Volume 2, No. 4 | July–August 2019

RAIL

The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL

The Journal of Robotics, Artificial Intelligence & Law

Volume 2, No. 4 | July–August 2019

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. Copyright 2019 Fastcase, Inc. No part of this journal may be reproduced in any form—by microfilm, xerography, or otherwise—or incorporated into any information retrieval system without the written permission of the copyright owner. For customer support, please contact Fastcase, Inc., 711 D St. NW, Suite 200, Washington, D.C. 20004, 202.999.4777 (phone), 202.521.3462 (fax), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Morgan Morrissette Wright
Journal Designer: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

This publication is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If legal advice or other expert assistance is required, the services of a competent professional should be sought.

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

711 D St. NW, Suite 200, Washington, D.C. 20004
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 711 D St. NW, Suite 200, Washington, D.C. 20004.

Articles and Submissions

Direct editorial inquires and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc.,
26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@meyerowitzcommunications.com, 646.539.8300.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please call:

Morgan Morrissette Wright, Publisher, Full Court Press at mwright@fastcase.com or at 202.999.4878

For questions or Sales and Customer Service:

Customer Service
Available 8am–8pm Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)
ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

The Robo-Criminal

Nina Scholten*

How can robots be punished for criminal offenses to best protect the values of criminal law? This article analyzes the conditions of criminal liability and the possibilities of holding a robot criminally liable. It explores the ways punishment can be imposed on robots for committing a criminal offense in order to best protect the values of criminal law, and critically reflects on the results thereof.

Robots are taking on a more important role in our society. In the coming years, self-driving cars will begin to take over our roads and the doctor performing surgery is increasingly less likely to be made out of flesh and blood. This increasing engagement of robots in our daily life brings about new questions, including many legal issues.

In March 2018, one of Uber's autonomous testing vehicles killed a pedestrian.[1] Although this vehicle already runs on relatively intelligent algorithms, it had a difficult time identifying a woman on a bike and failed to break on time. In this situation, the system confused itself, which was the reason for the accident.[2] Situations like this bring up important legal issues. Robots are getting more sophisticated and are no longer programmed to follow a certain set of rules. Instead, they are programmed with learning rules and adapt to environments in ways unpredictable to their original programmer. We are no longer able to know beforehand the exact ways in which a robot will act in a certain situation. And before writing this off as science fiction, such robots already exist.

So what does happen if such a robot commits a criminal offense? Would it be possible to hold robots criminally liable? And if convicted, what punishment could we apply?

Throughout the analysis in this article, it is important to be aware of *The Android Fallacy*: "the idea that robots are 'just like people' and that there is a meaningful difference between humanoid and non-humanoid robots."[3] Just because robots can look like people, does not mean they should be treated different from other "machines."[4] To prevent falling into this fallacy and anthropomorphizing machines it is useful to think of a robot as, for example, a self-driving car.

Robotics, Artificial Intelligence & Law / July–August 2019, Vol. 2, No. 4, pp. 263–283.

ISSN 2575-5633 (print) /ISSN 2575-5617 (online)

How Is "Robot" Defined?

There are many ways in which people imagine a "robot." Defining this concept is therefore important for the purpose of avoiding any confusion throughout this article. To draw a picture, consider the following scenario:

> A fully autonomous self-driving car with a person in the passenger seat is driving on a bridge. All of a sudden, a wire falls onto the bridge in front of the self-driving car. There are three things the car can decide to do: it can swerve into the oncoming traffic on the other lane, hitting a school bus full of children. Alternatively it can pull onto the sidewalk, hitting two elderly ladies. Finally, the car can decide to do nothing and as a result hit the wire, careen off the bridge and kill the passenger of the car, ultimately destroying the vehicle itself.

This scenario is an ethical thought experiment: the "trolley problem." The car has to make a difficult decision. A programmer can input some general moral considerations or preferences, but he has not foreseen the specific situation and cannot predict the precise actions of the car. In a situation like the above, legal responsibility is a difficult issue. Robots have become so autonomous that we are not able to just point a finger at the programmer anymore as the responsible party.[5] There is no longer a causal link between the programmers or designers of the robot and the final effect.[6]

As is clear from the example, what is meant by robot is not "simple" electronic equipment. We are not talking about machines programmed to perform a certain defined thought process. We are not talking about robots directly controlled by operators, because in all such cases the robot will not be criminally liable. Instead, the operator or manufacturer will be. In this context, what we are talking about are smart robots, robots with learning capabilities—machines that can think.[7] This type of robot has the capability of changing algorithms informing its actions in response to interactions with people or objects in its environment without immediate instructions to make those changes.[8]

In 1987, Roger Schank had already attempted to define artificial intelligence ("AI"). A general definition is difficult, because it depends on the specific type and goals of the AI in question. Therefore, Schank approached "an intelligent entity" as having five attributes. The first is communication. The more intelligent the entity, the better people can communicate with it. The second attribute is mental (internal) knowledge, meaning an entity will have some knowledge about itself. It needs to know what it knows. Third is world (external) knowledge: an intelligent entity will be capable of learning—and will have knowledge about the outside world and can use that information. It has a memory and can use past experiences to guide it in new experiences. The fourth attribute is intentionality (goal-driven behavior), which refers to the action that an intelligent entity will take to achieve certain goals. Last is creativity, meaning that when an initial action fails, an intelligent entity should be creative enough to take alternate action.[9]

Professor Ryan Calo defines robots as machines capable of sensing their environment, processing the information they sense and acting directly upon their environment.[10] Professor Jack Balkin expands the definition by also including unembodied robots—artificial intelligence agents and machine learning algorithms. In the context of criminal law, since both embodied and unembodied robots can commit criminal offenses, it makes sense to include both in the definition.[11]

Humans can play different roles in relation to robots and therefore bear different levels of responsibility. First there is the person, or persons, who created and trained the algorithms and built the physical components: the "programmer" or "designer." Then there is the role of the operator of the robot, who instructs it to carry out certain actions. You also have a person who is the legal owner of the robot. All those three roles could be fulfilled by one or by different people. When referring to "people behind the robot," I mean the collective of programmers or designers, operators and owners.

How Can a Robot Satisfy the Conditions of Criminal Liability?

The criminal law system was originally designed around humans. However, the pragmatic legal system of the United States has already integrated acts of legal persons, such as corporations. This implies that criminal liability is not limited to humans, but can also be applied to nonhuman entities.[12] Could a robot, as defined herein, be held liable for a criminal offense?

There are several criminal liability models that can be applicable in the context of robots, depending on their level of autonomy.[13] One is the *perpetrator-via-another* model, wherein the robot is used as an instrument by the perpetrator. The robot completely depends on the programmer, operator, or owner, who has criminal intent. This model will not be used in this article, because here the robot is innocent and therefore not the subject of punishment.

A second approach is the *natural-probable-consequence* model. In these instances, the offense is a natural and probable consequence of the robot's conduct. The programmer, operator, or owner could have reasonably foreseen the criminal conduct of the robot. However, the person has been negligent in preventing that conduct and therefore he or she can be held accountable. This theory applies both in the situation when the person had no criminal intent but was negligent, and to the situation where a person had criminal intent for one offense, but the robot instead or additionally committed another. If the robot was not just an innocent agent it can also be held liable (together with the programmer, owner, or operator). The third model, which is most relevant for this article, is the *direct liability* model.

To be able to impose criminal liability on any person or kind of entity, two components of a crime have to be proven: *actus reus* and *mens rea*.[14] To hold a robot criminally liable for an offense under this model, we need to look into how robots can satisfy those two requirements.

Actus Reus

Actus reus is most commonly interpreted as the conduct and the harmful result of a criminal offense.[15] It is expressed by acts or omissions.[16] It is not difficult to see how a robot can satisfy this component, since the robot can act—or fail to act—in the physical world. A self-driving car running someone over, thereby injuring or killing that person, satisfies the *actus reus* requirement.

Mens Rea

The *mens rea* requirement is the mental element of the crime, the criminal intent. This is more challenging to prove in the context of robots. It is useful to look at how *mens rea* is approached in the context of a corporation, since this is also a non-human entity that can accrue liability under criminal law.

The *intentional stance approach* is one of the approaches used to attribute a psychological state to a corporation. That approach treats a corporation as if it is capable of having mental states. Following this approach, corporations have a unique culture, character or ethos, which can create the possibility of corporate wrongdoing.[17] The highly developed robots discussed in this article can have such complicated decision-making systems that it is not feasible to determine the reasons for the robot's action by questioning the people programming that robot. Even the programmers might not know precisely how they work. Therefore, it could be more useful to treat them as rational agents and try to find an explanation for the robot's actions. This can lead to the best interpretation and prediction of that objects behavior.[18]

Which mental state is needed to satisfy the *mens rea* requirement depends on the crime.[19] Either knowledge, (specific) intent, or negligence needs to be proven. Or none, if it is a strict liability offense.[20] The definition of knowledge used by Gabriel Hallevy is "sensory reception of factual data and the understanding of that data."[21] Robots are capable of attaining and processing such knowledge: a self-driving car will receive information through its sensors about a pedestrian crossing the street, algorithms will process and analyze this data and the car will hit the break.

Specific intent implies the necessity of a purpose or an aim that a factual event will occur. Robots as defined in this article, are capable of goal-driven behavior: they are capable of executing action A to reach goal B. Therefore, a robot is capable of such intent. To prove intent when it comes to humans, we try to determine the purpose for acting or if that person knew or could have been reasonably expected to know that the result was almost certainly going to occur. We are not able to look into a person's mind, so rather we will look at circumstances and ask witnesses about facts or observations. Based on such factors, we try to show the person intended to reach a certain aim. Without a confession it is difficult to prove an internal state of mind.[22] However, in the case of a robot, this is arguably easier. We may be able to look into its "black box" in which we can see for what reason a robot executed an action.[23] Robots are still in development, with the example of self-driving cars, in which certain morals are being programmed so that they can make ethical decisions. This will allow intelligent robots to engage in moral reasoning.[24]

There are some crimes that require more, such as race- or gender-based crimes, where the element of "hatred" is required. For that small category of crimes, a robot is most likely not able to satisfy the requirements. However, this is for a minor amount of offenses. Since *actus reus* and *mens rea* are the only criteria for criminal liability and those can, in most situations, be satisfied by the highly autonomous robots, this leads to the conclusion that robots can be criminally liable for their acts and punishment can accordingly be imposed for such criminal conduct.

What is Meant by "Punishment" in Criminal Law?

Definition

Punishment can be described as having the following features:

1. It is being performed by, and directed at, agents who are responsible in some sense;
2. It involves deliberate harmful or unpleasant consequences;
3. Generally those harmful consequences are preceded by a conviction;
4. It is imposed by someone who has authority to do so;
5. It is imposed for a breach of an established rule of behavior;
6. It is imposed on an actual or supposed violator of that rule.[25]

When looking at the definition, most features are not at all problematic in the context of robots and can be easily satisfied. The "harmful or unpleasant consequences" do deserve attention in the context of robots. This will be further discussed hereafter.

Ratio

There are two important theories used to justify punishment. The first approach is utilitarian, which has been the dominant theory in American jurisprudence over the last century. According to utilitarians, punishment is justified because of the useful purposes that it serves. The justification of the punishment is based on supposed benefits that will accrue from its imposition.[26] It is about the positive consequences that punishment has, such as the protection of the values of criminal law.

According to the second theory, retributivism, punishment is justified because people deserve it. Offenders should be punished in a way proportionate to their wrongdoing; *an eye for an eye.* Punishment is intrinsically good, and necessary in a justice system, irrespective of its extrinsic benefits or its instrumental benefits.[27] In a human context, often both theories are relevant to warrant punishment.[28]

Humans are very sensitive to retributive considerations. Naturally, people often follow this logic in their behavior.[29] If we could satisfy all the utilitarian aims without the offender having any kind of unpleasant experience, many people feel like that is insufficient. One of the reasons why punishment should involve deliberate harmful or unpleasant consequences is to satisfy those human psychological desires of having an offender suffer in some way. It depends on the circumstances and on the person if such satisfaction is actually realized. Second, the fact that punishment is unpleasant is also why it is effective in protecting many values of criminal law. Therefore, it is important to address this in a robot context: in what way could a robot's experience be of punishment be "unpleasant"?

It is difficult to imagine how a robot would have an unpleasant experience, because generally people are of the view that a robot cannot "feel."[30] However, that issue can be overcome. According to Ying Hu, an experience can be unpleasant for a robot; "if we treat the algorithms of a smart robot as a collective consisting of individuals that are capable of influencing the content of those algorithms, then an imposition qualifies as punishment if it is considered unpleasant by at least some members of that collective."[31] This approach implies that punishment is unpleasant for a robot, if it is experienced as such by the people behind the robot.

This seems like a useful way of addressing this concept. No matter how autonomous the robot is, we cannot—and should not—ignore the people behind it. The way we treat robots will inevitably also have an effect on programmers, operators, and/or owners. With the imposition of punishment, this is an important factor to keep in mind. In this way, punishing a robot will indirectly be an unpleasant experience for the people behind the robot.

Which Values Does Criminal Law Try to Protect?

Values of Criminal Law

Criminal law generally has the function of preserving social order for the benefit and welfare of society.[32] From a utilitarian standpoint, the most important advantageous consequences that can be realized by punishment for a criminal offense are general deterrence, individual deterrence, incapacitation, or other forms of risk management, and reform.[33] Additionally, retributivists think punishment should be imposed as "payback," as retribution.[34]

General Deterrence

The theory of general deterrence articulates the thought that when you know that if you commit a crime, you will be punished, it is less likely that you will do so. The negative consequences—the threat of punishment—following a crime deters people from committing crimes. When people see that others who have committed a crime receive punishment, this has a general deterrent effect. According to Jeremy Bentham, the philosopher behind the theory of utilitarianism, a rational person would see that the benefits that came from a crime would be outweighed by the harm from the punishment.[35]

Individual Deterrence

The effect of individual deterrence follows the logic that if a person is punished for a crime, that same person is less likely to commit another crime. The offender is discouraged to engage in other criminal activity. The negative effects of the punishment should outweigh the positive effects of the crime. If an offender has been punished for one crime and after commits another crime, it can be justified to imply a more harsh punishment, since the first punishment did not reach the desired effect.[36]

Incapacitation and Other Forms of Risk Management

If an offender is put in prison, he cannot commit another crime for the time he is incarcerated. If an offender is charged with the death penalty, he will not be able to ever commit another crime after that punishment has been executed. Also, modes of punishment such as probation or parole, which can be combined with additional requirements or prohibitions can help manage risks. Examples of such additional requirements or prohibitions are random drug testing, prohibition on alcohol use, or possession of firearms.[37]

Reform

Being punished for an offense can have the effect of making the offender a "better person." The punishment itself can have the effect of making the offender realize that he acted wrongfully. It usually will be necessary to do more than just punish an offender to actually make sure this person will not engage in misconduct in the future. Examples of additional measures are rehabilitation such as psychotherapy or medication. Also, training programs or education offered with the goal to provide better alternatives to crime can be effective.[38]

Retribution

Retribution simply means that offenders get what they deserve. Punishment for the fact that they committed a crime.

Restitution

Restitution is a compensation for the committed offense; for example, through monetary compensation, restitution in kind, or moral compensation.

The Protection of the Values of Criminal Law Through Punishment

When a sentence is imposed by a judge, objectives to consider are the protection of society, punishment of the defendant for committing a crime, encouragement for the defendant to lead a law-abiding life, deterring others, isolating the defendant to prevent them from committing other crimes, securing of restitution for the victim, and seeking uniformity in sentencing.[39]

Capital Punishment

Capital punishment, or the death sentence, is the deprivation of one's right to life.[40] This permanently prevents a person from ever committing any other crime and therefore makes recidivism impossible.[41]

It therefore has an individual and in theory, a general deterrent effect. Additionally, it eliminates the risk the offender was posing and protects the value of retribution.

Imprisonment

Imprisonment is the general term used to signify the deprivation of human liberty, severe limitations on human free behavior, freedom of movement, and freedom to manage one's personal life.[42] Incarceration is an unpleasant experience for a human being. This has an individual and general deterrent effect. Also it manages the risk of an offender committing another crime for the time he or she is locked up. Prisons generally also have the goal of rehabilitation, but the effectiveness depends on the program in place. It also works into the goal of retribution.

Suspended Sentencing

Suspended sentencing comes in two forms. Either the imposition or the enforcement of the sentence can be suspended.[43] If a certain type of crime is committed nonetheless, the person will be sentenced to imprisonment for the first offense in addition to sentencing for the second offense. The threat of imprisonment is meant to have an increased deterrent effect on offenders. This discourages recidivism.[44] It can also have the effect of risk management, because suspended sentencing can come with additional obligations, such as random drug testing.[45]

Fine

In criminal law, fines are imposed on humans instead of or in conjunction with another punishment, such as imprisonment or suspended sentencing.[46] A fine constitutes deprivation of someone's property and can be monetary or forfeiture. It also protects the value of retribution, restitution, and has a general and individual deterrent effect.[47]

Community Service

Community service makes it mandatory for a human being to contribute labor to the community. This can also be imposed instead of a fine or imprisonment.[48] This should have a general and individual deterrent effect. It is also retributive and can contribute to the effect of reform.[49]

How Can Punishment Be Applied to Robots to Effectively Protect the Values of Criminal Law?

Keeping in mind the values that the system of criminal law is trying to protect, what would be the best way to translate the human punishments to the robot context?

Capital Punishment

For a human, the death sentence means that a person will never be able to engage in any kind of misconduct after the sentence is carried out. The main goal of capital punishment is deterrence. Recidivism of the individual will be impossible and because of the graveness of this punishment, there are good reasons to believe it will also *generally deter*.

Translated to a robot context, it might depend on the specifics of the robot if it will be necessary for it to be physically destroyed. If the software controlling the robot can be deleted, this will have the same *individual deterrent effect*. With completely new software, the danger of recidivism has been overcome and *risks are managed*. The general deterrent effect here comes from the negative consequences on the people behind the robot. If your self-driving car kills a person and capital punishment is imposed, the car will either be destroyed or you will be left with the car without any software controlling it. Either way, you will have to buy a new car or a new software package. This results in unhappy customers, negative attention, and declining sales. The designers of this car will most likely be forced to take the car off the market to *manage the risks*, until they have improved the software to guarantee this will not happen again. This will result in substantial financial losses. This is an example wherein the people behind the robot cannot be ignored.

Installing new software protects the value of *reform*. A robot can be reactivated and put to use, since the original risks are eliminated. Either the physical destruction of the car, or the negative (financial) effects of people that have programmed, operated, or owned this car can protect the value of *retribution*.

Imprisonment

For humans, the general and individual deterrent effect of incarceration comes from the unpleasant experience. If we put a robot in jail what would the effect be? A robot is deprived of its liberties and its freedom, when it cannot act anymore in the area for which it is designed. In that way, imprisonment of a robot could also mean that it is deactivated for a certain period of time.[50] If a robot is just put in a cell without any other additional measures, this is a form of *risk management*.

Is this unpleasant for a robot? Putting a robot in prison would have an unpleasant effect on its programmers, operators, or owners. Because of the way unpleasantness is approached in this context, this punishment can be seen as unpleasant for a robot. If a robot is put in a cell or deactivated, the people behind the robot are no longer able to utilize it. For a designer, that could be an incentive to program a robot or change its algorithm to make sure they will not engage in this misconduct again. The imprisonment can discourage the operator or owner to steer a robot toward misconduct.

Also, an owner who has purchased a robot that is later deactivated for a certain amount of time because of its misconduct will not be happy. When it turns out certain types robots are likely to engage in misconduct, sales are likely to drop, which will again incentivize designers to make sure their robots comply to legal rules.

However, since all values can be most effectively protected through some form of reprogramming, it is not logical to waste time and resources on a "robo-jail," but rather have an obligation to keep the robot deactivated until it is fully reprogrammed. To deter a robot from repeating misconduct, reprogramming will be enough for *individual deterrence*. *General deterrence* will be secured through the unpleasant effects that it has on programmers, owners, or operators when their robot cannot be put to use anymore. This will provide for incentives to not design or buy robots anymore that engage in misconduct. The market will play a role in the protection of criminal law values.

Reform is also a goal of imprisonment. You want to make sure that when prisoners regain their freedom, they will be better people who will not commit any other crimes. To reform a robot you do not need extensive programs or therapy. We have the opportunity to reform robots in a much more efficient and quicker way, without having to use resources to the same extent as you would with a person in prison, and with much more certainty of reform. Reprogramming seems like an effective way to protect the value of reform. *Retribution* is realized here through the negative consequences that this punishment for the people behind it.

Suspended Sentence

When a suspended sentence is imposed on humans, it is registered in the legal records. Nothing else happens until a second offense is committed. This has an increased deterrent effect on humans. To the contrary, a robot will not be aware of the aggravated consequences of specific misconduct. However, the people behind the robot will be. Programmers will be incentivized to start working on improving the robot to avoid the possible misconduct. Operators will want to avoid the negative consequences. Suspended punishment can have the effect of not only *individual deterrence*, but also *general deterrence*, when the improved software will be updated on all the robots that are running on it.

Fine

In criminal law, fines are imposed on humans instead of or in conjunction with another punishment, such as imprisonment or suspended sentencing. It has the goal of individual and general deterrence.

How does a robot pay a fine? It does not have a bank account or otherwise suffer monetary harm. If it is insured, fines can be paid out of the insurance fund.[51] This has negative consequences for the people behind the robot paying for this insurance, since their fee can rise. It is not unlikely that people will be less satisfied with their purchase of this robot because of the additional costs that they might not have foreseen. Sales can drop and programmers are again incentivized to improve and *reform* their robot. This is a *general deterrent* and, assuming the software controlling the specific robot that has committed the offense will also be updated, also *individually deterring*. Depending on the severity of the misconduct, it may be necessary that the robot itself, and other robots running on the same software, will not be activated until they have been reprogrammed to *manage the risks*. The unpleasant effects of time, effort, rising insurance fees, and people not being able to use their robot, protects the value of *retribution*.

What if a robot is not insured? If a human being cannot pay a fine, another penalty can be imposed on them instead. In the same way, if a robot is not insured, a fine could be substituted for another punishment. Examples of substitutions are imprisonment, probation, or community service. As we have seen, those punishments will also protect the values that fines protect.

Community Service

A person who is sentenced to perform community service has to work for the benefit of the community for a set amount of hours. In the case of a robot, instead of performing its usual tasks, it will be used for the benefit of the community. Reprogramming will in this case be necessary prior to putting a robot toward the good of the community. This is important *risk management*, otherwise there is no guarantee that the robot will not engage in the same misconduct.

It will depend on the specific robot if community service is a realistic possibility. It is not too hard to imagine how a self-driving car could be used in such a way, for example, by driving children with disabilities to school. However, with other types of robots it might be more difficult to find ways to use them for the communities' benefit.

If a robot is used for the good of the community the programmer, operator, or owner will no longer be able to use the robot for their own purposes. The benefits that normally go to a private individual or to a company will now go to the community. Being deprived from using your property freely for your own benefit is unwanted. As with the other types of punishment, it will therefore incentivize improvements in the technology by the programmers and discourage operators steering robots toward misconduct. When a robot is not suitable to be deployed for the benefit of the community, community service could be replaced, for example, by imprisonment (through temporary deactivation) or a fine.

What Are the Results of Imposing Punishment on Robots?

Punishment can be imposed on robots in ways that effectively protect the values of criminal law. Imposing criminal punishment on a robot has the effect of creating negative consequences for the people behind the robot. This can potentially lead to "regulation by design," because of the incentives to improve robot software and ensure its safety.

The negative attention that can come with punishment will most likely damage the reputation of the designers. This will lead to fewer sales and therefore financial loss. Operators could lose their jobs and credibility if they work with robots that are subject to criminal punishment. For the owners, if a robot gets deactivated or destroyed, they will likely suffer financial loss. Even if financial loss will be compensated by insurers or designers, this might not be sufficient. Depending on the robot, the data collected, possibly after years of interaction, could be difficult to replace. Owners might also have to spend time and resources on the investigation into the cause of the misconduct of their robot.[52] All such reasons can be incentives for the people behind the robot to monitor its performance and prevent misconduct.

The possible benefits for the designers or programmers who create a "robo-criminal" must now be weighed against the negative consequences when the robot is punished for criminal conduct.

The people that mostly carry the negative consequences, specifically the designers or programmers, are at the same time the people that are most capable of improving robots and making them safer. The punishments therefore affect the right people. This effect could in the future be strengthened if robots are programmed to adjust their behavior in response to the punishment of other robots. If robots can be interconnected in some way, punishing one robot could have a general deterrent effect. Robots would "learn" from each other's mistakes.[53]

The negative consequences that punishment has on the people behind the robot leads to a positive effect on the community as a whole. Robots are reformed into better versions through reprogramming. In the meantime, robots can be deactivated to prevent them from constituting any risks.

The punishment can be justified from a utilitarian standpoint because of the benefit that it provides to the community. The unpleasant or negative effects that the punishment of the robot has on the programmers, owners, and operators of the robot, can satisfy the human need for some form of retribution.

Conclusion

There is much debate about the desirability of holding robots criminally liable. Legally, there are convincing arguments that robots can and should be liable. Robots are capable of fulfilling the requirements of *actus reus* and *mens rea*. Punishments designed for humans can—where necessary, modified—be imposed on robots in a way that protects the values of criminal law. However, those desired effects come from the negative consequences that the punishment has on the people behind the robot. So who are we actually punishing here? The current nature of robots makes it impossible to apply traditional criminal punishment in a way that has the same direct effects as when punishing humans. This is not surprising, because of the undeniable differences between humans and robots. As a result of those differences, the desired effects are also achieved in a different way. Is that an issue?

Imposing punishments on robots in this way is effective and justifiable. Even though the criminal punishment of robots seems like a disguised punishment of the people behind that robot, it is arguably beneficial for society to do so. By classifying a robot as an appropriate subject of criminal liability, we bridge the liability gap (and impunity) that develops when robots have such a high level of autonomy that we can no longer point a finger at the people behind the robot. This way, we solve the legal issue of criminal liability in the context of robots and simultaneously impose punishment in a way that affects the right people and as such protects the values of criminal law.

Notes

* Nina Scholten holds a master's degree in information law from the University of Amsterdam. She may be contacted at nacescholten@gmail.com.

[1] Caitlin O'Hara, *Self-driving Uber car kills pedestrian in Arizona, where robots roam,* N.Y. Times, Mar. 19, 2018, https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

[2] T.S., *Why Uber's self-driving car killed a pedestrian,* The Economist, May 29, 2018, https://www.economist.com/the-economist-explains/2018/05/29/why-ubers-self-driving-car-killed-a-pedestrian.

[3] Neil M. Richards & William D. Smart, *How the Law Will Think About Robots (and Why You Should Care), in* 2014 IEEE International Workshop on Advanced Robotics and Its Social Impacts (September 2014), at 50.

[4] Ryan Calo, *Robots as Legal Metaphors,* 30 Harvard JOLT 209, 215 (2016).

[5] John Danaher, *Robots, Law and the Retribution Gap,* 18 Ethics and Info. Tech. 299, 302-03 (2016).

[6] Model Penal Code § 6.03 (1) (Am. Law. Inst., Proposed Official Draft 1962).

[7] Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities—From Science Fiction to Legal Social Control,* 4 Akron Intell. Prop. J. 171, 172 (2010).

[8] Ying Hu, *Robot Criminals,* U. Mich. J.L. Reform, Volume 52, issue 2, p. 487-531 (Winter 2019), https://prospectusmjlr.files.wordpress.com/2019/04/robot-criminals.pdf.

[9] Roger C. Schank, *What is AI, Anyway?,* 8 AI Magazine 59, 59-61 (1987).

[10] Ryan Calo, *Robots in American Law,* (University of Washington School of Law Legal Studies Research Paper No. 2016-04), http://euro.ecom.cmu.edu/program/law/08-732/AI/Calo.pdf, at. 6.

[11] Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data,* 78 Ohio State L.J. 1217, 1219 (2017).

[12] Sabine Gless, Emily Silverman & Thomas Weigend, *If Robots Cause Harm, Who is to Blame: Self-Driving Cars and Criminal Liability,* 19 New Crim. L. Rev. 412, 416 (2016).

[13] Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 10 (Spring 2010).

[14] Joshua Dressler & Stephen P. Garvey, Criminal law—cases and materials 133 (West Academic Publishing, 7th ed. 2016).

[15] *Id.* at 133.

[16] Model Penal Code § 2.01 (Am. Law. Inst., Proposed Official Draft 1962).

[17] Pamela H. Bucy, *Corporate Ethos: A Standard for Imposing Corporate Criminal Liability,* 75 Minn. L. Rev. 1095, 1121 (1990).

[18] Ying Hu, *Robot Criminals,* U. Mich. J.L. Reform, Volume 52, issue 2, p. 487-531 (Winter 2019), https://prospectusmjlr.files.wordpress.com/2019/04/robot-criminals.pdf.

[19] Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 7-8 (Spring 2010).

20. Model Penal Code § 2.02 (3)(d) (Am. Law. Inst., Proposed Official Draft 1962).

21. Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities—From Science Fiction to Legal Social Control,* 4 Akron Intell. Prop. J. 171, 188 (2010).

22. Joshua Dressler & Stephen P. Garvey, Criminal law—cases and materials 164 (West Academic Publishing, 7th ed. 2016).

23. See Ian Sample, *Give robots an "ethical black box" to track and explain decisions, say scientists,* The Guardian, Jul 19, 2017, for reasons we should fit robots with such a black box in the future.

24. Sabine Gless, Emily Silverman & Thomas Weigend, *If Robots Cause Harm, Who is to Blame: Self-Driving Cars and Criminal Liability,* 19 New Crim. L. Rev. 412, 423 (2016).

25. Joshua Dressler & Stephen P. Garvey, Criminal law—cases and materials 35 (West Academic Publishing, 7th ed. 2016).

26. *Id.* at 34-35.

27. *Id.* at 41.

28. *Id.* at 34.

29. John Danaher, *Robots, Law and the Retribution Gap,* 18 Ethics and Info. Tech. 299, 301 (2016).

30. *Id.* at 319.

31. Ying Hu, *Robot Criminals,* U. Mich. J.L. Reform, Volume 52, issue 2, p. 487-531 (Winter 2019), https://prospectusmjlr.files.wordpress.com/2019/04/robot-criminals.pdf.

32. Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 36 (Spring 2010).

33. Joshua Dressler & Stephen P. Garvey, Criminal law—cases and materials 38 (West Academic Publishing, 7th ed. 2016).

34. *Id.* at 42.

35. *Id.* at 38.

36. *Id.* at 38-39.

37. *Id.* at 39.

38. Joshua Dressler & Stephen P. Garvey, Criminal law—cases and materials 39 (West Academic Publishing, 7th ed. 2016).

39. *Id.* at 58.

40. Model Penal Code § 210.6 (Am. Law. Inst., Proposed Official Draft 1962). Withdrawn from the Model Penal Code on the basis of the *Report of the Council to the Membership of the American Law Institute On the Matter of the Death Penalty* in 2009.

41. Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 30 (Spring 2010).

42. *Id.* at 31.

[43] Herbert C. Parson, *Probation and Suspended Sentence,* 8 J. Crim. L. and Criminology 694, 702 (1918).

[44] Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 32 (Spring 2010).

[45] Model Penal Code § 6.02 (3) (d) (Am. Law. Inst., Proposed Official Draft 1962).

[46] Model Penal Code § 6.03 (Am. Law. Inst., Proposed Official Draft 1962).

[47] Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 34 (Spring 2010).

[48] Model Penal Code: Sentencing (Am. Law Inst., Proposed Final Draft, April 10 2017), at 62.

[49] Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1, 33 (Spring 2010).

[50] *Id.* at 31-32.

[51] Ying Hu, *Robot Criminals,* U. Mich. J.L. Reform, Volume 52, issue 2, p. 487-531 (Winter 2019), https://prospectusmjlr.files.wordpress.com/2019/04/robot-criminals.pdf.

[52] *Id.* at 17.

[53] *Id.* at 16.

Bibliography

Cited

Jack M. Balkin, *2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data,* 78 Ohio State L.J. 1217 (2017).

Pamela H. Bucy, *Corporate Ethos: A Standard for Imposing Corporate Criminal Liability,* 75 Minn. L. Rev. 1095 (1990).

Ryan Calo, *Robots as Legal Metaphors,* 30 Harvard JOLT 209 (2016).

Ryan Calo, *Robots in American Law,* (University of Washington School of Law Legal Studies Research Paper No. 2016-04), http://euro.ecom.cmu.edu/program/law/08-732/AI/Calo.pdf.

John Danaher, *Robots, Law and the Retribution Gap,* 18 Ethics and Info. Tech. 299 (2016).

Joshua Dressler & Stephen P. Garvey, *Criminal law—cases and materials* (West Academic Publishing, 7th ed. 2016).

Sabine Gless, Emily Silverman & Thomas Weigend, *If Robots Cause Harm, Who is to Blame: Self-Driving Cars and Criminal Liability,* 19 New Crim. L. Rev. 412 (2016).

Gabriel Hallevy, *I, Robot, I, Criminal—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses,* 22 Syracuse Science & Tech. L. Rep. 1 (Spring 2010).

Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities—From Science Fiction to Legal Social Control,* 4 Akron Intell. Prop. J. 171 (2010).

Ying Hu, *Robot Criminals,* U. Mich. J.L. Reform, Volume 52, issue 2, p. 487-531 (Winter 2019), https://prospectusmjlr.files.wordpress.com/2019/04/robot-criminals.pdf.

Model Penal Code (Am. Law. Inst., Proposed Official Draft 1962).

Model Penal Code: Sentencing (Am. Law Inst., Proposed Final Draft, April 10 2017).

Caitlin O'Hara, *Self-driving Uber car kills pedestrian in Arizona, where robots roam,* N.Y. Times, Mar. 19, 2018, https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

Herbert C. Parson, *Probation and Suspended Sentence,* 8 J. Crim. L. and Criminology 694, 702 (1918).

Neil M. Richards & William D. Smart, *How the Law Will Think About Robots (and Why You Should Care), in* 2014 IEEE International Workshop on Advanced Robotics and its Social Impacts (September 2014).

T.S., *Why Uber's self-driving car killed a pedestrian,* The Economist, May 29, 2018, https://www.economist.com/the-economist-explains/2018/05/29/why-ubers-self-driving-car-killed-a-pedestrian.

Ian Sample, *Give robots an 'ethical black box' to track and explain decisions, say scientists,* The Guardian, Jul 19, 2017, https://www.theguardian.com/science/2017/jul/19/give-robots-an-ethical-black-box-to-track-and-explain-decisions-say-scientists.

Roger C. Schank, *What is AI, Anyway?,* 8 AI Magazine 59 (1987).

Consulted

Rachel Charney, *Can Androids Plead Automatism?* A Review of *When Robots Kill: Artificial Intelligence Under the Criminal Law by Gabriel Hallevy,* 73 U. Toronto Fa L.Rev. 69 (2015).

Brent Fisse, *Reconstructing Corporate Criminal Law: Deterrence, Retribution, Fault, and Sanctions,* 56 S. Cal. L. Rev. 1141 (1983).

Pedro M. Freitas, Fransisco Andrade & Paulo Novais, *Criminal Liability of Autonomous Agents: from the unthinkable to the plausible in* 8929 LNCS (Springer 2013).

Gabriel Hallevy, *Virtual Criminal Responsibility,* 6 Orig. L. Rev. 6 (2010).

Mark C. Materni, *Criminal Punishment and the Pursuit of Justice,* 2 Br. J. Am. Leg. Studies 263 (2013).

Ugo Pagallo, *Apples, oranges, robots: four misunderstandings in today's debate on the legal status of AI systems,* 376 Phil. Trans. R. Soc. (2018).

Ugo Pagallo, *Robots of Just War: A Legal Perspective,* 24 Philos. Technol 307 (2011).

Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies,* 29 Harv. J.L. & Tech 353 (2016).

Kip Schlegel, *Desert, Retribution, and Corporate Criminality,* 5 Just. Quart. 615 (December 1988).