44

Ensuring Accountability for Robots and AI under Criminal Law

Monika Simmler

Autonomous robots and Artificial Intelligence (AI) are increasingly involved in the commission of criminal offenses, resulting in questions as to who is accountable for such crimes and how human–machine interactions influence criminal responsibility. This chapter first elaborates upon the conditions of machine responsibility and explains why technical systems cannot be granted personhood under today's criminal law. It then discusses the challenges that complex human–machine interactions bring in terms of the attribution of criminal responsibility, before outlining options regarding how they might be met. Human–machine interactions should be understood as a form of distributed agency. It is therefore essential to define what a legitimate delegation of agency to a machine is. As this chapter shows, this could lead to a modified understanding of the attribution of action under criminal law, or even new norms specifically designed to address automation. Ensuring accountability under criminal law for robots and AI means holding actors accountable who cede agency to such technologies in bad or unjustified ways.

44.1 INTRODUCTION

Potential legal personhood for machines in general and robots in particular has been a subject of academic interest for some time.[1] Even in times when robots were primarily known in fiction and technologically as yet not advanced, the idea that Artificial Intelligence (AI) could one day have rights and duties was intriguing.[2] In criminal law, however, this idea has been widely regarded as unrealistic.[3] The concept of the subject and especially the one of culpability in this legal field is simply too demanding.[4] It is people who act, and thus it is people to whom

[1] See among many others Hallevy, G., *When Robots Kill: Artificial Intelligence under Criminal Law*. Boston: Northeastern University Press, 2013; Pagallo, U., *The Laws of Robots: Crimes, Contracts, and Torts*. Law, Governance and Technology Series, Vol. 10. Wiesbaden: Springer, 2013; Beck, S., Intelligent agents and Criminal Law – Negligence, Diffusion of Liability and Electronic Personhood. *Robotics and Autonomous Systems* 86, 138–143, 2016 (hereafter Beck, Intelligent agents); Balkin, M. J., The Path of Robotics Law. *California Law Review Circuit* 6, 45–60, 2015; Calo R., Robotics and the Lessons of Cyberlaw. *California Law Review* 103(3), 513–563, 2015.

[2] See, for example, Solum, B. L., Legal Personhood for Artificial Intelligences. *North Carolina Law Review* 70(4), 1239, 1992.

[3] See, for example, Gless, S., Silverman, E., and Weigend, T., If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability. *New Criminal Law Review* 19(3), 412, 2016 (hereafter Gless, Silverman, and Weigend, If Robots Cause Harm).

[4] Ibid.

blame is ascribed. How can an artificial entity ever truly be personally blamed to the extent that punishment is justified? Criminal law has traditionally been based on a strictly dualistic perspective: People act; machines function.[5] However, the emergence of increasingly capacitated machines, socially interacting robots, and above all, technical systems based on AI challenge this foundational dualism.

It quickly becomes evident when researching the legal implications of advanced technical systems that there exists not one kind of "machine" or "robot," but rather a wide variety of technical artifacts with very different functions and capabilities. Thus, when this article discusses different technical systems, be it robots or others, it does so aware that in each individual case the peculiarities of the concrete system would have to be considered and that there are hardly any uniform definitions for these terms. However, what distinguishes modern technology that is of interest here from earlier systems is the significant increase in technical autonomy. Such autonomous (or partly autonomous) technology is not completely determined, but can, for example, react to changes in the environment, interact with other systems, expand its database, follow preferences, and choose between alternative courses of action.[6] Autonomy now is traditionally regarded as reserved for humans, and serves as a justification for holding them accountable for their actions. Doctrines of criminal responsibility still rely to a substantial extent on an understanding of human beings as autonomous creatures driven by free will.[7] Someone is at fault only because they could have acted otherwise. More and more autonomous robots now have the potential to challenge that conception as they increasingly take over agency and expand their power to act. This renders them particularly interesting for the doctrine of imputation under criminal law, raising questions about new artificial agents and novel forms of distributed agency in human–machine interaction. The difficulty is further accentuated when high levels of automation are added.

An article with a colleague in 2018[8] discussing the possibility of "guilty robots" began with the examples of Nao and Pepper, two humanoid robots capable of basic social interaction. We explored the question of whether the rise of such robots could result in them being acknowledged as subjects of criminal law and concluded that it would not. Today, just five years later, these examples of cute robots used for education and entertainment seem somewhat absurd, given the rapid pace of technological progress. Recently, the chatbot "ChatGPT"[9] explained to me that "criminal law requires that the accused party has the ability to understand the nature

[5] From a general sociological standpoint, Rammert, W., Distributed Agency and Advanced Technology: Or How to Analyze Constellations of Collective Inter-Agency. In J. Passoth, B. Peuker and M. Schillmeier (eds.), *Agency without Actors? New Approaches to Collective Action*. London: Routledge, 89, 2012, hereafter Rammert, Distributed Agency; see also Simmler, M., Automation. In P. Caeiro, S. Gless, V. Mitsilegas, M. J. Costa, J. De Snaijer and G. Theodorakakou (eds.), *Elgar Encyclopedia of Crime and Criminal Justice*. Cheltenham: Elgar Publishing, 2023, hereafter Simmler, Automation; on this dichotomy, furthermore see Hilgendorf, E., Können Roboter schuldhaft handeln? In S. Beck (ed.), *Jenseits von Mensch und Maschine – ethische und rechtliche Fragen zum Umgang mit Robotern, Künstliche Intelligenz und Cyborgs*. Baden-Baden: Nomos, 120, 2012, hereafter Hilgendorf, Roboter.

[6] Simmler, M. and Frischknecht, R., A Taxonomy of Human-Machine Collaboration: Capturing Automation and Technical Autonomy. *AI & Society* 36, 239, 2020 (hereafter Simmler and Frischknecht, Taxonomy); Müller-Hengstenberg, C. D. and Kirn, S., *Rechtliche Risiken autonomer und vernetzter Systeme*. Berlin: De Gruyter Oldenbourg, 5, 2016; Simmler, Automation (n 4).

[7] On the foundations of placing blame in criminal law, see (among many others) Simmler, M., The Importance of Placing Blame: Criminal Law and the Stabilization of Norms. *Criminal Law Forum* 31(2), 147–178, 2020 (hereafter Simmler, The Importance of Placing Blame).

[8] Simmler, M. and Markwalder, N., Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Criminal Law Forum* 30, 1–31, 2019 (hereafter Simmler and Markwalder, Guilty Robots).

[9] The question for 'ChatGPT' was: "Can social robots be called to account under criminal law?" https://chat.openai.com

and consequences of their actions." It continued that this "would seem to exclude most current social robots" but that "as the technology advances, it is possible that social robots may become more advanced and autonomous, which could potentially lead to them being held accountable under criminal law." Then everything has been said, right? It is impressive how quickly disruptive new technologies spread. Robots benefit from the power of AI, whose various learning methods are becoming ever-more advanced. At the same time, the possibility of embodying such intelligent technology in increasingly mechanically sophisticated machines is expanding, supporting the tendency to recognize them as fully fledged interaction partners. The examples of Nao and Pepper now seem from another age, but the idea of a guilty robot against whom the state might impose punishment remains difficult to imagine. However, given the disruptive nature of AI, we must come to terms with the idea that these issues may no longer be so clearly delineated. The chatbot might therefore not be wholly incorrect.

In light of recent developments, this chapter first explores the ways in which robots can be involved in criminal offenses (Section 44.2). The conditions under which robots could be granted personhood under criminal law are then analyzed, establishing that such an attribution is not per se impossible, but unlikely in today's climate (Section 44.3). Who is to be called to account for damage caused by robots is then considered (Section 44.4). Regarding the challenges associated with such attribution, it is next argued that the traditional dualism between human and machine may no longer be appropriate, and incapable of fully capturing the variety and nature of modern human–robot interactions (Section 44.5). Understanding such interaction as a case of (gradually) distributed agency does more justice to today's social reality. The existing doctrine of criminal attribution to some degree does meet this reality, but reaches its limits in the case of highly autonomous technology or very complex automation. We must therefore turn to new concepts to ensure accountability in human–machine interactions. This chapter elucidates such approaches by discussing four possible regulatory strategies (Section 44.6), before concluding with final remarks (Section 44.7).

44.2 CRIMES AGAINST ROBOTS, ROBOT CRIMES, AND CRIMES BY ROBOTS

It is clear that robots and technical systems can be involved in criminal conduct. *First*, robots can be the target of crimes. Among other things, they can be intentionally destroyed or hacked, resulting in responsibility for property crimes. For now, they are treated as "things." However, the possibility that in the future robots could have rights has been discussed in scholarship.[10]

Second, robots can be used intentionally as the means of committing a crime. The imputation of intentional wrongdoing in this case is mostly unproblematic, since the accused commits the offense both knowingly and willingly. The robot is a mere tool. If one day advanced technical systems would be regarded as agents of their own, we might have to deal with cases of offenses using an innocent agent. Even then, existing doctrines would allow for ascribing the robot's behavior to the intentional actor, although the human is not using their own hands.[11]

Third, the use of robots can result in damage that was unintended but potentially negligently caused. This variant is already relevant today and will certainly become more so in the future. The allegation of negligent conduct in most legal orders usually is that someone has disregarded their duty of care and caused damage that was predictable and preventable. Developers,

[10] See, for example, with empirical findings regarding public opinion on the matter De Graaf, M. M. A., Hindriks, F. A., and Hindriks, K. V., Who Wants to Grant Robots Rights? *Frontiers in Robotics and AI* 8, 1–13, 2022.

[11] Simmler and Markwalder, Guilty Robots (n 7), 7; Gless, Silverman and Weigend, If Robots Cause Harm (n 3), 412.

manufacturers, distributors, implementers, and users of technical systems can thus be held liable under the negligence regime if such carelessness can be proven. Modern technology challenges this kind of attribution of responsibility. This is due to two factors. One, technical autonomy complicates the determination of the relevant actor. A single person rarely creates a robot on their own. Rather, a whole range of actors are involved before the complex technical system comes into use. In such cases, it may be difficult to determine who made the mistake. It may be the result of a specific interplay of actors that leads to a kind of "cumulative carelessness" resulting in harm. Criminal law is not designed to address such "collective fallibility."[12] This phenomenon has been discussed in the literature as the "problem of many hands"[13] and additionally – adapted to the current challenge – the[14] "problem of many things." Such issues are problematic because criminal law is individualistic in the sense that a particular accusation must be made to call someone to account for acting negligently. In many legal orders, corporate responsibility presents an exception to this rule, but still requires a very specific allegation, as well as legal personhood built upon the corporation's distinct capacity to act. Two, the negligence regime itself is challenged by technical autonomy. As stated earlier, the allegation of negligence usually requires that the harm was predictable and preventable (although of course there can be differences between jurisdictions and legal cultures). The more autonomous the technology becomes, the more difficult it will be to establish this element.[15] Scholars have therefore discussed whether and to what extent responsibility gaps arise in such cases.[16]

Fourth, robots equipped with AI could commit crimes for which no human actor (and no corporation) can be held responsible, but which still appear to be intentional. An original "machine responsibility" could be the answer to this attribution problem, perhaps meeting a (retributive) social need. Notions such as "action," "guilt," and "responsibility" do not have an invariable meaning in criminal law and could in principle be employed for technical systems.[17] However, social attribution of personhood and other capacities is not arbitrary.[18] Robots would need to be recognized as capable addressees of norms and accusations, as well as of punishment.[19] As said earlier, such a new reality is also somewhat absurd and difficult to imagine from today's perspective. Above all, it is essential to note that such a "machine responsibility" cannot simply be introduced because technical systems may then serve as an easy scapegoat allowing to solve attribution issues under the current regime. Placing blame on a robot can only be functional (and thus make any sense) if it follows the criminal law's clearly defined mechanisms. Blame is not attributed randomly. If it is, it does not serve criminal law's purpose.

[12] Simmler, Automation (n 4).

[13] Santoni de Sio, F. and Mecacci, G., Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology* 34, 1063, 2021 (hereafter Santoni de Sio and Mecacci, Four Responsibility Gaps).

[14] Coeckelbergh, M., Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics* 26, 2052, 2020.

[15] Beck, Intelligent agents (n 1), 139; Simmler, Automation (n 4).

[16] For a discussion of the risk of responsibility gaps in general (but not under criminal law exclusively), see Matthias, A., The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6, 175–183, 2004; Chinen, M. A., The Co-Evolution of Autonomous Machines and Legal Responsibility. *Virginia Journal of Law & Technology* 20(2), 338–393, 2016; Santoni de Sio and Mecacci, Four Responsibility Gaps (n 12), 1057–1084.

[17] Hilgendorf, Roboter (n 4), 122, 125.

[18] Simmler and Markwalder, Guilty Robots (n 7), 19.

[19] On calling robots to account in the light of theories of punishment, see Simmler and Markwalder, Guilty robots (n 7), 21.

44.3 CONDITIONS OF MACHINE RESPONSIBILITY

In criminal law, blame is placed only if a capable addressee of norms is present (personhood), this person was at the time of offending competent enough to question a norm (competence), and the person communicated a negation of this norm (message).[20] If these requirements are met, an act destabilizes a criminal norm and punishment becomes necessary. If someone is punished without blame being placed (as in cases of strict liability[21]), punishment does not stabilize a norm because no one has questioned its ongoing significance. It is not answering a person's fault, but rather bad luck or fate, rendering punishment dysfunctional. If technical artifacts are held accountable only because doing so serves criminal policy considerations but without making an adequate personal accusation against them, there is no social benefit over time.

Criminal law protects specific normative expectations that are important for social life. Normative expectations are characterized by the fact that we continue to adhere to them in the event of their disappointment. A reaction is therefore required to reinforce them. Cognitive expectations are different, as we usually relearn them in the event of disappointment.[22] The crucial question is therefore the kinds of normative expectations we direct toward robots. Personhood is constructed and constituted through social interplay. Persons cannot emerge and persist without society (and vice versa).[23] If we assume that autonomous robots are determined objects and the products of their human creators, we cannot impart normative expectations on them. If they behave differently from what is expected, we either adjust our behavior, or if we feel their creators made a mistake, address the criminal subjects "behind" the robot. This mechanism is clearly what operates today.[24] Admittedly, research has suggested that we have the need to blame computers when they do not perform as planned.[25] However, such spontaneous reactions and anthropomorphizing cannot be equated with the need for state punishment in the legal sense. For us to grant robots personhood and ascribe to them the competence to challenge norms, we would have to recognize them as full members of social interaction. The characteristics we interpret today as technical autonomy, such as nontransparency, indetermination, adaptability, and openness to external input,[26] are not sufficient for this. Rather, such an autonomous system would have to be able to reflect upon its own actions, truly interact in a social manner, and independently set higher-order goals. The last criterion might be the most important. The possibility of giving oneself laws ("self-legislation") in a Kantian sense is regarded as a prerequisite for true autonomy.[27] Although this demand seems somewhat idealistic, it may be exactly

[20] Simmler, The Importance of Placing Blame (n 6), 168.

[21] See Simmler, M., Strict Liability and the Purpose of Punishment. *New Criminal Law Review* 23(4), 516–564, 2020.

[22] Luhmann, N., *Social Systems*. Stanford, CA: Stanford University Press, 1995 (hereafter Luhmann, Social Systems); Günther, J., *Strafrecht, Allgemeiner Teil. Die Grundlagen und die Zurechnungslehre*, 2nd ed. Berlin: De Gruyter, 6–8, 1993; Simmler, The Importance of Placing Blame (n 6), 154.

[23] Luhmann, Social Systems (n 21), 59.

[24] Consider the empirical findings in Simmler, Automation (n 4) or Strafrechtliche Verantwortung im Zeitalter autonomer Technik: Vom Individual- zum Unternehmensstrafrecht? In Fink et al. (eds.), Strafjustiz zwischen künstlicher Intelligenz und prädiktiven Algorithmen, Basel: Helbing Lichtenhahn, 140 (hereafter Simmler, Strafrechtliche Verantwortung), further discussed in Section 44.6.

[25] Kahn Jr., P. H., et al., Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes? In: *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012, 33; Kirchkamp, O. and Strobel, C., Sharing Responsibility with a Machine. *Journal of Behavioral and Experimental Economics* 80, 27, 2018; Friedmann, B., "It's the computer's fault" – Reasoning about Computers as Moral Agents. *Conference Companion on Human Factors in Computing Systems (CHI 95)*, 1995, 226.

[26] Simmler and Frischknecht, Taxonomy (n 5), 245.

[27] On this, for example, see Kain, P., Self-Legislation in Kant's Moral Philosophy. *Archiv für Geschichte der Philosophie* 86(3), 258, 2004.

what we do not credit robots with now. Artificial Intelligence may follow intentions, and even adapt intentions to follow goals of a higher order, varying its behavior to do so. Nevertheless, its final purpose and goal are externally given by inputs from human creators. As long as this holds true, criminal legal norms will address the people responsible for such input, and not the robot.

Consequently, the Electronic Personhood ("E-Personhood") discussed in the literature[28] is most likely primarily an answer to tort law liability needs. This construct can ensure compensation for financial damage brought about by robot use if traditional liability regimes are no longer suitable. Criminal law's function is different. A mere artificially constructed E-Person that does not find support in the social views on robots would not allow for the functional placement of blame. Of course, an E-Person could be understood as a representation of the human and corporate subjects "behind" the robot (i.e., as a legal person that emerges due to setting a robot in action). In this case, an E-Person would be a variant of a legal person, following the assumption that putting a robot into operation is preceded by a collective action. This would not imply responsibility of the robot itself, but rather the entity behind the machine. No one would consider the robot guilty. This idea is further explained below, as it presents one approach to future regulation.

Our 2018 conclusion that robots cannot be guilty in a criminal sense is reaffirmed today. As was stated in that earlier work, this is something that could change.[29] Conceptions of personhood, culpability, and responsibility are not ontically or metaphysically predetermined, but evolve with society. If the capabilities of technical systems change even more and we direct normative expectations toward them as equal actors in social interaction, they could well be declared responsible, and consequently, punished. This would happen automatically. The mechanism of placing blame remains (functionally) the same. Its substance, however, depends on society.

44.4 RESPONSIBILITY FOR DAMAGES CAUSED BY ROBOTS

When robots are not responsible for the damage they cause, one naturally asks who is. The question of what other actors could, under which conditions, be called to account under criminal law is therefore the subject of an ongoing debate.[30] In most legal orders it is natural and legal persons, while in others it is just humans. Both intentional and negligent offenses are perceivable. As mentioned earlier, intentional wrongdoing does not cause substantial problems. In cases of unintended harm, however, it becomes more difficult to define who and if someone is in charge. It has already been pointed out how problems of complex and difficult-to-trace behavior in corporations and development cycles complicate imputation under criminal law, as do higher levels of autonomy. In general, the following actors are likely to be considered responsible or jointly responsible for damages caused by technical systems: (i) the human beings programming and developing the software, (ii) the software company, (iii) the human beings developing and manufacturing the hardware, (iv) the manufacturer, (v) the human beings working to distribute the technical system (i.e., sellers and people giving

[28] Beck, Intelligent agents (n 1), 138–143; Avila Negri, S. M. C., Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence. *Frontiers in Robotics and AI* 8, 1–10, 2021.

[29] Simmler and Markwalder, Guilty Robots (n 7), 29.

[30] See, among others, Hallevy, G., The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control. *Akron Intellectual Property Journal* 4(2), 171–201, 2010; Beck, Intelligent agents (n 1), 138–143; Beck, S., Google Cars, Software Agents, Autonomous Weapons Systems – New Challenges for Criminal Law? In E. Hilgendorf and U. Seidel (eds.), *Robotics, Autonomics, and the Law*. Baden-Baden: Nomos, 227–252, 2017; Gless, Silverman, and Weigend, If Robots Cause Harm (n 3), 412–436.

instructions to future users), (vi) the distributor, (vii) the human beings deciding on the deployment of a technical system within an institution or company, (viii) the implementer, and (ix) the human beings using the technical system. Of course, not all of these actors are present at all times. Likewise, it could be that additional players are sometimes involved. What remains clear is that the situation is complicated.

Human–machine interaction thus makes attribution processes more complex. In today's doctrine, when damage occurs (such as property damage or bodily injury caused by a robot), a specific accusation must be present that a human or company was at fault. If a person has violated due diligence standards in a manner obviously suited to cause such damage, this often does not pose a problem. If a robot is sold as a child's toy that has been programmed without considering that children can react unpredictably, this is considered acting carelessly. Anyone who does not instruct the purchasers of such a robot or who obviously does so insufficiently is also acting carelessly. Such behavior is recognized as socially inadequate. High levels of automation or autonomy are not required to enable accusation. Rather, care must be adjusted in view of the corresponding levels. If it is obvious that a person has poorly managed the automation and therefore caused preventable and predictable harm, the matter is clear. In most cases, however, the facts are more complicated. Of course, these difficulties primarily concern the provision of evidence. It must first be proven that someone has made a mistake. Yet, even if it is clear what led to the damage, difficulties can arise. It must be established that the error was actually the cause of the accident. In the legal sense, causality means that it is *conditio sine qua non* and the sequence of causal events corresponds to the usual course of events. Unusual causal courses release the actor from responsibility. Autonomous robots could present a "superseding" or "intervening" cause because they can significantly shape the course of action, due to their independence.[31] Establishing causality can therefore be problematic. In addition, it must be proven that the mistake that caused the harm is no longer within the range of so-called admissible risk.[32] When dealing with new technologies, it is rarely possible to say with certainty that no damage can result. Nevertheless, society is generally prepared to accept certain risks when it comes to innovation. What exactly the admissible risk is with AI is not yet clear and will have to be defined. What is certain, however, is that not every risk-taking justifies criminal liability if harm results.

These considerations already show that there are obstacles when trying to capture human–machine interaction in traditional doctrine. It is therefore appropriate to discuss how exactly such constellations are to be appraised. Even corporate criminal liability is not designed to grasp assemblages of highly complex (socio-technical) interplay. Companies are recognized as the subjects of criminal law because they bundle normative expectations. In society's perception, they act and can comply with norms, but can also challenge them. If within a software or hardware company no due diligence standards are taken into account and compliance requirements simply ignored, it will be possible – depending on the legal system – to hold that corporation accountable. However, such extreme cases are unlikely. More often, we see cases of cumulative carelessness. Various minor errors lead to a major error, which then causes damage. In such situations, it is difficult to hold the company accountable, nor does it make much

[31] See, from a civil law perspective, Scherer, M. U., Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology* 29(2), 354–400, 2016; Kowert, W., The Foreseeability of Human-Artificial Intelligence Interactions. *Texas Law Review* 96(1), 181–204, 2017; in the relevance of intervening causation in criminal law generally Moore, M. S., *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press, 229–253, 2009.

[32] Beck, Intelligent agents (n 1), 141.

sense to do so. Often, it will not be a single company responsible for a technical system. The increasing complexity of technology development has led to constellations of multiple corporations interacting in various ways. This is where things become especially complicated. The concern about responsibility gaps[33] becomes more justified because it is no longer comprehensible who exactly is responsible for what. This can lead to lawlessness, or at least the feeling of it. It is therefore important to address such complex interactions in socio-technical systems to secure accountability under (criminal) law. In many jurisdictions, very similar considerations were made when corporate criminal liability was first introduced.[34] This, too, was the response to a need to legally grasp complex corporate structures (i.e., larger and more global entities). At that time, the main concern was preventing international organized crime, corruption, and other white-collar offenses.[35] It was often impossible to determine who exactly did what wrong in a given corporation. Corporations should not be able to serve as a cover, concealing criminal activities. Socio-technical systems bring a new form of complex interplay suitable to diffuse accountability. Therefore, we may have to yet again reconsider who to address through criminal law. This entails more than discussing the question of robot responsibility alone.

44.5 LEAVING DUALISTIC BELIEFS BEHIND

Our current social system denies the criminal liability of technical artifacts, even if they are quite advanced. However, it would be too simplistic to conclude that the attribution of responsibility remains the same, given the advanced nature of our technical companions. We must be open to the idea that machines do not act in a criminal sense, but they do take on partial agency. Artificial Intelligence is explicitly being defined as "a growing resource of interactive, autonomous, self-learning agency, which enables computational artifacts to perform tasks that otherwise would require human intelligence to be executed successfully."[36] Here, being able to assume agency is what is considered constitutive of AI. Such partial agency is significant because every criminal responsibility relies on the attribution of agency. As technical autonomy increases, partial agency becomes more plausible and understanding human–machine interaction as distributed agency more adequate.[37] When ignoring the changed nature of intelligent robots and clinging rigidly to a human–machine dichotomy, one misses the chance to truly understand and grasp the more complex and graded constellations of interaction. The position of a categorical and irrevocable dichotomy of humans and machines significantly limits the ability to respond to new challenges related to autonomous robots.[38] The question of robot responsibility must therefore be transformed into a question of responsibility within human–robot interactions. In this context, distributed agency is not only to be understood as interaction between the user and technical system. When advanced

[33] See the remarks in n 15.

[34] On the establishment of corporate criminal responsibility, among many others see Pieth, M. and Ivory, R. (eds.), *Corporate Criminal Liability: Emergence, Convergence and Risk*. Wiesbaden: Springer, 2011 (hereafter Pieth and Ivory, Corporate Criminal Liability); De Maglie, C., Models of Corporate Criminal Liability in Comparative Law. *Washington University Global Studies Law Review* 4(3), 547–565, 2005; Quaid, J. A., The Assessment of Corporate Criminal Liability on the Basis of Corporate Identity: An Analysis. *McGill Law Journal* 43(4), 67–114, 1998.

[35] See, for example, Stewart, J. G., The Turn to Corporate Criminal Liability for International Crimes: Transcending the Alien Tort Statute. *New York University Journal of International Law and Politics* 47, 121–206, 2014; Pieth and Ivory, Corporate Criminal Liability, 3–60.

[36] Taddeo, M. and Floridi, L., How AI Can be a Force for Good. *Science* 361(6404), 751, 2018 (hereafter Taddeo and Floridi, AI).

[37] From a sociological standpoint, see Rammert, Distributed Agency (n 4), 89–112.

[38] Hilgendorf, Roboter (n 4), 120.

technology is in use, agency is distributed across many different parties, including developers, software companies, hardware companies, distributors, implementers, and users.[39] The key question, then, is as follows. How does the interaction of humans and technical systems influence the distribution of agency and, as a consequence, the attribution of responsibility under criminal law?

If a retirement home uses a care robot that memorizes and considers the residents' preferences on a self-learning basis and an accident occurs, its capabilities cannot simply be ignored. They are relevant to the context of action and its legal assessment. The first step would be to analyze whether the human caregiver who supervised the use of the technology had any opportunity to prevent the harm. For this purpose, the automation level must be determined (i.e., the distribution of tasks between human and machine). If the operator could and should have intervened, then he or she remains the decisive actor. If the operator could no longer intervene, the level of technical autonomy becomes important. Of course, this level is relevant to establishing the operator's responsibility anyway, but it is crucial once the operator could not determine the machine's behavior anymore. The question then is what influence on this (autonomous) behavior other actors had, such as the retirement home as implementer or manufacturer as programmer. If the levels of autonomy and/or automation are low, the human actors are still in the loop, either by acting situationally or by fully determining the technical system. However, if the system itself has evolved, this fact cannot be overlooked.

When the situational responsibility for action is ceded due to a high level of automation and/or autonomy, the question of "prior fault" emerges. Behaviors preceding the use of technology become the main issue. That is nothing unusual under criminal law. Very often, the criminal act does not immediately precede the damage. The causal chain can frequently extend over longer periods or across intermediate acts. In the context of negligence offenses, this is primarily the case. Violation of the duty of care can be considered an act according to criminal law, even if other causal factors are added and considered relevant with regards to damage. This is distinguishable from the genuine prior responsibility we know from legal concepts as the so-called *actio libera in causa*. Here, the offense itself did not take place consciously and controllably (usually due to intoxication), but this state of affairs is due to one's own (prior) fault.

Regarding the care robot in the above example, it is usually possible to reproach natural and legal persons, provided that their behavior constitutes in advance a violation of due diligence standards and determines the causal course in a legally significant way. Thereby, it is likely that high levels of autonomy and automation require a high degree of care and are regarded as an indicator (but not proof) that the ceding of authority and control (i.e., of [partial] agency) to machines in and of itself should not have taken place, at least in this way. Of course, other factors are also decisive in establishing due diligence standards. For example, the robot's social skills and certainly its specific area of application could be crucial. Still, for attributing responsibility, the primary focus must be on the distribution of agency. This requires an analysis of the human–machine interaction, rendering automation and autonomy levels pivotal because they structure the distribution of (situational and prior) control. However, if agency was ceded to autonomous robots and their unforeseeable behavior leads to unforeseeable damage, the cession of agency must be considered. Noone was at fault. Otherwise, delegating actions to machines would in itself always give rise to criminal responsibility.

In summary, by making such interaction the center of attention, collaboration between humans and technology is considered an actual inter*action*. Humans are ceding agency

[39] Taddeo and Floridi, AI (n 35), 751.

situationally (through automation) and generally (through technical autonomy). Holding them accountable is therefore complicated.

44.6 STRATEGIES FOR FUTURE REGULATION: NEW ACTORS, ACTS, OR NORMS?

Thus far, we have seen that in an age of increasingly autonomous technical systems, we are dealing with machines that are not themselves responsible, but nonetheless are gradually assuming agency. In cases of human–machine interaction resulting in distributed agency, this means that (at least) one human being decided to give up agency and let it be assumed by technology. This cession of power can lead to potential irresponsibility on the side of the delegating human, who has given up control and therefore is no longer considered accountable. Thus, identifying responsibility gaps as a risk is justified.[40] Such situations could result in agency (and accountability) effectively being ceded to an artifact that is not subject to law. Such machines can no longer be (fully) controlled, nor their actions predicted (due to their autonomy). If, for example, a manufacturer was in such a case nevertheless to be held responsible for negligently causing damage, the doctrine of negligence would become quite overstretched. In fact, all harmful behavior would then be predictable from an *ex post* point of view, which in turn would result in strict liability. This is not reasonable. If every cession of agency to autonomous machines served as grounds for liability, such ceding is *de facto* forbidden.

Meanwhile, the results of an empirical study[41] indicated that the use of technology at high levels of automation, and especially of intelligent, autonomous technology, seemed to have a significant impact on criminal responsibility attribution practices, at least according to social observers. In the case of mere automation, the attribution of responsibility shifted initially from the human acting either independently or as a user of the technology, to new actors such as human programmers and, in particular, software developers and companies responsible for the deployment of the technical system. Corporate liability gained significant importance once the technology became part of the equation. No responsibility gap was visible, but rather there was a clear shift in responsibility to new actors and a substantially higher complexity in terms of attribution practices. Mostly, not one but rather several actors were perceived as responsible for any damage that occurred. When technical autonomy was added, the attribution became somewhat arbitrary and relatively incomprehensible. The phenomenon seemed to overstrain laypeople. They did not refrain from placing blame but appeared to ascribe it randomly. This suggests that a responsibility gap might not necessarily emerge, but rather a confusion or diffusion of responsibility. At least, that seems to be the case when legal laypersons were consulted. Such mechanisms could lead to insecurity and destabilization, and not be any less dysfunctional than the overstretching of established doctrines. Hence, for human–machine interactions that involve distributed agency, a functional approach to attribution must be found. There are different options for doing so.

44.6.1 *Accepting Responsibility Gaps*

As a first and most obvious option, responsibility attribution gaps could simply be accepted. If there is no actor who can be called to account, the incident could be labeled a technical accident, and there may be no need for punishment at all. Thus, there are authors who have said

[40] See the references in n 15.

[41] For a summary of the study, see Simmler, Automation (n 4) or Simmler, Strafrechtliche Verantwortung (n 23), 140.

that no responsibility gaps will emerge.[42] This scenario is the most likely when there is truly noone to blame. In such situations, no norms are destabilized. This path is plausible as a first step. Criminal law is *ultima ratio*. Many problems can adequately be solved with administrative law in the sense of regulating market approval and licensing, or with tort law providing compensation for damages. As long as this is sufficient to maintain minimum standards, there is no need to ensure accountability by means of criminal law. The situation is different, however, when normative expectations are disappointed and society interprets this as a questioning of criminal norms, but at the same time does not know to whom the reproach should be directed. Then there is indeed a gap and random attribution must be prevented. Which of these two alternatives will materialize depends very much on the understanding of human–machine interactions we develop in society and our willingness to take risks in this regard. It depends in particular on what attitudes we have regarding the ceding of agency to technology. We cannot place blame on everyone using or bringing to use autonomous technology if it results in harm. As said earlier, the use of AI would then generally be forbidden, a condition that is neither appropriate nor realistic. Regulatory regimes that are too strict may deter innovation.[43] This is precisely why sound principles of accountability are necessary. If we cannot make allegations based on today's doctrine, we should not pursue doing so. Gaps, then, are more acceptable.

44.6.2 *Robots and Socio-Technical Systems as E-Persons*

On the other side of the spectrum is the introduction of robots as E-Persons. As stated earlier, this can be thought of in two ways. One could call to account the autonomous technical systems themselves. Conversely, the E-Person could be construed as a representation of the socio-technical system (i.e., the human–machine interaction as a whole). The first way is not reasonable from today's perspective. It would be inappropriate to punish technical systems that do not meet the requirements of a criminal subject. If one day robots are truly integrated into our social interaction to the extent that we place normative expectations on them, ascribe to them the competence to act otherwise, and independently and effectively question the validity of norms, this might be imaginable. However, we are far from that today. For now, the discussion about guilty robots remains interesting, but purely academic. Establishing an E-Person as a conglomerate of different actors in the sense of a socio-technical system would also require that we grant the system the independent capacity to act under criminal law. Just as with corporations, we would have to reach a point where we experience socio-technical systems as autonomous addressees of normative expectations. For this to happen, such systems would have to be firmly established in society as actors on their own. In the case of corporations, this is affirmed (though not by everyone) because they can develop a collective agency due to their specific organizational structure. However, the interaction of actors who shape a human–robot interaction is usually much less structured and not characterized by similar hierarchies and a self-contained mode of operation. The interaction of developers, manufacturers, and users is likely to be an example of distributed rather than collective agency. This distributedness does not lead to a perception of the socio-technical system as a unified subject that has newly emerged and can be addressed by punishment as an actual person would. Thus, while such an E-Person could readily provide a liability substrate in the civil law sense, it is unlikely that a penalty collectively directed at

[42] Tigard, D. W., There Is No Techno-Responsibility Gap. *Philosophy & Technology* 34, 589–607, 2021.
[43] Abbott, R., *The Reasonable Robot*. Cambridge University Press, 3, 2020.

such actors could be functionally enacted. Thus, for the time being, such an approach is not a promising solution for imputation but should still be further considered, as such a scheme would at least clear the view for the peculiarities of human–machine interaction as a possibly new form of collaborative action.

44.6.3 *Establishing New Norms and Offenses*

A third option is the creation of new criminal law norms that explicitly address the question of what is and is not allowed when using autonomous technology or extensive automation. This solution would not focus on the doctrine of imputation (i.e., on how and to whom we direct an allegation). Rather, it would address the question of which new and specific rules should apply. It would most likely involve so-called endangerment offenses.[44] These have been the trend in modern legal policy and often are used when one cannot or does not want to make an accusation in relation to harm but instead wants to consider dangerous behavior in itself as a reason for criminal liability. This can also be understood as a variant of prior responsibility. To return to our example, punishment would not be imposed because a care robot has injured a person. However, one would formulate an accusation that the retirement home or caregiver used autonomous technology in a dangerous way, regardless of the damage. As already mentioned, the assumption that simply every use of such technology is dangerous and justifies liability should be avoided. Risks may be taken. New norms would therefore have to define what is a dangerous and careless use of technology. Different from a "traditional" negligence offense, the standards would no longer relate to the occurrence of damage, which must not be predictable. Instead, the very act of delegating agency to technology in violation of certain duties would be punishable. Such an offense would present a crime of conduct and not of result. Delegation of agency would be blameworthy because of the nature of the delegation itself, regardless of whether and what exact harm results.

New norms could address careless or unjustified automation. New legislation would be introduced, as is often the case with such challenges. Such a new norm could read: "You should not cede agency to (autonomous) robots in a careless way!" However, corresponding standards have to be developed and are not yet socially defined. A second norm would refer not only to carelessness, but also to the question of whether one is allowed to cede agency to technology at all when within certain constellations. If one actually cedes agency, they can no longer be held accountable because they have dismissed themselves as an actor. Such a norm might read: "You should not cede control over action to technology if there are no good reasons for it!" In other words, technology should not be used abusively or unjustifiably. Everyone who gives up control must consider whether it is appropriate. This is the case, for example, if autonomous vehicles demonstrably lead to fewer accidents. This is not the case if a lower level of automation might actually have been more reasonable and the human being would still have had responsibility but did not want to bear it. Individuals should not be able to unjustifiably escape accountability. In brief, the third option is to introduce a specific "automation offense," establishing standards for good and justified automation.[45]

[44] Duff, A. and Hörnle, T., Crimes of Endangerment (Chapter 5). In K. Ambos, A. Duff, A. Heinze, J. Roberts and T. Weigend (eds.), *Core Concepts in Criminal Law and Criminal Justice*, Vol. 2. Cambridge University Press, 132, 2022; Kaiafa-Gbandi, M., Sachoulidou, A., and Chatzinikolaou, N., Offences of Endangerment (Actual and Abstract Danger). In P. Caeiro, S. Gless, V. Mitsilegas, M. J. Costa, J. De Snaijer and G. Theodorakakou (eds.), *Elgar Encyclopedia of Crime and Criminal Justice*. Cheltenham: Elgar Publishing, 2023.

[45] See the considerations in Simmler, Automation (n 4) or Simmler, Strafrechtliche Verantwortung (n 23), 158.

44.6.4 *Refining Doctrines of Agency and Imputation*

The spread of human–robot interaction and emergence of ever-more autonomous technical systems challenge what we regard as an "act" and "acting" in criminal law. The debate about this essentially revolves around questions of how to legally deal with the fact that we have technical systems that gradually take over agency. Distributed agency in increasingly complex socio-technical systems erodes our understanding of action and agency. The highly individualistic imputation of actions in criminal law is thereby tested. It is necessary to discuss how our doctrinal understanding of action can be further developed so that it has the capacity to better grasp and reflect such an interaction between humans and machines. Existing subjects should remain accountable, but it is a question of the extent to which the "coacting" of technology is included in the evaluation of their acts. At the same time, it should be examined how to appraise the effective delegation of agency to autonomous technical systems. Predictability is the key prerequisite of the existing forms of prior responsibility. This is precisely what is torpedoed by technical autonomy.

If an autonomous robot is used in a retirement home, then the human caregiver who installs and operates it will no longer act as if they were providing the care themselves. It would be inappropriate to not considering technical autonomy here in the evaluation of their actions. The "coacting" of the robot will be reflected in the legal appraisal of the human action. If the user no longer acts at all and comprehensively cedes agency, there is no act to attribute to them in the event of damage. At lower levels of automation (i.e., if the human user still has a role in operating or monitoring the robot), he or she still acts. A negligent act can be attributed to the user if they violate a duty of care (e.g., fail to perform their monitoring task). In this case, an accusation can be made against the user. However, it must be considered that the fault only consisted of the lack of monitoring and not an active action. In addition, it must be taken into account whether the system was difficult to monitor, due to its autonomy. However, if the situation was controllable by the user, the attribution of action would not fundamentally change and the existing doctrines could be utilized. The imputation is for the lack of control of an automated interaction. This principle applies not only to the caregiver, but also those actors who poorly designed the automation in advance and violated the duty of care that led to a predictable and preventable harm.

As discussed earlier, the situation is different when no subject can be addressed, due to full automation or a high degree of technical autonomy. Agency has been ceded. Damages that result are no longer predictable and the course of causality no longer controllable. Placing blame is complicated. A temporal advance of the blameworthy action could take place. The actors could be accused of no longer having acted situationally, but rather of having been able to anticipate that something might happen. The allegation would be that they absolved themselves of responsibility as agents. Similar to the "automation offense" outlined earlier, the accusation would then be based on the automation itself. Different from that, however, it would still refer to the later causation of the damage. It would be a matter of an *actio libera in causa* of a new kind (i.e., a situation in which actors in the moment in which they could decide otherwise choose to enter a state in which they are no longer capable of acting). In contrast to the conventional *actio libera in causa* in cases of intoxication, this state would consist of the cession of the authority to act in response to the technology. In an intentional *actio libera in causa*, the individual would consciously release him or herself from responsibility but still want the damage to occur. Conversely, in the case of a negligent *actio libera in causa*, he or she would not want the damage but could foresee that it could occur. Here, too, the occurrence of damage

must be recognizable, at least in its outlines and main features, which is why not all problems dealing with autonomous technology would be solved. Nevertheless, it would be a starting point to focus on the cession of agency itself and discuss the conditions under which the cession might already present a criminal act.

It would furthermore be possible to draw upon current forms of criminal liability for omission. In the case of an offense of omission, persons are punished who have a guarantor status and, in violation of the guarantor's legal duty, allow a criminal act to take place.[46] They are punished for doing nothing, though they could and should have done something. By analogy, a form of accountability can be thought of in which people who design human–machine interactions with a high level of automation and technical autonomy become guarantors by default, since such constellations are regarded as inherently dangerous. The guarantor's legal duty would consist, for example, of the duty to adequately select, instruct, and supervise the technical system.[47] In contrast to the offense of negligence, the accusation would not consist of a violation of due diligence, but rather of the guarantor's specific obligation to control the source of danger: that is, the robot. Such a construct is not without problems. For example, it is not immediately clear how *de facto* strict liability can be prevented and who, precisely, would be the guarantor. Nevertheless, the further development of the doctrine on offenses of omission could provide a foundation for establishing an understanding of how we are to deal with such situations. Either way, one must not only focus on the causation of harm. At the heart of the matter is the question of what a legitimate and proper delegation to machines is. This is in line with the ethical debate about AI that focuses on the question of responsible delegation.[48] In the end, we must simply decide what delegation is acceptable and what is not. As seen, this question can be approached in various ways in criminal law doctrine.

44.7 CONCLUSION

The examination undertaken in this chapter does not purport to be a complete and definitive analysis of the criminal law issues that arise in the context of human–machine interaction. That would be presumptuous and premature. I hope, however, that some of the challenges have become visible and their contours clearer. Not only was the idea of a self-reliant robot responsibility rejected (given the state of technological progress and current level of social integration), it was also determined that a focus on the machine itself is too narrow. An increased emphasis on interaction and socio-technical interplay is more appropriate in this regard. Therefore, one should question any rigid adherence to individualism in criminal law, since such an approach attempts to isolate overly individual acts by particular (human) actors from their social embedment. A graduated understanding of agency and recognition of the importance of human–machine interaction is key to an evolved understanding of criminal responsibility in the digital age.

As long as machines are both determined and predictable, traditional doctrine will capture constellations of human–robot interaction, though the high complexity of the interaction will

[46] See, for example, for Switzerland, Article 11 of the Swiss Penal Code. For a US-American perspective see, for example, Robinson, P. H., Criminal Liability for Omissions – A Brief Summary and Critique of the Law in the United States. *New York Law School Review* 29(1), 101–124, 1984.

[47] In Switzerland, these obligations are called the three "curae" ("cura in eligendo," "cura in instruendo," and "cura in custodiendo"). See, for example, Staffler, L., Criminal Liability Concepts for Entrepreneurs and Companies. In L. Staffler (ed.), *Business Criminal Law: A Primer for Management and Economics*. Wiesbaden: Springer, 179, 2022.

[48] Compare Taddeo and Floridi, AI (n 35), 751.

complicate the attribution processes and law enforcement. Further problems will arise in cases where no situational actor can be identified. This is the case in highly automated scenarios and even more so when autonomous technology is used. The causation of harm cannot simply be attributed to an actor acting intentionally or negligently. Responsibility gaps may emerge. Whether they should be closed or consciously accepted will depend on what other legal systems, especially that of tort law, can offer, and how sufficient they are in stabilizing norms by means other than punishment.

If criminal law is to be applied, it is important to establish functional mechanisms. The introduction of E-Persons does not seem promising, at least not for the time being. Other options reside in the embedding of new norms that clearly define what automation is legitimate and what is not. With such an automation offense, due diligence standards cannot be established in terms of the damage caused, but rather according to the bad or unjustified abandonment of agency. Of course, this would mean an extension of responsibility. Endangerment resulting from the use of such technology would be targeted, albeit only under certain conditions. However, criminal responsibility would be based on the act of automation itself (i.e., the focus on accountability would shift to the act of delegation). The same shift could result from a modified understanding of what constitutes an act in the criminal justice system. Delegation can be seen as an act that lays the groundwork for the attribution of responsibility. However, it must be ensured that the attribution of acts is not arbitrary and follows clear and recognizable rules. Establishing what a legitimate delegation of agency to technology is and what is not is essential for addressing challenges regarding technologically advanced systems, especially AI.[49] Misused delegations must be prevented. With this objective in mind, the doctrine of imputation should be revisited. Criminal law can and should play an important role in ensuring the responsible use of robots and AI. We are not yet at a point where we can clearly see what normative expectations society is developing regarding the use of either, but it is certainly our job to find out.

[49] Ibid.