1

The Challenges of Human–Robot Interaction for Substantive Criminal Law

Mapping the Field

TATJANA HÖRNLE*

I Mapping the Field: Preliminary Remarks

Technological innovations are likely to increase the frequency of human–robot interactions in many areas of social and economic relations and humans' private lives. Criminal law theory and legal policy should not ignore these innovations. Although the main challenge is to design civil, administrative, and soft law instruments to prevent harm in human–robot interactions and to compensate victims, the developments will also have some impact on substantive criminal law. Criminal laws[1] should be scrutinized and, if necessary, amendments and adaptations recommended, taking the two dimensions of criminal law and criminal law theory, the preventive and the retrospective, into account.

The prevention of accidents is obviously one of the issues that needs to be addressed, and regulatory offenses in the criminal law could contribute to this end. Regulatory offenses are part of a larger legal toolbox that can be called upon to prevent risks and harms caused by malfunctioning technological innovations and unforeseen outcomes of their interactions with human users (see Section II.A). In addition to the risk of accidents, some forms of human–robot interaction, such as automated weapon systems and sex robots, are also criticized for other reasons, which invites the

* I would like to thank Emily Silverman for improving the language of this chapter.

[1] The category "criminal law" is used here in a wide sense, encompassing all norms that prohibit conduct and prescribe sanctions for noncompliance. Details and distinctions, e.g., between criminal offenses in a narrow sense and administrative offenses (*Ordnungswidrigkeiten*) in German law, are not discussed here. They will, however, play a role once prohibitions are seriously considered, and then, notions such as proportionality or *ultima ratio* become relevant and the kind and seriousness of potential sanctions need more thought.

question of whether these types of robots should be banned (Section II.B). If we turn to the second, retrospective dimension of criminal law, the major question, again, is liability for accidents. Under what conditions can humans who constructed, programmed, supervised, or used a robot be held criminally liable for harmful outcomes caused by the robot (Section III.A)? Other questions are whether existing criminal laws can be applied to humans who commit crimes with robots as tools (Section III.B), how dilemmatic situations should be evaluated (Section III.C), and whether self-defense against robots is possible (Section III.D). From the perspective of criminal law theory, the scope of inquiry should be even wider and extend beyond questions of criminal liability of humans for harmful events involving robots. Might it someday be possible for robots to incur criminal liability (Section III.E)? Could robots be victims of crime (Section III.F)? And, as robots become increasingly involved in the day-to-day life of humans and become subject to legal responsibility, might this also have a long-term impact on how human–human interactions are understood (Section IV)?

The purpose of this introductory chapter is to map the field in order to structure current and future discussions about human–robot interactions as topics for substantive criminal law. Marta Bo, Janneke de Snaijer, and Thomas Weigend analyze some of these issues in more depth in their chapters. Before we turn to the mapping exercise, the term "robot" deserves some attention,[2] including delineation from the broader concept of artificial intelligence (AI). Per the Introduction to the volume, which references the EU AI Act, AI is "software that is developed with one or more of [certain] approaches and techniques … and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."[3] The consequences of the growing use of information technology (IT) and AI are discussed in many areas of law and legal policy.[4] In the field of criminal justice, AI systems can be utilized at the pre-trial and sentencing stages as well

[2] See also Monika Simmler & Nora Markwalder, "Guilty Robots? Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence" (2019) 30:1 *Criminal Law Forum* 1 ["Guilty Robots"] at 5–6.

[3] European Union, European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts, COM/2021/206 final (Brussels, Belgium: European Commission, April 21, 2021).

[4] See e.g., Horst Eidenmüller & Gerhard Wagner, *Law by Algorithm* (Heidelberg, Germany: Mohr Siebeck, 2021) [*Law by Algorithm*].

as for making decisions about parole, to provide information on the risk of reoffending.[5] Whether these systems analyze information more accurately and comprehensively than humans, and the degree to which programs based on machine learning inherit biases, are issues under discussion.[6] The purpose of this volume is not to examine the relevance of these new technologies to criminal law and criminal justice in general; the focus is somewhat narrower. Robots are the subject. Entities that are called robots can be based on machine learning techniques and AI, technologies already in use today, but they also have another crucial feature. They are designed to perform actions in the real word[7] and thus must usually be embodied as physical objects. It is primarily this ability to interact physically with environments, objects, and the bodies of humans that calls for safeguards.

II The Preventive Perspective: Regulating Human–Robot Interactions

II.A Preventing Accidents

Regulation is necessary to prevent accidents caused by malfunctioning robots and unforeseen interactive effects. Some of these rules might need to be backed up by sanctions. It is almost impossible to say much more on a general level about potential accidents and what should be prohibited or regulated to minimize the risk of harm, as a more detailed analysis would require covering a vast area. The exact nature of important "dos and don'ts" that might warrant enforcement by criminal laws obviously depends on the kinds of activities that robots perform, e.g., in manufacturing, transportation, healthcare, households, and warfare, and the potential risks involved. The more complex a robot's task, the more that can go wrong. The kind and size of potential harm depends, among other things,

[5] For such instruments, see Sheldon Zhang, Robert Roberts, & David Farabee, "An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures" (2014) 60:2 *Crime and Delinquency* 167; Carolyn McKay, "Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making" (2020) 32:1 *Current Issues in Criminal Justice* 22; Lucia Sommerer, *Personenbezogenes Predictive Policing* (Baden-Baden, Germany: Nomos, 2020).

[6] See e.g., Solon Barocas & Andrew Selbst, "Big Data's Disparate Impact" (2016) 104:3 *California Law Review* 671; Richard Berk, Hoda Heidari, Shahin Jabbari *et al.*, "Fairness in Criminal Justice Task Assessments: The State of the Art" (2017) 50:1 *Sociological Methods & Research* 3; John Kleinberg, Himabindu Lakkaraju, Jens Ludwig *et al.*, "Human Decisions and Machine Predictions" (2018) 133:1 *Quarterly Journal of Economics* 237.

[7] See Erico Guizzo, "What Is a Robot?" *IEEE* (August 1, 2018), https://robots.ieee.org/learn/what-is-a-robot/.

on the physical properties of robots, such as weight and speed, the frequency with which they encounter the general public, and the closeness of their operations to human bodies. Autonomous vehicles and surgical robots, e.g., require tighter regulation than robot vacuum cleaners.

The task of developing proper regulations for potentially dangerous human–robot interaction is challenging. It begins with the need to determine the entity to whom rules and prohibitions are addressed: manufacturers; programmers; those who rely on robots as tools, such as owners or users; third parties who happen to encounter robots, e.g., in the case of automated cars, other road users; or malevolent intruders who, e.g., hack computer systems or otherwise manipulate the robot's functions. Another question is who can – and who should – develop legal standards. Not only legislatures, but also criminal and civil courts can and do contribute to rule-setting. Their rulings, however, generally target a specific case. Systematic and comprehensive regulation seems to call for legislative action. But before considering the enactment of new laws, attention should be paid to existing criminal laws, i.e., general prohibitions that protect human life, bodily integrity, property, etc. These prohibitions can be applied to some human failures that involve robots, but due to their unspecific wording and broad scope, they do not give sufficient guidance for our scenarios. More specific norms of conduct, norms tailored to the production, programming, and use of robots, would certainly be preferable. This leads again to the question of what institution is best situated to develop these norms of conduct. This task requires constant attention to and monitoring of rapid technological developments and emerging trends in robotics. Ultimately, traditional modes of regulation by means of laws might not be ideally suited to respond effectively to emerging technologies. Another major difficulty is that regulations in domestic laws do not make much sense for products circulating in global markets. This may prompt efforts to harmonize national laws.[8] As an alternative, soft law in the form of standards and guidelines proposed by the private sector or regulatory agencies might be a way to achieve faster and perhaps also more universal agreement among the producers and users of robots.[9]

For legal scholars and legal policy, the upshot is that we should probably not expect too much from substantive criminal law as an instrument

[8] See *Feasibility Study of a Future Council or Europe Instrument on Artificial Intelligence and Criminal Law* (European Committee on Crime Problems, September 4, 2020).

[9] Gary Marchant & Brad Allenby, "Soft Law: New Tools for Governing Emerging Technologies" (2017) 73:2 *Bulletin of the Atomic Scientists* 108; Ryan Hagemann, Jennifer Huddleston, & Adam Thierer, "Soft Law for Hard Problems: The Governance of Emerging

to control the use of new technologies. Effective and comprehensive regulation to prevent harm arising out of human–robot interactions, and the difficult task of balancing societal interest in the services provided by robots against the risks involved, do not belong to the core competencies of criminal law.

II.B Beyond Accidents

Beyond the prevention of accidents, other concerns might call for criminal prohibitions. If there are calls to suppress certain conduct rather than to regulate it, the criminal law is a logical choice. Strict prohibitions would make sense if one were to fundamentally object to the creation of AI and autonomous robots, in part because the long-term consequences for humankind might be serious,[10] although it may be too late for that in some instances. A more selective approach would be to demand not a categorical decision against all research in the field of AI and the production of advanced robots in general, but rather efforts to suspend research[11] or to stop the production of some kinds of robots. An example of the latter approach would be prohibiting devices that apply deadly force against humans, such as remotely controlled or automated weapons systems, addressed in this volume by Marta Bo.[12] Not only is the possibility of accidents a particularly serious concern in this area, but also the reliability of target identification, the precision of application, and the control of access are of utmost importance. Even if autonomous weapon systems work as intended, they might in the long run increase the death toll in wars, and ethical doubts regarding war might grow if the humans responsible for aggressive military operations do not face personal risks.[13]

Technologies in an Uncertain Future" (2018) 17:1 *Colorado Technology Law Journal* 37; Anna Thaler, *Values and Ethical Principles for AI and Robotics: A Qualitative Content Analysis of EU Soft Law Initiatives* (Hamburg, Germany: Verlag Dr. Kovač, 2021).

[10] See, for possible future risks, Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (New York, NY: Oxford University Press, 2014).

[11] For a proposal signed by prominent AI researchers and entrepreneurs, see "Pause Giant AI Experiments: An Open Letter," *Future of Life*, https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[12] See Chapter 2 in this volume; see also: Jai Galliot, *Military Robots: Mapping the Moral Landscape* (Abingdon, UK: Routledge, 2017); Paul Springer, *Outsourcing War to Machines: The Military Robotics Revolution* (Santa Barbara, CA: Praeger, 2018); Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York, NY: W.W. Norton & Company, 2018) [*Army of None*].

[13] For an overview of the ethical issues, see Nehal Bhuta, Susanne Beck, Robin Geis *et al.* (eds.), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge, UK: Cambridge University Press, 2016); *Army of None*, note 12 above, at 271–296.

Arguments that point to the risk of remote harm are often based on moral concerns. This is most evident in the discussions about sex robots. Should sex robots in general or, more particularly, sex robots that imitate stereotypical characteristics of female prostitutes, be banned?[14] The proposition of such prohibitions would need to be supported by strong empirical and normative arguments, including explanations as to why sex robots are more problematic than sex dolls, whether it is plausible to expect such robots to have negative effects on a sizable number of persons, why sexual activity involving humans and robots is morally objectionable, and even if convincing arguments of this kind could be made, why the state should engage in the enforcement of norms regarding sexual morality.

For legal theorists, it is also interesting to ask whether, at some point, policy debates will no longer focus solely on remote harms to other human beings, collective human concerns such as gender equality, or human values and morals, but will instead expand to include the interests or rights of individual robots as well. Take the example of sex robots. Could calls to prohibit sexual interactions between humans and robots refer to the dignity of the robot and its right to dignity? Might we experience a re-emergence of debates about slavery? At present, it would certainly be premature to claim that humans and robots should be treated as equivalent, but discussions about these issues have already begun.[15] As long as robots are distinguishable from humans in several dimensions, such as bodies, social competence, and emotional expressivity, it is unlikely that the rights humans grant one another will be extended to them. As long as there are no truly humanoid robots, i.e., robots that resemble humans in all or most physiological and psychological dimensions,[16] tremendous cognitive abilities alone are unlikely to trigger widespread demands for equal treatment such as the recognition of robots' rights. For the purpose

[14] Campaign against Sex Robots website, https://campaignagainstsexrobots.org/; Oliver Bendel, "Love Dolls and Sex Robots in Unproven and Unexplored Fields of Application" (2020) 12:1 *Paladyn, Journal of Behavioral Robotics* 1.

[15] See e.g., Phil McNally & Sohail Inayatullah, "The Rights of Robots: Technology, Culture and Law in the 21st Century" (1988) 20:2 *Futures* 119; Mark Coeckelbergh, "Robot Rights? Towards a Social-Relational Justification of Moral Consideration" (2010) 12:3 *Ethics and Information Technology* 209; David Gunkel, *Robot Rights* (Cambridge, MA: MIT Press, 2018); Henry Shevlin, "How Could We Know When a Robot Was a Moral Patient?" (2021) 30:3 *Cambridge Quarterly of Healthcare Ethics* 459; John Danaher, "What Matters for Moral Status: Behavioural or Cognitive Equivalence?" (2021) 30:3 *Cambridge Quarterly of Healthcare Ethics* 472.

[16] See, for an example from fiction, Ian McEwan, *Machines Like Me* (London, UK: Penguin Books, 2019).

of this introductory chapter, it must suffice to point out that thinking in this direction would also be relevant to debates concerning the need to criminalize selected conduct in order to protect the interests of robots.

III The Retrospective Perspective: Applying Criminal Law to Human–Robot Interactions

The harmful outcomes of human–robot interactions not only provide an impetus to consider creating preventive regulation. Harmful outcomes can also give rise to criminal investigations and, ultimately, to proceedings against the humans involved. The criminal liability of robots is also discussed below.

III.A Human Liability for Unforeseen Accidents

III.A.1 Manufacturers and Programmers

If humans have been injured or killed through interaction with a robot, if property has been damaged, or if other legally protected rights have been disregarded, questions of criminal liability will arise. It could, of course, be argued that the more pressing issue is effective compensation, a goal achievable by means of tort law and mandatory insurance, perhaps in combination with the legal construct of robots as "electronic persons" with their own assets.[17] Serious accidents, however, are also likely to engage criminal justice officials who need to clarify whether a human suspect or, depending on the legal system, a corporation has committed a criminal offense.

The first group of potential defendants could be those who built and programmed the robot. If the applicable criminal law does not include a strict liability regulatory offense, criminal liability will depend on the applicability of general norms, such as those governing negligent or reckless conduct. The challenges for prosecutors and courts are manifold, and they include establishing causality, attributing outcomes to acts and

[17] See, for the idea of an electronic person, European Union, The European Parliament, Resolution of February 16, 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), OJ 2015 C 252 (EU: Official Journal of the European Union, 2017) at No. 59(f); Susanne Beck, "Intelligent Agents and Criminal Law – Negligence, Diffusion of Liability and Electronic Personhood" (2016) 86:4 *Robotics and Autonomous Systems* 138 ["Intelligent Agents"] at 141–142; Jacob Turner, "Legal Personality for AI" in Jacob Turner, *Robot Rules* (London, UK: Palgrave, 2018) ["Legal Personality for AI"] 173; *Law by Algorithm*, note 4 above, at 103–126.

omissions, and specifying the standard of care that applied to the defendant's conduct.[18] Determining the appropriate standard of care requires knowledge of what could have been done better on the technical level. In addition, difficult, wide-ranging normative considerations are relevant. How much caution do societies require, and how much caution may they require when innovative products such as robots are introduced?[19] As a general rule, standards of care should not be so strict as to have a chilling effect on progress, since manufacturers and programmers can relieve humans of manual, tiresome, and tedious work, robots can compensate for the lack of qualified employees in many areas, and the overall effect of robot use can be beneficial to the public, e.g., by reducing traffic accidents once the stage of automated driving has been reached. Such fundamental issues of social utility should be one criterion when determining the standards of care upon which the criminal liability of manufacturers and programmers are predicated.[20]

Marta Bo focuses on the criminal liability of programmers in Chapter 2, "Are Programmers in or out of Control? The Individual Criminal Responsibility of Programmers of Autonomous Weapons and Self-Driving Cars." She asks whether programmers could be accused of crimes against persons if automated cars or automated weapons cause harm to humans or if the charge of indiscriminate attacks against civilians can be made. She describes the challenges facing programmers of automated vehicles and autonomous weapons and discusses factors that can undermine their control over outcomes. She then turns her attention to legal assessments, including criteria such as *actus reus*, the causal nexus between programming and harm caused by automated vehicles and autonomous weapons, and negligence standards. Bo concludes that it is possible to use criminal law criteria for imputation to test whether programmers had "meaningful human control."

An obvious challenge for criminal law assessment is to determine the degree to which, in the case of machine learning, programmers can foresee developments in a robot's behavior. If the path from the original algorithm to the robot's actual conduct cannot be reconstructed, it might be worth considering whether the mere act of exposing humans to encounters with a somewhat unpredictable and thus potentially dangerous robot

[18] See "Intelligent Agents", note 17 above, at 139.

[19] See, for the notion of "admissible risk," "Intelligent Agents", note 17 above, at 141.

[20] Sabine Gless, Emily Silverman, & Thomas Weigend, "If Robots Cause Harm, Who is to Blame? Self-Driving Cars and Criminal Liability" (2016) 19:3 *New Criminal Law Review* 412 ["If Robots Cause Harm"] at 433–434.

could, without more, be labeled criminally negligent. While this might be a reasonable approach when such robots first appear on the market, the question of whether it would be a good long-term solution merits careful consideration. It seems preferable to focus on strict criteria for licensing self-learning robots, and on civil law remedies such as compensation that do not require proof of individual negligence, and abandon the idea of criminal punishment of humans just for developing and marketing robots with self-learning features.

III.A.2 Supervisors and Users

Humans who are involved in a robot's course of action in an active cooperative or supervisory way could, if an accident occurs, incur criminal liability for recklessness or negligence. Again, for prosecutors and courts, a frequent problem will be to identify the causes of an accident and the various roles of the numerous persons involved in the production and use of the robot. A "diffusion of responsibility"[21] is almost impossible to avoid. Also, the question will arise as to what can realistically be expected of humans when they supervise and use robots equipped with AI and machine learning technology. How can they keep up with self-learning robots if the decision-making processes of such robots are no longer understandable and their behavior hard to predict?[22]

In Chapter 3, "Trusting Robots: Limiting Due Diligence Obligations in Robot-Assisted Surgery under Swiss Criminal Law," Janneke de Snaijer describes one area where human individuals might be held criminally liable as a consequence of using robots. She focuses on the potential and the challenges of robot-assisted surgery. The chapter introduces readers to a technology already in use in operating rooms: that of automated robots helping surgeons achieve greater surgical precision. These robots can perform limited tasks independently, but are not fully autonomous. De Snaijer concentrates primarily on criminal liability for negligence, which depends on how the demands of due diligence are defined. She describes general rules of Swiss criminal law doctrine that provide some guidelines for requirements of due diligence. The major problem she identifies is how much trust surgeons should be allowed to place in the functioning of the robots with which they cooperate. Concluding that

[21] Susanne Beck, "Google Cars, Software Agents, Autonomous Weapons Systems – New Challenges for Criminal Law?" in Eric Hilgendorf & Uwe Seidel (eds.), *Robotics, Autonomics, and the Law* (Baden-Baden, Germany: Nomos, 2017) 227 ["Google Cars"] at 245.

[22] Ibid. at 243.

Swiss law holds surgeons accountable for robots' actions to an unreasonable degree, she diagnoses contradictory standards in that surgeons are held responsible but required by law to use new technology to improve the quality of surgery.

In other contexts, robots are given the task of monitoring those who use them, e.g., by detecting fatigue or alcohol consumption, and, if need be, issuing warnings. Under such circumstances, a human who fails to heed a warning and causes an accident may face criminal liability. Presuming negligence in such cases might have the effect of establishing a higher standard for humans carrying out an activity while under the surveillance of a robot than for humans carrying out the same activity without the surveillance function. It might also mean that the threshold for assuming recklessness, or, under German law, conditional intent,[23] will be lowered. An interesting question is the degree to which courts will allow leeway for human psychology, including perhaps a human disinclination to be bossed around by a machine.

III.A.3 Corporate Liability

In many cases, it will not be possible or very difficult to trace harm caused by a device based on artificial intelligence to the wrongful conduct of an individual human being who acted in the role of programmer, manufacturer, supervisor, or user. Thomas Weigend starts Chapter 4, entitled "Forms of Robot Liability: Criminal Robots and Corporate Criminal Responsibility," with the diagnosis of a "responsibility gap." He then examines the option of holding robots criminally liable before going a step further and considering the introduction of corporate criminal responsibility for the harmful actions of robots. Weigend begins with the controversial discussion of whether corporations should be punished for crimes committed by employees. He then develops the idea that the rationales used to justify the far-reaching attribution of employee conduct to corporations could be applied to the conduct of robots as well. He contends that criminal liability should be limited to cases in which humans acting on behalf of the corporation were (at a minimum) negligent regarding the designing, programming, or controlling of robots.

[23] See, for the notion of conditional intent in German criminal law: Michael Bohlander, *Principles of German Criminal Law* (Oxford, UK: Hart, 2009) [*German Criminal Law*] at 63–67; Tatjana Hörnle & Rita Vavra, "Criminal Law" in Joachim Zekoll & Gerhard Wagner (eds.), *Introduction to German Law*, 3rd ed. (Philadelphia, PA: Wolters Kluwer, 2019) ["Criminal Law"] 503 at 509.

III.B Human Liability for the Use of a Robot with the Intent to Commit a Crime

Robots can be purposefully used to commit crimes, e.g., to spy on other persons.[24] If the accused human intentionally designed, manipulated, used, or abused a robot to commit a crime, he or she can be held criminally liable for the outcome.[25] The crucial point in such cases is that the human who employs the robot uses it as a tool.[26] If perpetrators pursue their criminal goals with the use of a tool, it does not matter whether the tool is of the traditional, merely mechanical kind, such as a gun, or whether it has some features of intelligence, such as an automated weapon that is, e.g., reprogrammed for a criminal purpose.

While this is clearly the case for many criminal offenses, particularly those that focus on outcomes such as causing the death of another person, the situation with regard to other criminal offenses is not so clear. It will not always be obvious that a robot will be able to fulfil the definitional elements of all offenses. It could, e.g., be argued that sexual offenses that require bodily contact between offender and victim cannot be committed if the offender causes a robot to touch another person in a sexual way. In such cases, it is a matter of interpretation if wrongdoing requires the physical involvement of the human offender's body. I would answer this particular question in the negative, because the crucial point is the penetration of the victim's body. However, answers must be developed for different crimes separately, based on the legal terminology used and the kind of interest protected.

III.C Human Liability for Foreseen but Unavoidable Harm

In the situation of an unsolvable, tragic dilemma, in which there is no alternative harmless action, a robot might injure humans as part of a planned course of action. The most frequently discussed examples of these dilemmas involve automated cars in traffic scenarios in which all available options, such as staying on track or altering course, will lead to a crash with human victims.[27] If such events have been anticipated by human programmers, the question

[24] See, for the potential of service robots to be used this way, "Google Cars", note 21 above, at 231.

[25] "Legal Personality for AI", note 17 above, at 118; "If Robots Cause Harm", note 20 above, at 425.

[26] For a discussion of characterization of robots as a tool, see Chapter 13 in this volume.

[27] For this dilemma, see Dietmar Hübner & Lucie White, "Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us beyond Harm Minimisation" (2018) 21:3 *Ethical Theory and Moral Practice* 685; Rob Lawlor, "The Ethics of Automated

arises of whether they could perhaps be held criminally liable, should the dilemmatic situation in fact occur. When human drivers in a comparable dilemma knowingly injure others to save their own lives or the lives of their loved ones, criminal law systems recognize defenses that acknowledge the psychological and normative forces of strong fear, the will to survive, and personal attachments.[28] The rationale of such defenses does not apply, however, if a programmer, who is not in acute distress, decides that the automated car should always safeguard passengers inside the vehicle, and thus chooses the course that will lead to the death of humans outside the car.

If a human driver has to choose between swerving to save the lives of two persons on the road directly in front of the car, thus hitting and killing a single person on the sidewalk, or staying the course, thus hitting and killing both persons on the road, criminal law doctrine does not provide clear-cut answers. Under German doctrine, which displays a built-in aversion to utilitarian reasoning, the human driver who kills one person to save two would risk criminal punishment.[29] Whether this would change once the assessment shifts from the human driver at the wheel of the car at the crucial moment to the vehicle's programmer is an interesting question. German law is shaped by a strong preference for remaining passive, i.e., one may not become active in order to save the greater number of lives, but for the programmer, this phenomenological difference dissolves completely. At the time the automated car or other robot is manufactured, it is simply a decision between programming option A or programming option B for dilemmatic situations.[30]

Vehicles: Why Self-Driving Cars Should Not Swerve in Dilemma Cases" (2021) 28:1 *Res Publica* 193; and Chapter 15 in this volume.

[28] See *Strafgesetzbuch* (German Criminal Code) (StGB), Germany (November 13, 1998 (Federal Law Gazette I, p. 3322), as amended by Art. 2 of the Act of June 19, 2019 (Federal Law Gazette I, p. 844)) [StGB], §35 (excusing necessity); and David Ormerod & Karl Laird, *Smith, Hogan, and Ormerod's Criminal Law*, 15th ed. (New York, NY: Oxford University Press, 2018) at 364–367 for the "duress of circumstances" doctrine in English law.

[29] See StGB, note 28 above, §34; from the viewpoint of legal philosophy, Ivó Coca Vila, "Self-Driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law" (2018) 12:1 *Criminal Law and Philosophy* 59 at 64–66; see for a more critical perspective on the anti-utilitarian German stance, Eric Hilgendorf, "Automated Driving and the Law" in Eric Hilgendorf & Uwe Seidel (eds.), *Robotics, Autonomics, and the Law* (Baden-Baden, Germany: Nomos, 2017) 171 at 190; and for an empirical analysis that shows the human preference for saving the greater number of humans, Anja Faulhaber, Anke Dittmer, Felix Blind *et al.*, "Human Decisions in Moral Dilemmas Are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles" (2019) 25:2 *Science and Engineering Ethics* 399.

[30] Tatjana Hörnle & Wolfgang Wohlers, "The Trolley Problem Reloaded. Wie sind autonome Fahrzeuge für Leben-gegen-Leben-Dilemmata zu programmieren?" (The Trolley

III.D Self-Defense against Robots

If a human faces imminent danger of being injured or otherwise harmed by a robot, and the human knowingly or purposefully damages or destroys that robot, the question arises as to whether this situation is covered by a justificatory defense. In some cases, a necessity/lesser evil defense could be raised successfully if the danger is substantial. In other cases, it could be questioned if a lesser evil defense would be applicable, e.g., if someone shoots down a very expensive drone to prevent it from taking pictures.[31] Under such circumstances, another justificatory defense might be that of self-defense. In German criminal law, self-defense does not require a proportionality test.[32] In the case of an unlawful attack, it is permissible to destroy valuable objects even if the protected interest might be of comparatively minor importance. The crucial question in the drone case is whether an "unlawful attack"[33] or "unlawful force by another person"[34] requires that the attacker is a human being.

III.E Criminal Liability of Robots

In the realm of civil liability, robots could be treated as legal persons, and this status could be combined with the duty of producers or owners to endow robots with sufficient funds to compensate potential accident victims.[35] A different question is whether a case could also be

Problem Reloaded. How Should Autonomous Vehicles Be Programmed for the Case of a Life-against-Life Dilemma?) (2018) 165:1 *Goltdammer's Archiv für Strafrecht* 12 at 23–24; Thomas Weigend, "Notstandsrecht für Selbstfahrende Autos?" (Emergency Law for Self-Driving Cars?) (2017) 10 *Zeitschrift für Internationale Strafrechtdogmatik* 599.

[31] Regarding questions of self-defense, see Michael Froomkin & Zak Colangelo, "Self-Defense against Robots and Drones" (2015) 48:1 *Connecticut Law Review* 1; Severin Löffler, "Rechtswidrigkeit der Abwehr von Drohnen über privaten Wohngrundstücken" (Lawfulness of Defense against Drones above Private Property) in Susanne Beck, Carsten Kusche, & Brian Valerius (eds.), *Digitalisierung, Automatisierung, KI und Recht* (Baden-Baden, Germany: Nomos, 2020) 329.

[32] *German Criminal Law*, note 23 above, at 104.

[33] StGB, note 28 above, §32; "Google Cars", note 21 above, at 236 and 242; Wolfgang Mitsch, "Roboter und Notwehr" (Robots and Self-Defense) in Susanne Beck, Carsten Kusche, & Brian Valerius (eds.), *Digitalisierung, Automatisierung, KI und Recht* (Baden-Baden, Germany: Nomos, 2020) 365.

[34] American Law Institute, Model Penal Code: Official Draft and Explanatory Notes: Complete Text of Model Penal Code as Adopted at the 1962 Annual Meeting of the American Law Institute at Washington, DC, 24 May 1962 (Philadelphia, PA: American Law Institute, 1985), §3.04(1).

[35] See the citations stated in note 17 above.

made for the capacity of robots to incur criminal liability.[36] This is a highly contested proposal and a fascinating topic for criminal law theorists. Holding robots criminally liable would not be compatible with traditional features of criminal law: its focus on human agency and the notion of personal guilt, i.e., *Schuld*, which is particularly prominent in German criminal law doctrine. Many criminal law theorists defend these features as essential to the very idea of criminal law and thus reject the idea of permitting criminal proceedings against robots. But this is at best a weak argument. Criminal law doctrine is not set in stone; it has adapted to changes in the real world in the past and can be expected to do so again in the future.

The crucial question is whether there are additional principled objections to subjecting robots to criminal liability. Scholars typically examine the degree to which the abilities of robots are similar to those of humans[37] and ask whether robots fulfil the requirements of personhood, which is defined by means of concepts such as autonomy and free will.[38] These positions could be described as status-centered, anthropocentric, and essentialist. Traditional concepts of personhood rely on ontological claims about what humans are and the characteristics of humans *qua* humans. As possible alternatives, notions such as autonomy and personhood could also be described in a more constructivist manner, as the products of social attribution,[39] and it is worth considering whether the criminal liability of robots could at least be constructed for a limited subsection of criminal law, i.e., strict liability regulatory offenses, for legal systems that recognize such offenses.[40]

Instead of exploring the degree of a robot's human-ness or personhood, the alternative is to focus on the functions of criminal proceedings and punishments. In this context, the crucial question is whether some goals of criminal punishment practices could be achieved if norms of conduct

[36] See, for the argument that the categories of *actus reus* and *mens rea* could also be applied to robots, Gabriel Hallevy, *When Robots Kill* (Boston, MA: Northeastern University Press, 2013).

[37] Lawrence Solum, "Legal Personhood for Artificial Intelligences" (1992) 70:4 *North Carolina Law Review* 1231 ["Legal Personhood"] at 1255–1280.

[38] "Legal Personality for AI", note 17 above, at 416–417; see Chapter 15 in this volume.

[39] See Gunther Teubner, "Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law" (2006) 33:4 *Journal of Law & Society* 497; "Guilty Robots", note 2 above, at 13–21.

[40] See Mireille Hildebrandt, "Criminal Liability and 'Smart' Environments" in Antony Duff & Stuart Green (eds.), *Philosophical Foundations of Criminal Law* (New York, NY: Oxford University Press, 2011) 507 ["Criminal Liability"] at 525–526.

were explicitly addressed to robots and if defendants were not humans but robots. As we will see, it makes sense to distinguish between the preventive functions of criminal law, such as deterrence, and the expressive meaning of criminal punishment.

The purpose of deterring agents is probably not easily transferrable from humans to robots. Deterring someone presupposes that the receiver of the message is actually aware of a norm of conduct but is inclined not to comply with it, because other incentives seem more attractive or other personal motives and emotions guide his or her decision-making. AI will probably not be prone to the kind of multi-layered, sometimes blatantly irrational type of decision-making practiced by humans. For robots, the point is to identify the right course of conduct, not to avoid being sidetracked by greed and emotions. But preventive reasoning could, perhaps, be brought to bear on the humans involved in the creation of robots who might be indirectly influenced. They might be effectively driven toward higher standards of care in order to avoid public condemnation of their products' behavior.[41]

In addition to their potentially preventive effects, criminal law responses have expressive features. They communicate that certain kinds of wrongful conduct deserve blame, and more specifically they reassure crime victims that they were indeed wronged by the other party to the interaction, and not that they themselves made a mistake or simply suffered a stroke of bad luck.[42] Some of the communicative and expressive features of criminal punishment might retain their functions, and address the needs of victims, if robots were the addressees of penal censure.[43] Even if robots will not for a long time, if ever, be capable of feeling remorse as an emotional state, the practice of assigning blame could persist with some modifications.[44] It might suffice if robots had the cognitive capacity to understand what their environment labels as right and wrong and the reasons behind these judgments, and if they adapted their behavior to norms of conduct. Communication would be possible with smart robots that

[41] Ying Hu, "Robot Criminals" (2019) 52:2 *University of Michigan Journal of Law Reform* 487 ["Robot Criminals"] at 508–510.

[42] Tatjana Hörnle, "The Role of Victims' Rights in Punishment Theories" in Antje du Bois-Pedain & Anthony Bottoms (eds.), *Penal Censure: Engagements Within and Beyond Desert Theory* (London, UK: Hart, 2019) 207.

[43] "Guilty Robots", note 2 above, at 21–28.

[44] See "Robot Criminals", note 41 above, at 504–507; Karsten Gaede, *Künstliche Intelligenz – Rechte und Strafen für Roboter?* (Artificial Intelligences – Rights and Criminal Punishment for Robots?) (Baden-Baden, Germany: Nomos, 2019) [*Künstliche Intelligenz*] at 64.

are capable of explaining the choices they have made.[45] In their ability to respond and to modify parameters for future decision-making, advanced robots are distinguishable from others not held criminally liable, e.g., animals, young children, and persons with severe mental illness.

Admittedly, criminal justice responses to the wrongful behavior of robots cannot be the same as the responses to delinquent humans. It is difficult, e.g., to conceive of a "hard treatment" component of criminal punishment[46] that would apply to robots, and such a component, if conceived, might well be difficult to enforce.[47] It could, however, be argued that punishment in the traditional sense is not necessary. For an entirely rational being, the message that conduct X is wrongful and thus prohibited, and the integration of this message into its future decision-making, would be sufficient. The next question would be if blaming robots and eliciting responses could provide some comfort to human victims and thus fulfil their emotional needs. It is conceivable that a formal, solemn procedure might serve some of the functions that traditional criminal trials fulfil, at least in the theoretical model, but study would be required to determine whether empathy or at least the potential for empathy are prerequisites for calling a perpetrator to account. Criminal law theorists have argued that robots could only be held criminally liable if they were able to understand emotional states such as suffering.[48] In my view, a deeply shared understanding of what it means, emotionally, to be hurt is not necessarily essential for the communicative message delivered to victims who have been harmed by a robot.

Another question, however, is whether a merely communicative "criminal trial," without the hard treatment component of sanctions, would be so unlike criminal punishment practices as we know them that the general human public would consider it pointless and not worth the effort, or even a travesty. This question moves the inquiry beyond criminal law theory. Answers would require empirical insight into the feasibility and acceptance of formal, censuring communication with robots. If designing procedures with imperfect similarities to traditional criminal trials would make sense, the question of criminal codes for robots should perhaps also be addressed.[49]

[45] "Robot Criminals", note 41 above, at 499.
[46] For the distinction between blame and hard treatment, see Andrew von Hirsch, *Censure and Sanctions* (Oxford, UK: Clarendon, 1993) at 9–14.
[47] *Künstliche Intelligenz*, note 44 above, at 66–69.
[48] "Criminal Liability", note 40 above, at 530–531.
[49] "Robot Criminals", note 41 above, at 500–503.

III.F Robots as Victims of Crime

Another area that might require more attention in the future is the interpretation of criminal laws if the victim of the crime is not a human, as assumed by the legislators when they passed the law, but a robot. Crimes against personality rights, e.g., might lead to interesting questions. Might it be a criminal offense to record spoken words, a criminal offense under §201 of the *Strafgesetzbuch* (German Criminal Code), if the speaker is a robot rather than a human being? Thinking in this direction would require considering whether advanced robots should be afforded constitutional and other rights[50] and, should such a discussion gain seriousness, which rights these would be.

IV The Long-Term Perspective: General Effects on Substantive Criminal Law

The discussion in Section III above referred to criminal investigations undertaken after a specific human–robot interaction has caused or threatened to cause harm. From the perspective of criminal law theory, another possible development could be worth further observation. Over time, the assessment of human conduct, in general, might change, and perhaps we will begin to assess human–human interactions in a somewhat different light, once humanoid robots based on AI become part of our daily lives. At present, criminal laws and criminal justice systems are to different degrees quite tolerant with regard to the irrational features of human decision-making and human behavior. This is particularly true of German criminal law where, e.g., the fact that an offender has consumed drugs or alcohol can be a basis for considerable mitigation of punishment,[51] and offenders who are inclined to not consider possible negative outcomes of their highly risky behavior receive only a very lenient punishment or no punishment at all.[52] This tolerance of human imperfections might shrink if the more rational, de-emotionalized version of decision-making by AI has an effect on our expectations regarding careful behavior. At present, this is merely a hypothesis; it remains to be seen whether the willingness of criminal courts to accommodate human deficiencies really will decrease in the long term.

[50] For a discussion about the legal rights of robots, see "Legal Personhood", note 37 above.
[51] StGB, note 28 above, §21; *German Criminal Law*, note 23 above, at 135.
[52] The definition of conditional intent requires the defendant to be aware of the risk and to accept it: see *German Criminal Law*, note 23 above, at 63–67; "Criminal Law", note 23 above, at 509.