The Conflict Between People's Urge to Punish AI and Legal Systems

Gabriel Lima[1,2], Meeyoung Cha[1,2]*, Chihyung Jeon[3] and Kyung Sin Park[4]

[1]School of Computing, KAIST, Daejeon, South Korea, [2]Data Science Group, Institute for Basic Science, Daejeon, South Korea, [3]Graduate School of Science and Technology Policy, KAIST, Daejeon, South Korea, [4]School of Law, Korea University, Seoul, South Korea

Regulating artificial intelligence (AI) has become necessary in light of its deployment in high-risk scenarios. This paper explores the proposal to extend legal personhood to AI and robots, which had not yet been examined through the lens of the general public. We present two studies (*N* = 3,559) to obtain people's views of electronic legal personhood vis-à-vis existing liability models. Our study reveals people's desire to punish automated agents even though these entities are not recognized any mental state. Furthermore, people did not believe automated agents' punishment would fulfill deterrence nor retribution and were unwilling to grant them legal punishment preconditions, namely physical independence and assets. Collectively, these findings suggest a conflict between the desire to punish automated agents and its perceived impracticability. We conclude by discussing how future design and legal decisions may influence how the public reacts to automated agents' wrongdoings.

Keywords: artificial intelligence, robots, AI, legal system, legal personhood, punishment, responsibility

1 INTRODUCTION

Artificial intelligence (AI) systems have become ubiquitous in society. To discover where and how these machines[1] affect people's lives does not require one to go very far. For instance, these automated agents can assist judges in bail decision-making and choose what information users are exposed to online. They can also help hospitals prioritize those in need of medical assistance and suggest who should be targeted by weapons during war. As these systems become widespread in a range of morally relevant environments, mitigating how their deployment could be harmful to those subjected to them has become more than a necessity. Scholars, corporations, public institutions, and nonprofit organizations have crafted several ethical guidelines to promote the responsible development of the machines affecting people's lives (Jobin et al., 2019). However, are ethical guidelines sufficient to ensure that such principles are followed? Ethics lacks the mechanisms to ensure compliance and can quickly become a tool for escaping regulation (Resseguier and Rodrigues, 2020). Ethics should not be a substitute for enforceable principles, and the path towards safe and responsible deployment of AI seems to cross paths with the law.

The latest attempt to regulate AI has been advanced by the European Union (EU; (European Commission, 2021)), which has focused on creating a series of requirements for high-risk systems (e.g., biometric identification, law enforcement). This set of rules is currently under public and

---

[1]We use the term "machine" as a interchangeable term for AI systems and robots, i.e., embodied forms of AI. Recent work on the human factors of AI systems has used this term to refer to both AI and robots (e.g., (Köbis et al., 2021)), and some of the literature that has inspired this research uses similar terms when discussing both entities, e.g., (Matthias, 2004).

scholarly scrutiny, and experts expect it to be the starting point of effective AI regulation. This research explores one proposal previously advanced by the EU that has received extensive attention from scholars but was yet to be studied through the lens of those most affected by AI systems, i.e., the general public. In this work, we investigate the possibility of extending legal personhood to autonomous AI and robots (Delvaux, 2017).

The proposal to hold machines, partly or entirely, liable for their actions has become controversial among scholars and policymakers. An open letter signed by AI and robotics experts denounced its prospect following the EU proposal (http://www. robotics-openletter.eu/). Scholars opposed to electronic legal personhood have argued that extending certain legal status to autonomous systems could create human liability shields by protecting humans from deserved liability (Bryson et al., 2017). Those who argue against legal personhood for AI systems regularly question how they could be punished (Asaro, 2011; Solaiman, 2017). Machines cannot suffer as punishment (Sparrow, 2007), nor do they have assets to compensate those harmed.

Scholars who defend electronic legal personhood argue that assigning liability to machines could contribute to the coherence of the legal system. Assigning responsibility to robots and AI could imbue these entities with realistic motivations to ensure they act accordingly (Turner, 2018). Some highlight that legal personhood has also been extended to other nonhumans, such as corporations, and doing so for autonomous systems may not be as implausible (Van Genderen, 2018). As these systems become more autonomous, capable, and socially relevant, embedding autonomous AI into legal practices becomes a necessity (Gordon, 2021; Jowitt, 2021).

We note that AI systems could be granted legal standing regardless of their ability to fulfill duties, e.g., by granting them certain rights for legal and moral protection (Gunkel, 2018; Gellers, 2020). Nevertheless, we highlight that the EU proposal to extend a specific legal status to machines was predicated on holding these systems legally responsible for their actions. Many of the arguments opposed to the proposal also rely on these systems' incompatibility with legal punishment and pose that these systems should not be granted legal personhood because they cannot be punished.

An important distinction in the proposal to extend legal personhood to AI systems and robots is its adoption under criminal and civil law. While civil law aims to make victims whole by compensating them (Prosser, 1941), criminal law punishes offenses. Rights and duties come in distinct bundles such that a legal person, for instance, may be required to pay for damages under civil law and yet not be held liable for a criminal offense (Kurki, 2019). The EU proposal to extend legal personhood to automated systems has focused on the former by defending that they could make "good any damage they may cause." However, scholarly discussion has not been restricted to the civil domain and has also inquired how criminal offenses caused by AI systems could be dealt with (Abbott, 2020).

Some of the possible benefits, drawbacks, and challenges of extending legal personhood to autonomous systems are unique to civil and criminal law. Granting legal personhood to AI systems may facilitate compensating those harmed under civil law (Turner, 2018), while providing general deterrence (Abbott, 2020) and psychological satisfaction to victims (e.g., through revenge (Mulligan, 2017)) if these systems are criminally punished. Extending civil liability to AI systems means these machines should hold assets to compensate those harmed (Bryson et al., 2017). In contrast, the difficulties of holding automated systems criminally liable extend to other domains, such as how to define an AI system's mind, how to reduce it to a single actor (Gless et al., 2016), and how to grant them physical independence.

The proposal to adopt electronic legal personhood addresses the difficult problem of attributing responsibility for AI systems' actions, i.e., the so-called responsibility gap (Matthias, 2004). Self-learning and autonomous systems challenge epistemic and control requirements for holding actors responsible, raising questions about who should be blamed, punished, or answer for harms caused by AI systems (de Sio and Mecacci, 2021). The deployment of complex algorithms leads to the "problem of many things," where different technologies, actors, and artifacts come together to complicate the search for a responsible entity (Coeckelbergh, 2020). These gaps could be partially bridged if the causally responsible machine is held liable for its actions.

Some scholars argue that the notion of a responsibility gap is overblown. For instance, Johnson (2015) has asserted that responsibility gaps will only arise if designers choose and argued that they should instead proactively take responsibility for their creations. Similarly, Sætra (2021) has argued that even if designers and users may not satisfy all requirements for responsibility attribution, the fact that they chose to deploy systems that they do not understand nor have control over makes them responsible. Other scholars view moral responsibility as a pluralistic and flexible process that can encompass emerging technologies (Tigard, 2020).

Danaher (2016) has made a case for a distinct gap posed by the conflict between the human desire for retribution and the absence of appropriate subjects of retributive punishment, i.e., the retribution gap. Humans look for a culpable wrongdoer deserving of punishment upon harm and justify their intuitions with retributive motives (Carlsmith and Darley, 2008). AI systems are not appropriate subjects of these retributive attitudes as they lack the necessary conditions for retributive punishment, e.g., culpability.

The retribution gap has been criticized by other scholars, who defend that people could exert control over their retributive intuitions (Kraaijeveld, 2020) and argue that conflicts between people's intuitions and moral and legal systems are dangerous only if they destabilize such institutions (Sætra, 2021). This research directly addresses whether such conflict is real and could pose challenges to AI systems' governance. Coupled with previous work finding that people blame AI and robots for harm (e.g., (Kim and Hinds, 2006; Malle et al., 2015; Furlough et al., 2021; Lee et al., 2021; Lima et al., 2021)), there seems to exist a clash between people's reactive attitudes towards harms caused by automated systems and their feasibility. This conflict is yet to be studied empirically.

We investigate this friction. We question whether people would punish AI systems in situations where human agents would typically be held liable. We also inquire whether these reactive attitudes can be grounded on crucial components of legal punishment, i.e., some of its requirements and functions.

Previous work on the proposal to extend legal standing to AI systems has been mostly restricted to the normative domain, and research is yet to investigate whether philosophical intuitions concerning the responsibility gap, retribution gap, and electronic legal personhood have similarities with the public view. We approach this research question as a form of experimental philosophy of technology (Kraaijeveld, 2021). This research does not defend that responsibility and retribution gaps are real or can be solved by other scholars' proposals. Instead, we investigate how people's reactive attitudes towards harms caused by automated systems may clash with legal and moral doctrines and whether they warrant attention.

Recent work has explored how public reactions to automated vehicles (AVs) could help shape future regulation (Awad et al., 2018). Scholars posit that psychology research could augment information available to policymakers interested in regulating autonomous machines (Awad et al., 2020a). This body of literature acknowledges that the public view should not be entirely embedded into legal and governance decisions due to harmful and irrational biases. Yet, they defend that obtaining the general public's attitude towards these topics can help regulators discern policy decisions and prepare for possible conflicts.

Viewing the issues of responsibility posed by automated systems as political questions, Sætra (2021) has defended that these questions should be subjected to political deliberation. Deciding how to attribute responsibility comes with inherent trade-offs that one should balance to achieve responsible and beneficial innovation. A crucial stakeholder in this endeavor is those who are subjected to the indirect consequences of widespread deployment of automated systems, i.e., the public (Dewey and Rogers, 2012). Scholars defend that automated systems "should be regulated according to the political will of a given community" (Sætra and Fosch-Villaronga, 2021), where the general public is a major player. Acknowledging the public opinion facilitates the political process to find common ground for the successful regulation of these new technologies. If legal responsibility becomes too detached from the folk conception of responsibility, the law might become unfamiliar to those whose behavior it aims to regulate, thus creating the "law in the books" instead of the "law in action" (Brożek and Janik, 2019).

People's expectations and preconceptions of AI systems and robots have several implications to their adoption, development, and regulation (Cave and Dihal, 2019). For instance, fear and hostility may hinder the adoption of beneficial technology (Cave et al., 2018; Bonnefon et al., 2020), whereas a more positive take on AI and robots may lead to unreasonable expectations and overtrust—which scholars have warned against (Bansal et al., 2019). Narratives about AI and robots also inform and open new directions for research among developers and shape the views of both policymakers and its constituents (Cave and Dihal, 2019). This research contributes to the maintenance of the "algorithmic social contract," which aims to embed societal values into the governance of new technologies (Rahwan, 2018). By understanding how all stakeholders involved in developing, deploying, and using AI systems react to these new technologies, those responsible for making governance decisions can be better informed of any existing conflicts.
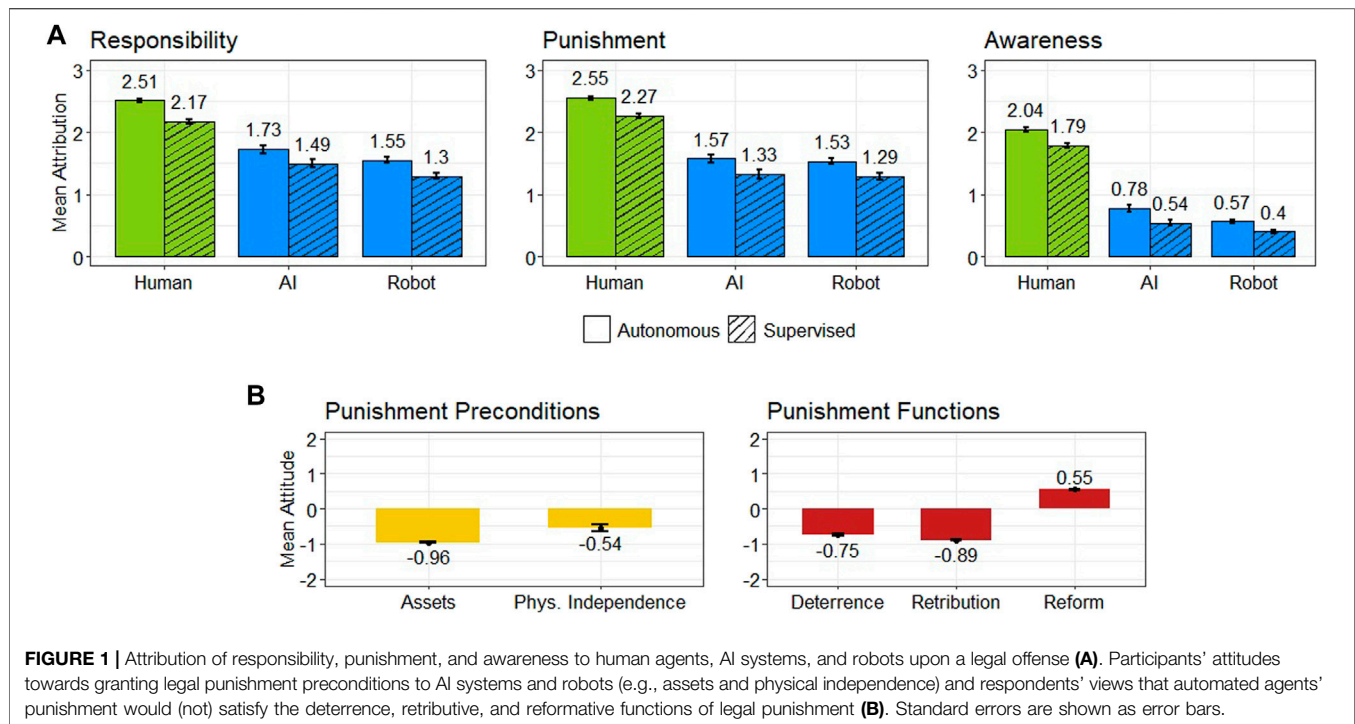
2 METHODS

Our research inquired how people's moral judgments of automated systems may clash with existing legal doctrines through a survey-based study. We recruited 3,315 US residents through Amazon Mechanical Turk (see SI for demographic information), who attended a study where they 1) indicated their perception of automated agents' liability and 2) attributed responsibility, punishment, and awareness to a wide range of entities that could be held liable for harms caused by automated systems under existing legal doctrines.

We employed a between-subjects study design, in which each participant was randomly assigned to a scenario, an agent, and an autonomy level. Scenarios covered two environments where automated agents are currently deployed: medicine and war (see SI for study materials). Each scenario posited three agents: an AI program, a robot (i.e., an embodied form of AI), or a human actor. Although the proposal of extending legal standing to AI systems and robots have similarities, they also have distinct aspects worth noting. For instance, although a "robot death penalty" may be a viable option through its destruction, "killing" an AI system may not have the same expressive benefits due to varying levels of anthropomorphization. However, extensive literature discusses the two actors in parallel, e.g., (Turner, 2018; Abbott, 2020). We come back to this distinction in our final discussion. Finally, our study introduced each actor as either "supervised by a human" or "completely autonomous."

Participants assigned to an automated agent first evaluated whether punishing it would fulfill some of legal punishment's functions, namely reform, deterrence, and retribution (Solum, 1991; Asaro, 2007). They also indicated whether they would be willing to grant assets and physical independence to automated systems—two factors that are preconditions for civil and criminal liability, respectively. If automated systems do not hold assets to be taken away as compensation for those they harm, they cannot be held liable under civil law. Similarly, if an AI system or robot does not possess any level of physical independence, it becomes hard to imagine their criminal punishment. These questions were shown in random order and answered using a 5-point bipolar scale.

After answering this set of questions or immediately after consenting to the research terms for those assigned to a human agent, participants were shown the selected vignette in plain text. They were then asked to attribute responsibility, punishment, and awareness to their assigned agent. Responsibility and punishment are closely related to the proposal of adopting electronic legal personhood, while awareness plays a major role in legal judgments (e.g., mens rea in criminal law, negligence in civil law). We also identified a series of entities (hereafter associates) that could be held liable under existing legal doctrines, such as an automated system's manufacturer under product liability, and asked participants to attribute the same variables to each of them. All questions were answered using a 4-pt scale. Entities were shown in random order and one at a time.

We present the methodology details and study materials in the SI. A replication with a demographically representative sample (N = 244) is also shown in the SI to substantiate all of the findings presented in the main text. This research had been

FIGURE 1 | Attribution of responsibility, punishment, and awareness to human agents, AI systems, and robots upon a legal offense **(A)**. Participants' attitudes towards granting legal punishment preconditions to AI systems and robots (e.g., assets and physical independence) and respondents' views that automated agents' punishment would (not) satisfy the deterrence, retributive, and reformative functions of legal punishment **(B)**. Standard errors are shown as error bars.

approved by the first author's Institutional Review Board (IRB). All data and scripts are available at the project's repository: https://bit.ly/3AMEJjB.

3 RESULTS

Figure 1A shows the mean values of responsibility and punishment attributed to each agent depending on their autonomy level. Automated agents were deemed moderately responsible for their harmful actions ($M = 1.48$, SD = 1.16), and participants wished to punish AI and robots to a significant level ($M = 1.42$, SD = 1.28). In comparison, human agents were held responsible ($M = 2.34$, SD = 0.83) and punished ($M = 2.41$, SD = 0.82) to a larger degree.
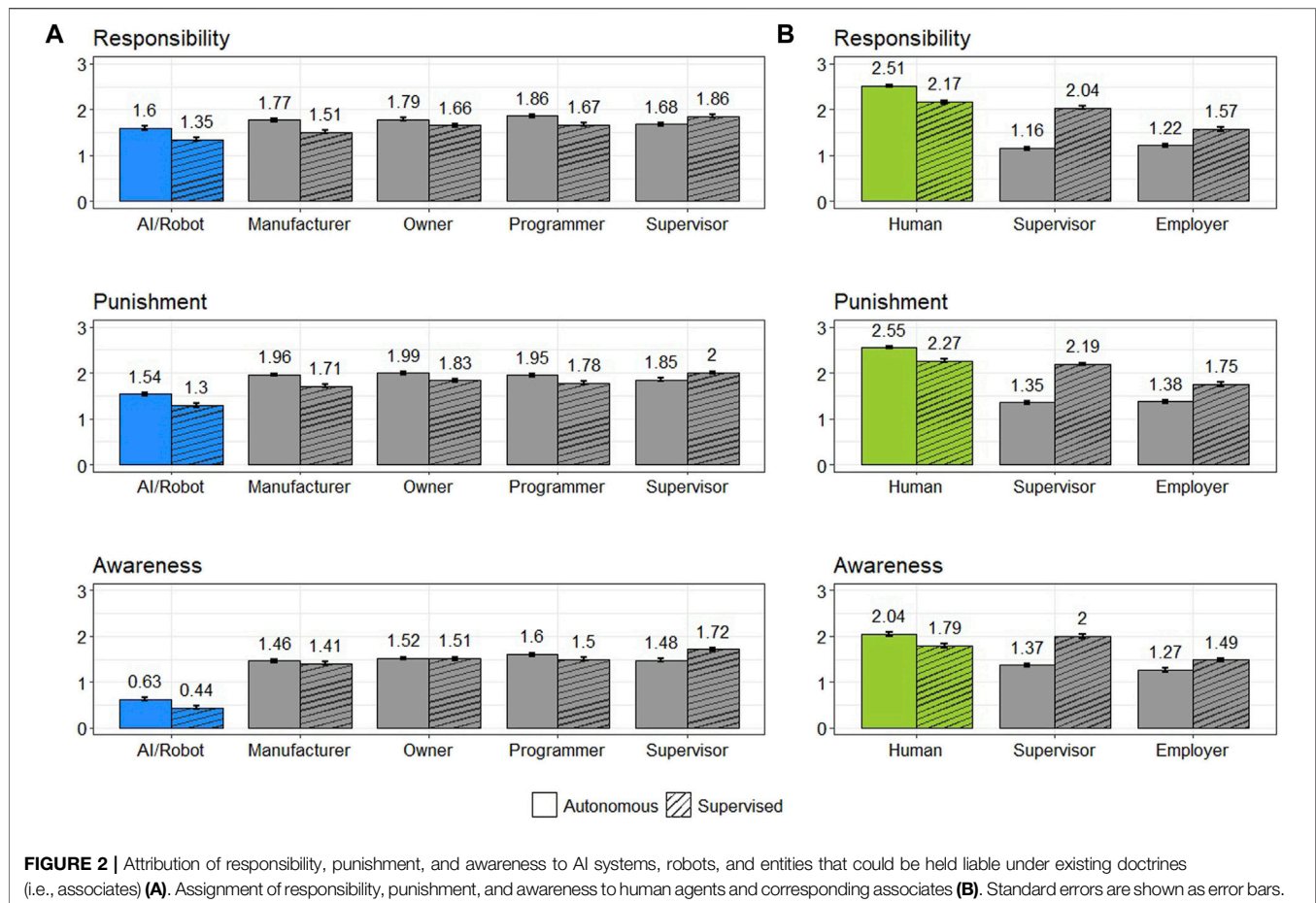
A 3 (agent: AI, robot, human) × 2 (autonomy: completely autonomous, supervised) ANOVA on participants' judgments of responsibility revealed main effects of both agent (F (2, 3309) = 906.28, $p < 0.001$, $\eta_p^2 = 0.35$) and autonomy level (F (1, 3309) = 43.84, $p < 0.001$, $\eta_p^2 = 0.01$). The extent to which participants wished to punish agents was also dependent on the agent (F (2, 3309) = 391.61, $p < 0.001$, $\eta_p^2 = 0.16$) and its autonomy (F (1, 3309) = 45.56, $p < 0.001$, $\eta_p^2 = 0.01$). The interaction between these two factors did not reach significance in any of the models ($p > 0.05$). Autonomous agents were overall viewed as more responsible and deserving of a larger punishment for their actions than their supervised counterparts. We did not observe noteworthy differences between AI systems and robots; the latter were deemed marginally less responsible than AI systems.

Figure 1A shows the mean perceived awareness of AI, robots, and human agents upon a legal offense. Participants perceived automated agents as only slightly aware of their actions ($M = 0.54$,

SD = 0.88), while human agents were considered somewhat aware ($M = 1.92$, SD = 1.00). A 3 × 2 ANOVA model revealed main effects for both agent type (F (2, 3309) = 772.51, $p < 0.001$, $\eta_p^2 = 0.35$) and autonomy level (F (1, 3309) = 43.87, $p < 0.001$, $\eta_p^2 = 0.01$). The interaction between them was not significant ($p = 0.401$). Robots were deemed marginally less aware of their offenses than AI systems. A mediation analysis revealed that perceived awareness of AI systems (coded as -1) and robots (coded as 1) mediated judgments of responsibility (partial mediation, coef = −0.04, 95% CI [−0.06, −0.02]) and punishment (complete mediation, coef = −0.05, 95% CI [−0.07, −0.02]).

The leftmost plot of **Figure 1B** shows participants' attitudes towards granting assets and some level of physical independence to AI and robots using a 5-pt scale. These two concepts are crucial preconditions for imposing civil and criminal liability, respectively. Participants were largely contrary to allowing automated agents to hold assets ($M = −0.96$, SD = 1.16) or physical independence ($M = −0.55$, SD = 1.30). **Figure 1B** also shows the extent to which participants believed the punishment of AI and robots might satisfy deterrence, retribution, and reform, i.e., some of legal punishment's functions. Respondents did not believe punishing an automated agent would fulfill its retributive functions ($M = −0.89$, SD = 1.12) or deter them from future offenses ($M = −0.75$, SD = 1.22); however, AI and robots were viewed as able to learn from their wrongful actions ($M = 0.55$, SD = 1.17). We only observed marginal effects ($\eta_p^2 \leq 0.01$) of agent type and autonomy in participants' attitudes towards preconditions and functions of legal punishment and present these results in the SI.

The viability and effectiveness of AI systems' and robots' punishment depend on fulfilling certain legal punishment's preconditions and functions. As discussed above, the incompatibility between legal punishment and automated

FIGURE 2 | Attribution of responsibility, punishment, and awareness to AI systems, robots, and entities that could be held liable under existing doctrines (i.e., associates) **(A)**. Assignment of responsibility, punishment, and awareness to human agents and corresponding associates **(B)**. Standard errors are shown as error bars.

agents is a common argument against the adoption of electronic legal personhood. Collectively, our results suggest a conflict between people's desire to punish AI and robots and the punishment's perceived effectiveness and feasibility.

We also observed that the extent to which participants wished to punish automated agents upon wrongdoing correlated with their attitudes towards granting them assets (r (1935) = 0.11, $p <$ 0.001) and physical independence (r (224) = 0.21, $p <$ 0.001). Those who anticipated the punishment of AI and robots to fulfill deterrence (r (1711) = 0.34, $p <$ 0.001) and retribution (r (1711) = 0.28, $p <$ 0.001) also tended to punish them more. However, participants' views concerning automated agents' reform were not correlated with their punishment judgments (r (1711) = −0.02, $p =$ 0.44). In summary, more positive attitudes towards granting assets and physical independence to AI and robots were associated with larger punishment levels. Similarly, participants that perceived automated agents' punishment as more successful concerning deterrence and retribution punished them more. Nevertheless, most participants wished to punish automated agents regardless of the punishment's infeasibility and unfulfillment of retribution and deterrence.

Participants also judged a series of entities that could be held liable under existing liability models concerning their responsibility, punishment, and awareness for an agent's wrongful action. All of the automated agents' associates

were judged responsible, deserving of punishment, and aware of the agents' actions to a similar degree (see **Figure 2**). The supervisor of a supervised AI or robot was judged more responsible, aware, and deserving of punishment than that of a completely autonomous system. In contrast, attributions of these three variables to all other associates were larger in the case of an autonomous agent. In the case of human agents, their employers and supervisors were deemed more responsible, aware, and deserving of punishment when the actor was supervised. We present a complete statistical analysis of these results in the SI.

4 DISCUSSION

Our findings demonstrate a conflict between participants' desire to punish automated agents for legal offenses and their perception that such punishment would not be successful in achieving deterrence or retribution. This clash is aggravated by participants' unwillingness to grant AI and robots what is needed to legally punish them, i.e., assets for civil liability and physical independence for criminal liability. This contradiction in people's moral judgments suggests that people wish to punish AI and robots even though they believe that doing so would not be successful, nor are they willing to make it legally viable.

These results are in agreement with Danaher's (2016) retribution gap. Danaher acknowledges that people might blame and punish AI and robots for wrongful behavior due to humans' retributive nature, although they may be wrong in doing so. Our data implies that Danaher's concerns about the retribution gap are significant and can be extended to other considerations, i.e., deterrence and the preconditions for legal punishment. Past research shows that people also ground their punishment judgments in functions other than retribution (Twardawski et al., 2020). Public intuitions concerning the punishment of automated agents are even more contradictory than previously advanced by Danaher: they wish to punish AI and robots for harms even though their punishment would not be successful in achieving some of legal punishment's functions or even viable, given that people would not be willing to grant them what is necessary to punish them.

Our results show that even if responsibility and retribution gaps can be easily bridged as suggested by some scholars (Sætra, 2021; Tigard, 2020; Johnson, 2015), there still exists a conflict between the public reaction to harms caused by automated systems and their moral and legal feasibility. The public is an important stakeholder in the political deliberation necessary for the beneficial regulation of AI and robots, and their perspective should not be rejected without consideration. An empirical question that our results pose is whether this conflict warrants attention from scholars and policymakers, i.e., if they destabilize political and legal institutions (Sætra, 2021) or leads to lack of trust in legal systems (Abbott, 2020). For instance, it may well be that the public may need to be taught to exert control over their moral intuitions, as suggested by Kraaijeveld (2020).

Although participants did not believe punishing an automated agent would satisfy the retributive and deterrence aspects of punishment, they viewed robots and AI systems as capable of learning from their mistakes. Reform may be the crucial component of people's desire to punish automated agents. Although the current research might not be able to clear this inquiry, we highlight that future work should explore how participants imagine the reform of automated agents. Reprogramming an AI system or robots can prevent future offenses, yet it will not satisfy other indirect reformative functions of punishment, e.g., teaching others that a specific action is wrong. Legal punishment, as it stands, does not achieve the reprograming necessary for AI and robots. Future studies may question how people's preconceptions of automated agents' reprogramming influence people's moral judgments.

It might be argued that our results are caused by how the study was constructed. For instance, participants who punished automated agents might have reported being more optimistic about its feasibility so that their responses become compatible. However, we observe trends that methodological biases cannot explain but can only result from participants' a priori contradiction (see SI for detailed methodology). This work does not posit this contradiction as a universal phenomenon; we observed a significant number of participants attributing no punishment whatsoever to electronic agents. Nonetheless, we observed similar results in a demographically representative sample of respondents (see SI).

We did not observe significant differences between punishment judgments of AI systems and robots. The differences in responsibility and awareness judgments were marginal and likely affected by our large sample size. As discussed above, there are different challenges when adopting electronic legal personhood for AI and robots. Embodied machines may be easier to punish criminally if legal systems choose to do so, for instance through the adoption of a "robot death penalty." Nevertheless, our results suggest that the conflict between people's moral intuitions and legal systems may be independent of agent type. Our study design did not control for how people imagined automated systems, which could have affected how people make moral judgments about machines. For instance, previous work has found that people evaluate the moral choices of a human-looking robot as less moral than humans' and non-human robots' decisions (Laakasuo et al., 2021).

People largely viewed AI and robots as unaware of their actions. Much human-computer interaction research has focused on developing social robots that can elicit mind perception through anthropomorphization (Waytz et al., 2014; Darling, 2016). Therefore, we may have obtained higher perceived awareness had we introduced what the robot or AI looked like, which in turn could have affected respondents' responsibility and punishment judgments, as suggested by Bigman et al. (2019) and our mediation analysis. These results may also vary by actor, as robots are subject to higher levels of anthropomorphization. Past research has also shown that if an AI system is described as an anthropomorphized agent rather than a mere tool, it is attributed more responsibility for creating a painting (Epstein et al., 2020). A similar trend was observed with autonomous AI and robots, which were assigned more responsibility and punishment than supervised agents, as previously found in the case of autonomous vehicles (Awad et al., 2020b) and other scenarios (Kim and Hinds, 2006; Furlough et al., 2021).

4.1 The Importance of Design, Social, and Legal Decisions

Participants' attitudes concerning the fulfillment of punishment preconditions and functions by automated agents were correlated with the extent to which respondents wished to punish AI and robots. This finding suggests that people's moral judgments of automated agents' actions can be nudged based on how their feasibility is introduced.

For instance, to clarify that punishing AI and robots will not satisfy the human need for retribution, will not deter future offenses, or is unviable given they cannot be punished similarly to other legal persons may lead people to denounce automated agents' punishment. If legal and social institutions choose to embrace these systems, e.g., by granting them certain legal status, nudges towards granting them certain perceived independence or private property may affect people's decision to punish them. Future work should delve deeper into the causal relationship between people's attitudes towards the topic and their attribution of punishment to automated agents.

Our results highlight the importance of design, social, and legal decisions in how the general public may react to automated agents. Designers should be aware that developing systems that are perceived as aware by those interacting with them may lead to heightened moral judgments. For instance, the benefits of automated agents may be nullified if their adoption is impaired by unfulfilled perceptions that these systems should be punished. Legal decisions concerning the regulation of AI and their legal standing may also influence how people react to harms caused by automated agents. Social decisions concerning how to insert AI and robots into society, e.g., as legal persons, should also affect how we judge their actions. Future decisions should be made carefully to ensure that laypeople's reactions to harms caused by automated systems do not clash with regulatory efforts.

5 CONCLUDING REMARKS

Electronic legal personhood grounded on automated agents' abilities to fulfill duties does not seem a viable path towards the regulation of AI. This approach can only become an option if AI and robots are granted assets or physical independence, which would allow civil or criminal liability to be imposed, or if punishment functions and methods are adapted to AI and robots. People's intuitions about automated agents' punishment are somewhat similar to scholars who oppose the proposal. However, a significant number of people still wish to punish AI and robots independently of their a priori intuitions.

By no means this research proposes that robots and AI should be the sole entities to hold liability for their actions. In contrast, responsibility, awareness, and punishment were assigned to all associates. We thus posit that distributing liability among all entities involved in deploying these systems would follow the public perception of the issue. Such a model could take joint and several liability models as a starting point by enforcing the proposal that various entities should be held jointly liable for damages.

Our work also raises the question of whether people wish to punish AI and robots for reasons other than retribution, deterrence, and reform. For instance, the public may punish electronic agents for general or indirect deterrence (Twardawski et al., 2020). Punishing an AI could educate humans that a specific action is wrong without the negative consequences of human punishment. Recent literature in moral psychology also proposes that humans might strive for a morally coherent world, where seemingly contradictory judgments arise so that the public perception of agents' moral qualities match the moral qualities of their actions' outcomes (Clark et al., 2015). We highlight that legal punishment is not only directed at the wrongdoer but also fulfills other functions in society that future work should inquire about when dealing with automated agents. Finally, our work poses the question of whether proactive actions towards holding existing legal persons liable for harms caused by automated agents would compensate for people's desire to punish them. For instance, future work might examine whether punishing a system's manufacturer may decrease the extent to which people punish AI and robots. Even if the responsibility gap can be easily solved, conflicts between the public and legal institutions might continue to pose challenges to the successful governance of these new technologies.

We selected scenarios from active areas of AI and robotics (i.e., medicine and war; see SI). People's moral judgments might change depending on the scenario or background. The proposed scenarios did not introduce, for the sake of feasibility and brevity, much of the background usually considered when judging someone's actions legally. We did not control for any previous attitudes towards AI and robots or knowledge of related areas, such as law and computer science, which could result in different judgments among the participants.

This research has found a contradiction in people's moral judgments of AI and robots: they wish to punish automated agents, although they know that doing so is not legally viable nor successful. We do not defend the thesis that automated agents should be punished for legal offenses or have their legal standing recognized. Instead, we highlight that the public's preconceptions of AI and robots influence how people react to their harmful consequences. Most crucially, we showed that people's reactions to these systems' failures might conflict with existing legal and moral systems. Our research showcases the importance of understanding the public opinion concerning the regulation of AI and robots. Those making regulatory decisions should be aware of how the general public may be influenced or clash with such commitments.

DATA AVAILABILITY STATEMENT

The datasets and scripts used for analysis in this study can be found at https://bitly.com/3AMEJjB.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) at KAIST. The patients/participants provided their informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors designed the research. GL conducted the research. GL analyzed the data. GL wrote the paper, with edits from MC, CJ, and KS.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.756242/full#supplementary-material

REFERENCES

Abbott, R. (2020). *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge University Press.

Asaro, P. M. (2011). 11 a Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. *Robot Ethics ethical Soc. implications robotics*, 169–186.

Asaro, P. M. (2007). Robots and Responsibility from a Legal Perspective. *Proc. IEEE* 4, 20–24.

Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2020a). Crowdsourcing Moral Machines. *Commun. ACM* 63, 48–55. doi:10.1145/3339904

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The Moral Machine experiment. *Nature* 563, 59–64. doi:10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., et al. (2020b). Drivers Are Blamed More Than Their Automated Cars when Both Make Mistakes. *Nat. Hum. Behav.* 4, 134–143. doi:10.1038/s41562-019-0762-8

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). "Beyond Accuracy: The Role of Mental Models in Human-Ai Team Performance," in Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2–11.

Bigman, Y. E., Waytz, A., Alterovitz, R., and Gray, K. (2019). Holding Robots Responsible: The Elements of Machine Morality. *Trends Cognitive Sciences* 23, 365–368. doi:10.1016/j.tics.2019.02.008

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2020). *The Moral Psychology of AI and the Ethical Opt-Out Problem*. Oxford, UK: Oxford University Press.

Brożek, B., and Janik, B. (2019). Can Artificial Intelligences Be Moral Agents. *New Ideas Psychol.* 54, 101–106. doi:10.1016/j.newideapsych.2018.12.002

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: the Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Carlsmith, K. M., and Darley, J. M. (2008). Psychological Aspects of Retributive justice. *Adv. Exp. Soc. Psychol.* 40, 193–236. doi:10.1016/s0065-2601(07)00004-4

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., et al. (2018). *Portrayals and Perceptions of Ai and Why They Matter*.

Cave, S., and Dihal, K. (2019). Hopes and Fears for Intelligent Machines in Fiction and Reality. *Nat. Mach Intell.* 1, 74–78. doi:10.1038/s42256-019-0020-9

Clark, C. J., Chen, E. E., and Ditto, P. H. (2015). Moral Coherence Processes: Constructing Culpability and Consequences. *Curr. Opin. Psychol.* 6, 123–128. doi:10.1016/j.copsyc.2015.07.016

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Sci. Eng. Ethics* 26, 2051–2068. doi:10.1007/s11948-019-00146-8

Danaher, J. (2016). Robots, Law and the Retribution gap. *Ethics Inf. Technol.* 18, 299–309. doi:10.1007/s10676-016-9403-3

Darling, K. (2016). "Extending Legal protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects," in *Robot Law* (Edward Elgar Publishing).

de Sio, F. S., and Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them. *Philos. Tech.*, 1–28. doi:10.1007/s13347-021-00450-x

Delvaux, M. (2017). Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (Inl)). European Parliament Committee on Legal Affairs.

Dewey, J., and Rogers, M. L. (2012). *The Public and Its Problems: An Essay in Political Inquiry*. Penn State Press.

Epstein, Z., Levine, S., Rand, D. G., and Rahwan, I. (2020). Who Gets Credit for Ai-Generated Art. *Iscience* 23, 101515. doi:10.1016/j.isci.2020.101515

European Commission (2021). Proposal for a Regulation of the European Parliament and of the council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain union Legislative Acts).

Furlough, C., Stokes, T., and Gillan, D. J. (2021). Attributing Blame to Robots: I. The Influence of Robot Autonomy. *Hum. Factors* 63, 592–602. doi:10.1177/0018720819880641

Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law (Edition 1)*. Routledge.

Gless, S., Silverman, E., and Weigend, T. (2016). If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability. *New Criminal L. Rev.* 19, 412–436. doi:10.1525/nclr.2016.19.3.412

Gordon, J. S. (2021). Artificial Moral and Legal Personhood. *AI Soc.* 36, 457–471. doi:10.1007/s00146-020-01063-2

Gunkel, D. J. (2018). *Robot Rights*. mit Press.

Jobin, A., Ienca, M., and Vayena, E. (2019). The Global Landscape of Ai Ethics Guidelines. *Nat. Mach Intell.* 1, 389–399. doi:10.1038/s42256-019-0088-2

Johnson, D. G. (2015). Technology with No Human Responsibility. *J. Bus Ethics* 127, 707–715. doi:10.1007/s10551-014-2180-1

Jowitt, J. (2021). Assessing Contemporary Legislative Proposals for Their Compatibility With a Natural Law Case for AI Legal Personhood. *AI Soc.* 36, 499–508. doi:10.1007/s00146-020-00979-z

Kim, T., and Hinds, P. (2006). "Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction," in ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication (IEEE), 80–85.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. (2021). Bad Machines Corrupt Good Morals. *Nat. Hum. Behav.* 5, 679–685. doi:10.1038/s41562-021-01128-2

Kraaijeveld, S. R. (2020). Debunking (The) Retribution (gap). *Sci. Eng. Ethics* 26, 1315–1328. doi:10.1007/s11948-019-00148-6

Kraaijeveld, S. R. (2021). Experimental Philosophy of Technology. *Philos. Tech.*, 1–20. doi:10.1007/s13347-021-00447-6

Kurki, V. A. (2019). *A Theory of Legal Personhood*. Oxford University Press.

Laakasuo, M., Palomäki, J., and Köbis, N. (2021). Moral Uncanny valley: a Robot's Appearance Moderates How its Decisions Are Judged. *Int. J. Soc. Robotics*, 1–10. doi:10.1007/s12369-020-00738-6

Lee, M., Ruijten, P., Frank, L., de Kort, Y., and IJsselsteijn, W. (2021). "People May Punish, but Not Blame Robots," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–11. doi:10.1145/3411764.3445284

Lima, G., Grgić-Hlača, N., and Cha, M. (2021). "Human Perceptions on Moral Responsibility of Ai: A Case Study in Ai-Assisted Bail Decision-Making," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–17. doi:10.1145/3411764.3445260

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). "Sacrifice One for the Good of many? People Apply Different Moral Norms to Human and Robot Agents," in 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), 117–124.

Matthias, A. (2004). The Responsibility gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics Inf. Technol.* 6, 175–183. doi:10.1007/s10676-004-3422-1

Mulligan, C. (2017). Revenge against Robots. *SCL Rev.* 69, 579.

Prosser, W. L. (1941). *Handbook of the Law of Torts*. West Publishing.

Rahwan, I. (2018). Society-in-the-loop: Programming the Algorithmic Social Contract. *Ethics Inf. Technol.* 20, 5–14. doi:10.1007/s10676-017-9430-8

Resseguier, A., and Rodrigues, R. (2020). Ai Ethics Should Not Remain Toothless! a Call to Bring Back the Teeth of Ethics. *Big Data Soc.* 7, 2053951720942541. doi:10.1177/2053951720942541

Sætra, H. S. (2021). Confounding Complexity of Machine Action: a Hobbesian Account of Machine Responsibility. *Int. J. Technoethics (Ijt)* 12, 87–100. doi:10.4018/IJT.20210101.oa1

Sætra, H. S., and Fosch-Villaronga, E. (2021). Research in Ai Has Implications for Society: How Do We Respond. *Morals & Machines* 1, 60–73. doi:10.5771/2747-5174-2021-1-60

Solaiman, S. M. (2017). Legal Personality of Robots, Corporations, Idols and Chimpanzees: a Quest for Legitimacy. *Artif. Intell. L.* 25, 155–179. doi:10.1007/s10506-016-9192-3

Solum, L. B. (1991). Legal Personhood for Artificial Intelligences. *NCL Rev.* 70, 1231.

Sparrow, R. (2007). Killer Robots. *J. Appl. Philos.* 24, 62–77. doi:10.1111/j.1468-5930.2007.00346.x

Tigard, D. W. (2020). There Is No Techno-Responsibility gap. *Philos. Tech.*, 1–19. doi:10.1007/s13347-020-00414-7

Turner, J. (2018). *Robot Rules: Regulating Artificial Intelligence*. Springer.

Twardawski, M., Tang, K. T. Y., and Hilbig, B. E. (2020). Is it All about Retribution? the Flexibility of Punishment Goals. *Soc. Just Res.* 33, 195–218. doi:10.1007/s11211-020-00352-x

van den Hoven van Genderen, R. (2018). "Do we Need New Legal Personhood in the Age of Robots and Ai," in *Robotics, AI and the Future of Law* (Springer), 15–55. doi:10.1007/978-981-13-2874-9_2

Waytz, A., Heafner, J., and Epley, N. (2014). The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *J. Exp. Soc. Psychol.* 52, 113–117. doi:10.1016/j.jesp.2014.01.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.