Will Punishing Robots Become Imperative in the Future?

Gabriel Lima
School of Computing
KAIST
gabriel.lima@kaist.ac.kr

Meeyoung Cha
Data Science Group
Institute for Basic Science
mcha@ibs.re.kr

Chihyung Jeon
Science and Technology Policy
KAIST
cjeon@kaist.edu

Kyungsin Park
Department of Law
Korea University
kyungsinpark@korea.ac.kr

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3383006

Abstract
The possibility of extending legal personhood to artificial intelligence (AI) and robots has raised many questions on how these agents could be held liable given existing legal doctrines. Intending to promote a broader discussion, we conducted a survey (N=3315) asking online users' impressions of electronic agents' liability. Results suggest the existence of what we call the *punishment gap* that refers to the public's demand to punish automated agents upon a legal offense, even though their punishment is currently not feasible. Participants were also negative in granting assets or physical independence to electronic agents, which are crucial liability requirements. We discuss possible solutions to this punishment gap and present how legal systems might handle this contradiction while maintaining existing legal persons liable for the actions of automated agents.

Author Keywords
Robots; Artificial Intelligence; Punishment Gap; Legal Issues; Legal Personhood

CCS Concepts
•**Applied computing** → **Psychology;** *Law;* •**Social and professional topics** → *Governmental regulations;*

Introduction

Law was originally designed to enforce rules and doctrines among natural persons. It did, nevertheless, adapt itself to accommodate other non-human entities like corporations and NGOs over time. Legal systems around the world are now challenged to better accommodate newly emerging entities such as artificial intelligence (AI) and robots. Autonomous electronic agents have begun to raise numerous legal issues [4] as they become widely deployed. An important topic that has been discussed by many scholars is how liability should be assigned in the case of damages caused by an AI or robot [11], as the current doctrines do not accommodate autonomous and self-learning products.

A long-established proposal is to extend some level of legal personhood to AI and robots [17]. This possibility remains controversial, with opinions across a broad spectrum ranging from extreme support [18] to complete opposition [6]. Granting legal personhood has had the difficulty of imposing liability[1] to AI and robots as one of its biggest hurdles, as it would not allow these agents to be punished for their actions had they caused any damage. Scholars have debated whether punishing these systems is indeed possible, and it is currently agreed upon that existing legal doctrines and punishment methods can not be applied to electronic agents [7, 4].

We present early results from a study that aims to observe online users' first impressions of the necessity and viability towards AI and robot's liability. The study results indicate that participants find electronic agents deserving of punishment for damages caused, although they oppose any possibility of granting assets or physical independence to electronic agents (i.e., requirements of civil and criminal liability). This contradiction, which we define as the punishment gap, needs to be solved regardless of the legal standing of AI and robots.

We offer and analyze possible solutions to the punishment gap, by giving attention to the public's desire for electronic punishment. Punishing automated agents might be beneficial to society as a whole, as it might decrease retributivist sentiments by transferring such negative feelings to inanimate agents. Electronic legal personhood is not currently viable, given the existence of the punishment gap; the punishment gap, however, exists regardless of the legal status of AI and robots. We highlight the fact that solving the punishment gap should not lead to the extinction of the liability of existing legal persons for damages caused by AI and robots; the punishment of electronic agents must coexist with existing liability models.

How AI and Robots Challenge the Current Legal System

The rapid deployment of AI and robots raises various issues in far-reaching areas of law [4]. Regulation of innovations always creates turmoil as legal systems are unable to rely on existing frameworks and precedents. Similar to how the internet development raised various questions at the start of the 21st century, robots and AI will bring lawyers, policymakers, developers, and many other stakeholders to the drawing board.

These electronic agents, namely AI and robots, have the potential to challenge how current regulation treats manufactured goods due to their distinctive characteristics. For instance, robots and AI will be in a position to make moral choices. For example, given an inevitable accident, should an autonomous car prioritize the life of a passenger or a pedestrian? [5] These systems can also learn by

[1] For this research, "punishment" and "liability" mean both civil liability and criminal liability.

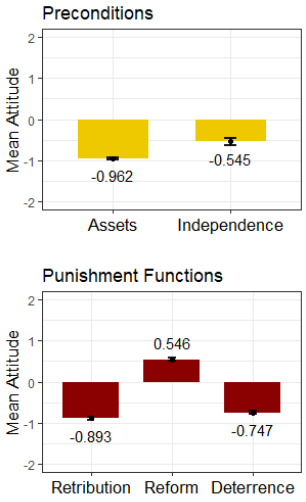

Figure 1: Comparing the perceived fulfillment of preconditions and punishment functions of electronic agents. Positive values indicate a supportive mean attitude towards granting assets and physical independence to AI and robots or a positive belief that their punishment fulfills its functions; negative values, however, represents a position against the fulfillment of punishment functions and requirements.

themselves based on evolutionary and training algorithms that allow AI and robots to modify themselves to better suit users' needs (e.g., directed advertisement), making the imposition of product liability more difficult [18]. Legal scholars have also raised the question of whether an autonomous AI or robots could break the chain of causation needed for liability assignment [11]. Nevertheless, it is agreed that electronic agents will challenge existing legal doctrines due to their autonomy and self-learning capabilities, which prevent manufacturers and users from predicting the behavior of their products.

In an earlier report, the European Parliament announced an idea to extend some level of legal standing to AI and robots [1]. This EU report proposed to create *"the status of electronic persons with specific rights and obligations"* to highly autonomous robots. This legal status would be applied similarly to the concept of legal personhood, which is used with many entities, such as humans, corporations, and nations, and grants legal rights and obligations to its holders. This proposal, if adopted, would completely change how AI and robots are to be treated in a particular legal system, transforming them from a single product to an entity with their own rights and obligations.

The possibility of extending some level of legal personality to AI and robots involves a multifaceted discussion with scholars and policymakers in a spectrum from completely support towards the proposal to the belief that electronic legal personhood must be avoided at all costs. Van Genderen, for instance, uses utilitarian theories to defend that legal personhood could be granted if legal systems find it beneficial to do so [21]. Koops et al. believe that issues of responsibility and punishment assignment would be more easily solved if electronic agents could be held liable and accountable for their actions [16]. Adversaries to the proposal, on the other hand, argue that AI and robots are not human, and extending such concept to them could confront with humans rights [2]. For instance, Bryson argues that electronic legal personhood would create problems both at an individual and institutional level [6].

Alongside the discussion of extending legal personhood to AI and robots is the issue of their punishment. Liability imposition, under current doctrines, takes the right to hold assets and physical independence as critical requirements. If a legal person does not own assets or physical freedom to be taken away, punishment, at its current form, is not feasible. This precondition is one of the driving arguments against the proposal [7, 4]. Nevertheless, some scholars believe that granting assets or some level of independence to electronic agents would be beneficial by allowing a developer to better shape their behavior by teaching them to value such concepts, and thus facilitating the adoption of electronic legal personhood [18].

Defining the Punishment Gap

Punishment fulfills three main functions: deterrence, retribution, and reform [17]. Therefore, the possibility of extending legal personhood raises the question of whether the punishment of AI and robots can indeed be achieved and whether it fulfills its functions.

Inspired by participatory policymaking [19, 20] and descriptive ethics [15], we conducted a survey-based study (N=3315) and asked "what do online users think of the possibility of punishing AI and robots?". We attended the survey on Amazon Mechanical Turk (N=3315), an online survey platform which has shown results equally as or better than survey panels [8, 14], asking whether survey participants 1) agree to grant assets and physical independence

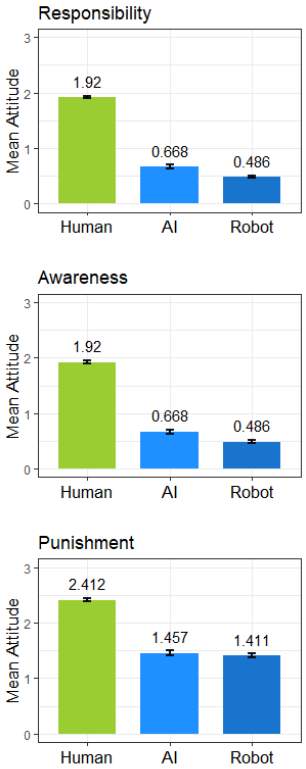

Figure 2: Attribution of responsibility, awareness, and punishment to a human or electronic wrongdoer in the case of a legal offense. The scale on the y-axis represents: 0=Not at all, 1=A little, 2=Some, 3=Very.

to AI and robots and 2) believe that their punishment fulfills its primary functions.

The initial results (Fig. 1) indicate that online users do not agree on granting assets or physical independence to AI and robots, which would prove to be a hurdle in extending legal personhood to AI and robots. Fig. 1 also shows that respondents consider that the punishment of AI and robots does not fulfill the retributive and deterrence aspects of punishment.

In the same study, we also asked the participants to assign responsibility, awareness, and punishment to human and electronic wrongdoers in real-life adapted scenarios. These three variables are important aspects of legal liability. If an agent is responsible for an action that caused harm, he or she is held liable. Punishment is the method through which society is kept in check and makes good any damage caused. Finally, awareness is often defined as the foreseeability of an action (e.g., in civil law) or the guilty mind of an agent (e.g., in criminal law). The question presented did not ask survey participants whether the wrongdoer was aware of its actions. but focuses on the normative aspect of whether the agents should be attributed awareness for their efforts.

The results (Fig. 2) show that online users assign moderate levels of responsibility and punishment to AI and robots, even though they should not be attributed awareness. Earlier studies indicate that punishment assignment among humans is a two-step process, initiated by the causality aspect of the damage caused and supported by the state of mind of the wrongdoer [12]; robots and AI were assigned lower levels of punishment because they were awarded a low level of awareness.

Nevertheless, survey results indicate that online users desire to punish robots for the damages caused. They do not, however, consider that punishing them is possible and useful. Participants are also not willing to grant liability preconditions to AI and robots. We define this public contradiction as the *punishment gap*. Expanding on the previously proposed concept of retribution gap [13], which focuses on the fact that the AI and robots will not be proper subjects of retributive blame if victims choose to assign them responsibility for damages, this gap is more comprehensive than previously thought by also encompassing the deterrence aspect of liability imposition.

In conclusion, participants' intuition led them to consider electronic entities as causally responsible agents that must be punished for their actions. Therefore, even though electronic legal personhood is hard to conceptualize and adopt in the short run, the punishment of AI and robots might become imperative, given the human desire to punish these systems.

Possible Solutions to the Punishment Gap

Existing forms of punishment would not apply to robots since they do not satisfy the preconditions for liability, i.e., assets or physical independence. This does not, however, change online users' desire to punish them. The survey findings indicate that even though participants are aware that robots and AI do not satisfy these preconditions, they, nonetheless, believe electronic agents should be punished. This set of conflicting interests may require a broader legal reframing by changing how current punishment methods are applied and even lead to the development of new forms of punishment to hold AI and robots liable.

Nevertheless, the discussion above raises the question of how one can punish electronic agents should they satisfy

the preconditions of punishment. While some scholars argue that this question still does not provide any solution [7], others believe that adapting AI, robots, and society to electronic punishment could contribute to safer systems [18].

A partial solution to the punishment gap is to embed more "realistic interests" into AI and robots [18]. By granting such interests to electronic agents and teaching them to value such interests, their behaviors could be shaped more desirably, which resolves the deterrence issue of the punishment gap. The retributive aspect of the punishment gap could also be resolved. For instance, punishing an AI or robot that has similar interests and values to humans would be seen as legitimate by the public. What had been suggested as "realistic interests" is hard to conceptualize under the punishment gap, and this concept might need to be better defined by future researchers. For instance, take a robot that had been granted assets. If this robot causes any damage, their assets could be taken away and used to compensate those harmed. Even though this might sound somewhat inconceivable due to the nature of the robot, we conduct financial transactions daily with corporations and governments, all of which are non-human legal entities.

Implementing the right to hold assets to electronic agents might present itself as a hard question, but could be adopted through some remuneration for an AI or robot's work. Take, as an example, an autonomous taxi owned by a corporation. The taxi could receive a part of the taxi fare and keep it as a balance in the event of an accident. Another possibility would be to allow these systems to hold insurance policies [22]. Such a model would require all autonomous systems to hold mandatory insurance policies with premiums depending on their safety record. Systems with excellent track records would qualify for low incentives, while less safe AI and robots would only be counted for policies that would make them financially unfeasible. It is essential to add that this proposal could also deal with the deterrence aspect of the punishment gap as it promotes systems to hold a sterling safety record. If such policies are implemented in a way that the general public perceived the insurance premiums as a burden (i.e., punishment) to the autonomous system, the retributive aspect could also be arguably dealt with.

The possible solutions presented above aim to punish manufacturers, users, and other entities alongside electronic agents so that AI and robots do not become human liability shields. This work, by no means, posits that existing legal persons should be exempted for the actions of their products; rather, AI and robots might need to be punished alongside them due to public demand. All entities, including manufacturers, users, and programmers involved in the deployment of electronic agents, should be held accountable for the actions of an AI or robot. The degree of liability, however, might vary given the circumstance and electronic agent in question (e.g., the manufacturer of a completely autonomous robot might be held liable to a smaller degree). The most important aspect of any solution to the punishment gap is the general public having the perception that the causally responsible agents are being punished, dealing with their demands, instead of shifting liability from other entities to AI and robots.

Concluding Remarks

Regardless of the effectiveness of the solutions discussed in the previous section, electronic legal personhood is not yet viable without broad legal reframing or development of systems that satisfy liability requirements. Although granting a certain legal status to electronic agents might currently be unfeasible, the punishment of AI and robots might

become imperative as these systems are widely deployed and become the causes of damages.

Our study results suggest that the general public may find AI and robots causally responsible for their actions and believe that they should be punished for their actions. Our participant sample, however, do not believe that their liability fulfills its main functions and do not agree in granting liability requirements, namely assets and physical independence, to electronic agents. This public contradiction leads to what we define as the punishment gap. The punishment gap exists regardless of the legal status of AI and robots since it stems from the public perception of the actions of electronic agents and their consequences.

Observations in this paper are also based on the fact that humans are empirically found to be retributivists when dealing with other humans [9]. Whether or not this finding is also accurate when tackling the actions of AI and robots is still an open question that must be tackled in the near future. If results are found to match existing previous work, the punishment gap might support the argument that legal systems should consider how public desire for electronic punishment can be achieved. The results presented in this paper must also be validated with more representative samples and different scenarios so that the existence of the punishment gap can indeed be verified, as the sample of the current study is not representative nor its scenarios exhaustive.

The punishment of electronic agents might be psychologically beneficial to individuals and society as a whole. Scholars have previously criticized the fact that retributivism is often the leading aspect when assigning punishment to wrongdoers [3]. Psychological evidence also indicates that such a retributivist feeling (i.e., revenge) is held more strongly if the wrongdoer is indeed punished. In contrast, the need for vengeance is more easily forgotten if the wrongdoer does not suffer the consequences [10]. Therefore, it might be possible to use the punishment of electronic agents as a form of motivating victims to "move on" from their retributive sentiment, leading to improvements of individual post-traumatic sentiments and a society that focus its punishment on reform, rather than vengeance.

Assigning legal liability to electronic agents by no means imply that manufacturers, users, and other entities involved in the deployment of AI and robots should not be held liable for the actions of AI and robots. The punishment of electronic agents might need to coexist with existing liability models given the public demand for punishment of AI and robots. By applying current doctrines alongside the possible punishment of electronic agents, concerns of AI and robots becoming human liability shields [7] are reduced as humans are not shielded from liability.

The possible solutions for the punishment gap discussed in this article are based on the human perception that AI and robots are being punished, regardless of the inquiry of whether such punishment fulfills its functions. If the public believes that AI and robots are being punished even though the legal system does not believe so, the punishment gap is nevertheless dealt with.

REFERENCES

[1] 2017. *European Parliament report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*.

[2] 2017. *Open Letter to the European Commission on Artificial Intelligence and Robotics*. Available at `http://www.robotics-openletter.eu/`.

[3] Devika Agrawal. 2015. The Impulse to Punish: A Critique of Retributive Justice. (2015).

[4] Peter M Asaro. 2011. 11 A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. *Robot ethics: The ethical and social implications of robotics* (2011), 169.

[5] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59.

[6] Joanna J Bryson. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (2010), 63–74.

[7] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.

[8] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[9] Kevin M Carlsmith and John M Darley. 2008. Psychological aspects of retributive justice. *Advances in experimental social psychology* 40 (2008), 193–236.

[10] Kevin M Carlsmith, Timothy D Wilson, and Daniel T Gilbert. 2008. The paradoxical consequences of revenge. *Journal of personality and social psychology* 95, 6 (2008), 1316.

[11] Samir Chopra and Laurence F White. 2011. *A legal theory for autonomous artificial agents*. University of Michigan Press.

[12] Fiery Cushman. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 2 (2008), 353–380.

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

[14] Marc Dupuis, Barbara Endicott-Popovsky, and Robert Crossler. 2013. An analysis of the use of amazons mechanical Turk for survey research in the cloud. In *proc. of the International Conference on Cloud Security Management (ICCSM)*. 10.

[15] Bernard Gert and Joshua Gert. 2002. The definition of morality. (2002).

[16] Bert-Jaap Koops, Mireille Hildebrandt, and David-Olivier Jaquet-Chiffelle. 2010. Bridging the accountability gap: Rights for new entities in the information society. *Minn. JL Sci. & Tech.* 11 (2010), 497.

[17] Lawrence B Solum. 1991. Legal personhood for artificial intelligences. *NCL Rev.* 70 (1991), 1231.

[18] Jacob Turner. 2018. *Robot Rules: Regulating Artificial Intelligence*. Springer.

[19] Sybille Van den Hove. 2000. Participatory approaches to environmental policy-making: the European Commission Climate Policy Process as a case study. *Ecological Economics* 33, 3 (2000), 457–472.

[20] L van Dijk, A Hayton, DCJ Main, A Booth, A King, DC Barrett, HJ Buller, and KK Reyher. 2017. Participatory Policy Making by Dairy Producers to Reduce Anti-Microbial use on Farms. *Zoonoses and public health* 64, 6 (2017), 476–484.

[21] Robert van den Hoven van Genderen. 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.

[22] David C Vladeck. 2014. Machines without principals: liability rules and artificial intelligence. *Wash. L. Rev.* 89 (2014), 117.