.354)	.253 (.355)
	Cat. 3	-.078 (.337)	-.048 (.338)
Income	Cat. 1 (*Ref*)	.	.
	Cat. 2	.376 (.330)	.390 (.332)
	Cat. 3	-.089 (.313)	-.108 (.315)
	Cat. 4	.826** (.393)	.823** (.395)
Treatment##Age		-.024 (.016)	.
Treatment##Gender		.	-.237 (.463)
Constant		1.774** (.838)	2.330*** (.749)
Observations		209	209
F		3.70***	3.50***
R2		.1846	.1765

Education categories: baccalaureate or less, 1-3 years of university studies, and 4 years or more of university studies. Income categories: ≤ €800/month, €801-€1300/month, €1301-€2300/month, and ≥ €2301/month. ***, ** and * refer to significance at the levels of 1%, 5% and 10% respectively.

Appendix E. Testing the interaction between age and fairness (Dep. Variable: Deterrence)

Variables		Coefficients and significance
Treatment	Judge (*Ref*)	.
	AI	.522** (.236)
Fair		.035 (.161)
Seriousness		-.033 (.085)
Justification		.094 (.062)
Age (Continuous)		-.058** (.023)
Fair##Age		.008* (.004)
Gender (=1 if Female)		-.107 (.229)
Education	Cat. 1 (*Ref*)	.
	Cat. 2	.228 (.351)
	Cat. 3	-.129 (.337)
Income	Cat. 1 (*Ref*)	.
	Cat. 2	.386 (.329)
	Cat. 3	-.139 (.312)
	Cat. 4	.925** (.395)
Constant		3.957*** (1.116)
Observation		209
F		3.85***
R2		.1908

Education categories: baccalaureate or less, 1-3 years of university studies, and 4 years or more of university studies. Income categories: ≤ €800/month, €801-€1300/month, €1301-€2300/month, and ≥ €2301/month. ***, ** and * refer to significance at the levels of 1%, 5% and 10% respectively. The values between brackets correspond to Standard Errors.