Asymmetries in the Moral Judgements for Human Decision-Makers and Artificial Intelligence Systems (AI) Delegated to Make Legal Decisions

Oronzo Parlangeli
oronzo.parlangeli@unisi.it
DISPOC - Università di Siena
Siena, Italy

Francesco Curro'
francesco.curro2@unisi.it
DISPOC - Università di Siena
Siena, Italy

Paola Palmitesta
palmitesta@unisi.it
DISPOC - Università di Siena
Siena, Italy

Stefano Guidi
stefano.guidi@unisi.it
DISPOC - Università di Siena
Siena, Italy

ABSTRACT

The evaluation of the use of Artificial Intelligence (AI) in legal decisions may concern several factors. We structured a study conducted by administering an online questionnaire in which the participants had to consider different scenarios in which a decision-maker, human or artificial, made an unintentionally benevolent or malevolent error of judgement for offences punishable by a fine (Civil Law infringement) or years in prison (Criminal Law infringement). We found that humans who delegate AIs are blamed less than solo humans. In addition, people consider the error more serious if committed by a human being when a sentence for a crime of the penal code is mitigated, and for an AI when a penalty is aggravated for an infringement of the civil code.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

moral judgments; responsibility; equity; decision delegation; ethics; responsibility attributions

ACM Reference Format:
Oronzo Parlangeli, Francesco Curro', 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 *European Conference in Cognitive Ergonomics (ECCE '23), September 19–22, 2023, Swansea, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3605655.3605676

1 INTRODUCTION

Nowadays, the use of AI systems is becoming more and more common in the legal field [2, 10, 16, 17]. Thanks to the exponential advancement of the efficiency of these technologies, AI also found its way into trials [5, 9].

'A central worry about such deployments of AI systems concerns responsibility attributions' [13]. Efficient as they are, artificial intelligence systems are, nonetheless, systems that can lead to discriminatory results [18].

Furthermore, it remains to be clarified how we assess the seriousness of a decision error made by an AI [1, 7, 8, 19].

These issues appear only partially resolved to date since studies have often attempted to address them only one at a time without trying to analyse the complexity of their relationships. The study presented here has been structured to provide some answers to these relevant questions.

2 RELATED WORKS

The relationship of human beings with the tools that have enabled cultural evolution is absolutely special [14]. This relationship takes on particular characteristics regarding Artificial Intelligence Systems or Big Data [3]. To date it still seems unclear under what circumstances one tends to delegate an artificial agent with greater or lesser confidence, and even more so, how the consequences of this delegation are assessed [4]. As suggested by Chugunova and Sele [4], conflicting findings on this issue can be reconciled with a general willingness to include intelligent systems in decision-making processes, even to an excessive degree, though not allocating them to the decision [4]. Some experiments underline that human participants prefer human decisions over AI, even if they know that AI outperforms humans in some decision-making processes [4].

The attributions of responsibility have been studied for the *objective of the delegation*, which often involves a specific goal and implies an aspect called "Blame Avoidance": the delegation could, in fact, be used to reach a better performance, but also to circumvent the harmful consequences of a wrong decision (also defined as "strategic scapegoating") [6, 15].

Malle and colleagues [11] provided evidence that people seek to apply similar moral norms to human and artificial agents. However, it may prove challenging to translate blame placed on artificial agents into punishment. [6].

ECCE '23, September 19–22, 2023, Swansea, United Kingdom

ACM ISBN 979-8-4007-0875-6/23/09…$15.00
https://doi.org/10.1145/3605655.3605676

Table 1: List of the scenarios.

Typology of offence	Offence	Punishment range	Fine/sentence	Mitigation	Aggravation
Civil code	Running a red light	€167,00 - €665,00	€416,00	↓ to €277,00	↑ to €555,00
Civil code	Excessive speeding	€173,00 - €695,00	€434,00	↓ to €289,00	↑ to €579,00
Criminal code	Killing someone with the car	2 - 7 yrs	4 yrs 5 mths	↓ to 3 yrs	↑ to 6 yrs
Criminal code	Hit-and-run while driving	1,5 yrs - 6 yrs	3 yrs 8 mths	↓ to 2,5 yrs	↑ to 5 yrs

The major issue in assigning responsibility to artificial agents may be the distribution of responsibility across the various decision-makers involved in the process [12]. Matthias [12] underlines that the decision of the AI could not be attributed to the manufacturer, the programmer, or the operator because it is often impossible to recover the decision processes made by the machines backwards. Many studies have highlighted an asymmetry in the evaluations we make when a human being or an artificial intelligence system makes some mistakes [1, 8]. Due to these results, Hidalgo and colleagues [8] formulated the hypothesis that there is a dual mode of judgement. Artificial agents would be judged essentially by the results of their actions, more or less harmful. On the other hand, human decision-makers would be assessed essentially by the level of intentionality/accidentality of their actions.

In another study, scenarios were developed in which intelligent robots, or humans, could cause the death of either other robots or other people by detonating a bomb. Compared to humans, the results indicated that robots are more blamed in cases where their actions cause harm to other robots [7]. Even in this case, *intentionality* is crucial for attributing responsibility.

The attribution of responsibility is also affected by contextual factors that appear to influence judgement, first and foremost, the severity of the consequences caused [8] and, related to this, whether the harm affect other human beings or artificial entities [7].

3 THE STUDY

The objectives of the study concerned the following research questions:

- RQ1 Are errors made by AIs that are delegated to make legal judgments rated more or less serious than the same errors made by a human judge?
- RQ2 Are human beings and AI systems held equally responsible if they make errors of judgement?

4 METHOD

4.1 Participants

We recruited 288 participants through Prolific, a commercial online access platform (survey published on 20 Mar 2023, 17:51). Participants were not given a time limit to complete the survey.

Participants were informed of the objective of the study, and they were invited to fill in an online questionnaire voluntarily, receiving a reward for their participation (they would receive £1.5, approximately corresponding to €1.70). Participants have to express informed consent before starting to fill in the survey. The study was approved by the Ethical Board Committee for Research in Human and Social Sciences (CAREUS) of the University of Siena, Italy (act n. 04/2023). Participants' age ranged from 20 to 60 years ($M = 30.32, SD = 9.01$), with a predominance of men ($n = 164, 56.90\%$).

4.2 Design and Procedure

In structuring the questionnaire, we manipulated three factors: the decision-maker (2 levels: Human or AI decision maker), the typology of error (2 levels: mitigating or aggravating the decision) and the type of offence (2 levels defined by fine or imprisonment). The study, therefore, had a 3-way mixed experimental design. To not overload the subjects' task, eight versions of the same questionnaire were created, varying the order in which the scenarios were presented so that each participant only considered four scenarios. A counterbalanced Latin square design was used to control the order of scenarios.

Questionnaires were composed of two sections.

The first section of the survey regarded socio-demographic questions about gender (M, F, non-binary, prefer not to answer), age (by year), citizenship and level of education.

In the second section, participants were randomly assigned to one of the eight different survey conditions, thus having to consider the description of four events in which someone had committed an offence. The scenarios had all the same structure. First, the offence was described, e.g. "A person ran a red light". Then the range of possible punishment was given. This was followed by information about who decided the amount of the fine/sentence. In all scenarios, the value of the punishment was the intermediate figure between the two extremes. After reading the scenario, the participant had to answer the first question: "Do you think this penalty is fair?". The answer was on a 7-point Likert scale from "totally disagree" to "totally agree". Then, the scenario continued in a second part, in which it was reported that the decision made was wrong, either because it had aggravated the fine/years in prison or because it had mitigated the fine/years in prison.

In table 1 are reported the different scenarios, with the erroneous decision and the amount of their mitigating/aggravating effects. The wrong decision always led to an increase or decrease of the punishment that was 1/3 of what had been formerly established.

Participants then rated the severity of the error made by the decision-maker on a 7-point Likert scale, from "very little" to "very much". Another question was asked to assess the participant's opinion on the level of responsibility of the decision-maker concerning the error, in particular evaluating how much responsibility should be accounted to the AI system or/and to the human decision-maker (7-point Likert scale, from "not at all" to "completely"). Lastly, there were two questions for a self-report about knowledge in legal aspects.

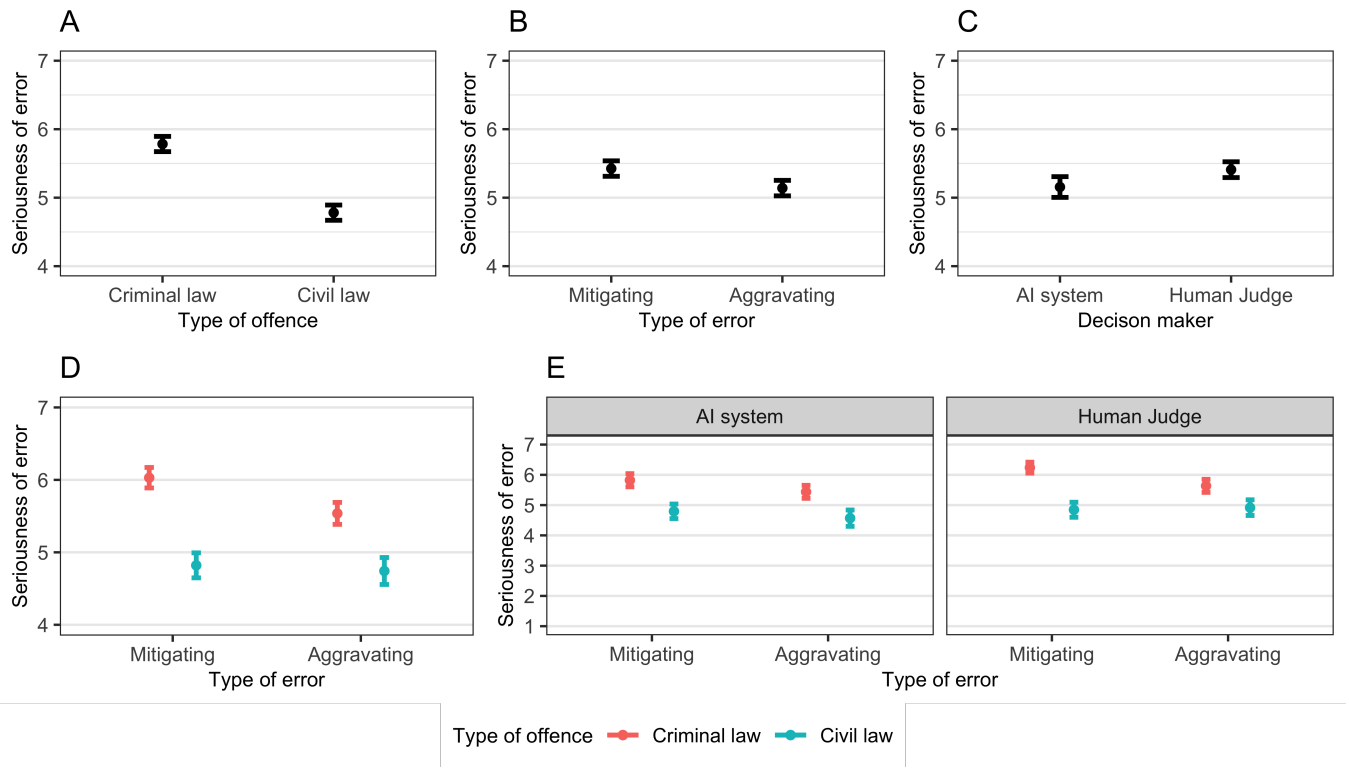

Figure 1: Plots of the marginal means of the seriousness ratings by (A) Type of offence, (B) Type of error, (C) Type of decision maker, (D) Type of error and Type of offence. In (E) the average seriousness ratings for the errors by the AI system (left panel) and by the human judged are plotted as a function of Type of error and Type of offence. Error bars are within-subject confidence intervals for the means.

5 RESULTS

5.1 Seriousness ratings

We analysed the ratings for the seriousness of the judicial error in a 3-way mixed ANOVA, including 2 within-subjects factors, each with 2 levels (*type of offence*: criminal law *vs* civil law; *type of error*: aggravating *vs* mitigating the decision), and one between-subject factor (*type of decision maker*: human judge *vs* AI system). The results for the analysis showed significant main effects of *decision-maker* ($F_{1,287} = 6.87, p = .009$), *type of offence* ($F_{1,287} = 156.63, p < .001$) and *type of error* ($F_{1,287} = 12.27, p < .001$). Moreover, the 2-way *type of offence by type of error* interaction was significant ($F_{1,287} = 5.53, p = .019$). Pairwise comparisons showed that the error was judged as significantly more serious in criminal law scenarios ($M_{criminal} = 5.8, SE = 0.06$) than in civil law ones ($M_{civil} = 4.8, SE = 0.07, t_{287} = 12.5, p < .001$), when the error was mitigating the decision ($M_{mitig.} = 5.4, SE = 0.06$) than when it was aggravating it ($M_{aggr.} = 5.1, SE = 0.06, t_{287} = 3.5, p = .0005$), and when the judicial decision was made by a human judge ($M_{judge} = 5.4, SE = 0.07$) than when it was made by an AI system ($M_{AI} = 5.2, SE = 0.07, t_{287} = 2.62, p = .009$). The analysis of the simple effects of the *type of error* for the different types of offences that we conducted following the significant interaction, however, showed that only in criminal law offences the error was judged as significantly more serious when it was mitigating the sentence ($M_{mitig.} = 6.0, SE = 0.07$) for the individual on trial than when it was aggravating it ($M_{aggr.} = 5.5, SE = 0.08, t_{287} = 4.87, p < .0001$), while for civil law offences the difference was not significant ($t_{287} = 0.57, p = .58$). The plots of the marginal means for the main effects and for the 2-way type of offence x type of error and 3-way type of scenario x type of error x decision maker interactions are presented in figure 1.

5.2 Responsibility attributions

We first analysed the ratings for the degree to which the human judge in the scenarios was considered responsible for the error in a 3-way, mixed ANOVA with the same within- and between-subjects factors included in the analysis of the severity ratings. The results showed, first of all, that the main effects were significant: Delegation ($F_{1,287} = 5.35, p = .02,$), Type of offence ($F_{1,287} = 54.3, p < .001$), Type of error ($F_{1,287} = 13.2, p < .001$) as well as the 2-way type of offence by type of error interaction, mirroring exactly the results for the severity ratings. However, if the pattern of marginal means for the effects of type of offence and of error and their interaction replicates the one found in the seriousness ratings (as confirmed by pairwise comparisons, whose details are reported on OSF), the main effect of delegating the decision to an AI system had an opposite direction: the judge was rated significantly less responsible when

the decision was delegated to the AI system than when it was made directly by the judge.

We then conducted a further 3-way ANOVA only of the data for the scenarios in which the judicial decision was delegated. The type of scenario and type of error were included as in the previous analyses as within-subjects factors along with another within-subjects factor (*Actor*) whose two levels corresponded to the degree of responsibility that each participant attributed, respectively, to the judge and to the AI system to which the judge had delegated the decision. The main effect of *Actor* was significant ($F_{1,145} = 24.1, p < .001$), and so were all the 2-way interactions *Actor x Type of offence* ($F_{1,145} = 40.9, p < .001$), *Actor x Type of error* ($F_{1,145} = 6.2, p = .014$), and *Type of offence x Type of error* ($F_{1,145} = 10.7, p = .001$). Although the 3-way interaction was not significant, since all the 2-way interactions were significant, we examined the *Type of offence x Type of error* interaction separately for ratings of the responsibility attributed to the different actors.

For the attributions to the judge, the simple effects of type of error across the different types of offences followed the same pattern found for the seriousness ratings. This pattern, however, is reversed when it comes to the attributions of responsibility to the AI system. In this case, for criminal law scenarios, the responsibility ratings did not vary significantly across types of errors (mitigating vs aggravating). In contrast, for civil law scenarios, the responsibility attributions were significantly higher when the error aggravated the sentence ($M = 4.97, SE = 0.15$) than when it mitigated it. ($M = 4.47, SE = 0.15, t_{145} = 3.46, p < .001$). Concerning the main effect of the type of actor, pairwise comparisons showed that averaging across types of scenarios and errors, the judge was rated as significantly more responsible ($M = 5.32, SE = 0.10$) than the AI system ($M = 4.55, SE = .12, t_{145} = 4.91, p < .001$).

6 DISCUSSION AND CONCLUSION

As already pointed out [19], and quite predictably, errors in judgement are rated more serious when a human being commits them (RQ1). An asymmetry that, with good reason, can be attributed to the fact that human beings are recognised as having a higher degree of intentionality [8]. Here we find that it is for human judges that the error is considered to be more serious in cases where the error implies a reduction in sentence (RQ1). This, however, is specifically attributable to cases of offences for which the criminal code is infringed (RQ1).

Reasonably linked again to the possibility of attributing agency and intentionality is the result that human decision-makers are more responsible for errors than the AIs they have delegated (RQ2). This asymmetry, however, implies a sharing of responsibility between the actors involved in the judgement. In cases where a judgement involves an AI that makes mistakes, the degree of responsibility of the delegating human is decreased.

For judgements concerning serious events, a human judge is expected to decide and to be severe (RQ2). However, judgments resulting in the determination of fines might also be dealt with by artificial agents, and benevolence is expected from them (RQ2). These results may be due to the recognition of a different capacity in knowledge processing: context-sensitive for human beings and efficient and algorithmic for artificial systems.

REFERENCES

[1] Daniele Amoroso and Guglielmo Tamburrini. 2021. The Human Control Over Autonomous Robotic Systems: What Ethical and Legal Lessons for Judicial Uses of AI? In *New Pathways to Civil Justice in Europe: Challenges of Access to Justice*, Xandra Kramer, Alexandre Biard, Jos Hoevenaars, and Erlis Themeli (Eds.). Springer International Publishing, Cham, 23–42. https://doi.org/10.1007/978-3-030-66637-8_2

[2] Sarah Barrington and Hany Farid. 2023. A comparative analysis of human and AI performance in forensic estimation of physical attributes. *Scientific Reports* 13, 1 (2023), 4784. https://doi.org/10.1038/s41598-023-31821-3

[3] Cindy Candrian and Anne Scherer. 2022. Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior* 134 (2022), 107308. https://doi.org/10.1016/j.chb.2022.107308

[4] Marina Chugunova and Daniela Sele. 2022. We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics* 99 (2022), 101897. https://doi.org/10.1016/j.socec.2022.101897

[5] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580

[6] Till Feier, Jan Gogoll, and Matthias Uhl. 2022. Hiding behind machines: Artificial agents may help to evade punishment. *Science and Engineering Ethics* 28, 2, Article 19 (2022), 19 pages. https://doi.org/10.1007/s11948-022-00372-7

[7] Stefano Guidi, Enrica Marchigiani, Sergio Roncato, and Oronzo Parlangeli. 2021. Human Beings and Robots: Are There Any Differences in the Attribution of Punishments for the Same Crimes?. In *Proceedings of the 32nd European Conference on Cognitive Ergonomics* (Siena, Italy) *(ECCE '21)*. Association for Computing Machinery, New York, NY, USA, Article 21, 6 pages. https://doi.org/10.1145/3452853.3452864

[8] César A. Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines*. MIT Press, Floor Cambridge, MA. https://doi.org/10.7551/mitpress/13373.001.0001

[9] Danielle Leah Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. *Responsive Communities Initiative, Berkman Klein Center for Internet and Society, Harvard Law School* 2, 7 (2017), 37. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041

[10] Massimo Luciani. 2018. La decisione giudiziaria robotica. *Rivista AIC* 3, 3 (2018), 22. https://www.rivistaaic.it/images/rivista/pdf/3_2018_Luciani.pdf

[11] Bertram F. Malle, Stuti Thapa Magar, and Matthias Scheutz. 2019. AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma. In *Robotics and Well-Being*, Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar (Eds.). Springer International Publishing, Cham, 111–133. https://doi.org/10.1007/978-3-030-12524-0_11

[12] Andreas Matthias. 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6, 3 (2004), 175–183. https://doi.org/10.1007/s10676-004-3422-1

[13] Lauritz Munch, Jakob Mainz, and Jens Christian Bjerring. 2023. The value of responsibility gaps in algorithmic decision-making. *Ethics and Information Technology* 25, 1, Article 21 (2023), 11 pages. https://doi.org/10.1007/s10676-023-09699-6

[14] Jordan Navarro, François Osiurak, Sandrine Ha, Guillaume Communay, Eleonore Ferrier-Barbut, Arnaud Coatrine, Vivien Gaujoux, Emma Hernout, Julien Cegarra, William Volante, and Peter A. Hancock. 2022. Development of the Smart Tools Proneness Questionnaire (STP-Q): an instrument to assess the individual propensity to use smart tools. *Ergonomics* 65, 12 (2022), 1639–1658. https://doi.org/10.1080/00140139.2022.2048895 PMID: 35243968.

[15] Daniel B. Shank, Alyssa DeSanti, and Timothy Maninger. 2019. When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society* 22, 5 (2019), 648–663. https://doi.org/10.1080/1369118X.2019.1568515

[16] Richard Susskind. 2012. Technology and the Law. In *A Companion to the Philosophy of Technology*, Jan Kyrre Berg Olsen Friis, Stig Andur Pedersen, and Vincent F. Hendricks (Eds.). Wiley-Blackwell, London, England, United Kingdom. https://philpapers.org/rec/SUSTAT-2

[17] Richard E. Susskind. 1990. Artificial intelligence, expert systems and law. *Denning LJ* 5 (1990), 105.

[18] Alexander Tischbirek. 2020. *Artificial Intelligence and Discrimination: Discriminating Against Discriminatory Systems*. Springer International Publishing, Cham, 103–121. https://doi.org/10.1007/978-3-030-32361-5_5

[19] Abigail Wilson, Courtney Stefanik, and Daniel B. Shank. 2022. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations? *Computers in Human Behavior Reports* 8 (2022), 100229. https://doi.org/10.1016/j.chbr.2022.100229

Received 17 April 2023