Criminal Law Forum
https://doi.org/10.1007/s10609-018-9360-0

MONIKA SIMMLER* and NORA MARKWALDER**

GUILTY ROBOTS? – RETHINKING THE NATURE OF CULPABILITY AND LEGAL PERSONHOOD IN AN AGE OF ARTIFICIAL INTELLIGENCE

ABSTRACT. Robots and Artificial Intelligence are conquering our world. Just as any progress, this development is expected to have a relevant impact on law in general as well as on criminal law in particular. It involves the potential of transforming our conception of criminal responsibility, as notions of personhood, capacity and culpability will not stay unaffected. This article aims at giving an overview of the potential conversion criminal law is facing due to the increased importance of robotics and of artificial intelligence in our everyday lives. The discussion starts with an overview of different scenarios of criminal liability in the context of robotics. While some of them can be faced with existing doctrines, others demand a more far reaching assessment of the question if robots could ever gain legal personhood and therefore be originally called to account. While the picture of robots as liable perpetrators seems implausible at first sight, the present analysis reveals that blameworthiness is inherently socially constructed. However, it is not randomly constituted and follows social interaction and social meaning in fulfilling a certain function. Enabling the possibility of robots' criminal liability therefore would require that robots are regarded as a suitable agent of responsibility. The article lights up the conditions for such social and legal change in rethinking the very nature of culpability having the overall function of criminal law in mind. It can be concluded that the on-going technological progress definitely has the potential of testing the theory of criminal responsibility while more clearly unveiling its foundations and its sociological implications. A guilty robot, however, as fictional as that appears today, may be nothing unrealistic nor unlikely in the future.

* Monika Simmler, PhD, Postdoctoral Fellow of Criminal Law, Law of Criminal Procedure and Criminology at the University of St. Gallen, Tigerbergstrasse 21, 9000 St. Gallen, Switzerland. E-mail: monika.simmler@unisg.ch.

** Nora Markwalder, Assistant Professor of Criminal Law, Law of Criminal Procedure and Criminology at the University of St. Gallen, Bodanstrasse 4, 9000 St. Gallen, Switzerland. E-mail: nora.markwalder@unisg.ch.

Published online: 04 December 2018

I INTRODUCTION

1.1 *Robots as a Challenge for Criminal Law Theory*

When Nao is sad, his shoulders cave forward and he looks towards the floor. When he is happy, he throws his arms into the air and wants a hug.[1] Pepper is similar. He shows emotion and is a lovable companion. Pepper, however, is afraid of the dark and feels a little bit lonely when he is being ignored for too long.[2] Pepper and Nao are no boys next door or pets, as one might expect. They are humanoid robots of the kind that are increasingly conquering our world and – probably quicker than expected – our everyday lives. Nao and Pepper are not singular cases. The impressive and revolutionary progress in the context of robotics is a reality and it is going to leave a mark on our future. Just as any form of technological progress, this progress is expected to have an impact on the law.[3]

The exceeding number of headlines regarding innovations in the field of robotics, however, do not only engender enthusiastic reactions. Increasingly, they are a source of worries and concerns regarding the social implications of this progress and – of particular interest to us – regarding the consequences of these developments for the law. Whereas questions of civil liability have been raised quickly, criminal law has been lagging behind. It might have been perceived as absurd for criminal lawyers to address the question of a possible criminal responsibility of robots. In the meantime, however, these questions have also been the subject of some publications[4] and in the

[1] Jha, 'First robot able to develop and show emotions is unveiled', *The Guardian* (09 August 2010), available at: https://www.theguardian.com/technology/2010/aug/09/nao-robot-develop-display-emotions (last visited 01 November 2017).

[2] Schwiegershausen, 'The World's First Robot with Feelings is a Big Hit', N.Y. *Magazine* (22 June 2015), available at: http://nymag.com/thecut/2015/06/worlds-first-robot-with-feelings-is-a-big-hit.html (last visited 01 November 2017).

[3] Beck, 'Grundlegende Fragen zum rechtlichen Umgang mit der Robotik', *Juristische Rundschau* 6 (2009), 230 [hereinafter Beck, Robotik].

[4] *See, eg, for the criminal Law:* Hallevy, *Liability for Crimes Involving Artificial Intelligence* (2015) [hereinafter Hallevy, Artificial Intelligence]; Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law* (2013); Calo, 'Robotics and the Lessons of Cyberlaw', *California Law Review* 103 (2015); Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013); Beck, 'Intelligent agents and criminal law – Negligence, diffusion of liability and electronic personhood, *Robotics and Autonomous Systems* 86 (2016) [hereinafter Beck, Intelligent agents and criminal law].

context of the emerging field of 'the law of robotics',[5] they are no longer just of interest for science fiction fans, but also for lawyers and legal scholars.[6]

Questions of criminal law, or legal questions in general, regarding autonomous vehicles are presently receiving the bulk of the attention.[7] This is, of course, due to the fact that these vehicles are projected to become an essential part of our everyday lives in the near future. Especially incidents like the collision of a google car with a bus in early 2016[8] or the first lethal accident with a partially autonomous vehicle[9] only a few months later got the attention of the media and raised some questions regarding legal responsibility, which are not only relevant in theory but already in present-day practice.

It is hardly a matter of dispute that robots can be relevant as a means to a criminal end. Neither is it a contested matter that robots can cause harm and can infringe on legally protected interests and that questions of criminal responsibility can arise in this context.[10] The groundbreaking and therefore disputed question, however, is whether robots can be the perpetrators (or possibly victims) of a criminal offence, ie if they can be regarded as subjects of criminal law. According to existing traditional accounts, robots cannot be made criminally responsible. This, according to Gless, Silverman, and

[5] This area of law is not yet clearly defined. For the moment, this term simply applies to the area of research that is dedicated to a legal analysis of subject matters related to robots. *See* Müller, 'Roboter und Recht. Eine Einführung', *Aktuelle Juristische Praxis* 5 (2014), 597.

[6] With a reference to Star Trek: Gless, Silverman and Weigend, 'If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability', *New Criminal Law Review* 19(3) (2016), 415.

[7] *See, for example, id., and furthermore* Douma and Palodichuk, 'Criminal Liability Issues Created by Autonomous Vehicles', *Santa Clara Law Review* 52(4) (2012); Gurney, 'Driving Into the Unknown: Examining the Crossroads of Criminal Law and Autonomous Vehicles, *Wake Forest J.L. & Pol'y* 5 (2015).

[8] Davies, 'Google's Self-Driving Car Caused Its First Crash', *Wired* (29 February 2016), available at: https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash (last visited 01 November 2017).

[9] Klein, 'Tesla driver dies in first fatal autonomous car crash in US', *New Scientist* (01 July 2016), available at: https://www.newscientist.com/article/2095740-tesla-driver-dies-in-first-fatal-autonomous-car-crash-in-us (last visited 01 November 2017).

[10] cf Calo, *supra* note 4, 513, with the remark that 'robotics combines, for the first time, the promiscuity of data with the capacity to do physical harm'.

Weigend, is due to the fact that they are not conceived as morally responsible agents and that they cannot be the addressees of retribution in the form of punishment, or – put slightly differently – that they don't have the capacity to understand the concept of punishment.[11] The subject of the present paper is whether this traditional account can be maintained or whether it is still adequate (or for how long it will stay adequate) and which questions arise in the context of the theory of criminal responsibility. At the outset, it is beyond doubt that the question of criminal responsibility of robots will (once again) push the boundaries of criminal law scholarship[12] and that robotics could prove an 'exceptional' occasion for changes to the law and legal theory.[13]

Of course, many of the developments and theoretical consequences discussed in this paper can be characterized as pertaining to a distant future. However, the tendency to wait and not to address the consequences of current developments early is particularly hazardous for legal scholarship.[14] This could result in the consequence that our discipline can only react to existing technological developments after the fact. It is therefore desirable to analyze questions regarding robotics at an early stage, not least because of the fact that the legal system is "the organ of society that is used for turning a change in public opinion into a legal form".[15] Furthermore, and maybe more importantly, this debate, which is slowly but surely gaining the attention of criminal lawyers, sheds new light on well-known problems.[16] Namely, it is unfolding its potential in the context of a question that has been debated for a long time, although may have been getting too little attention in the recent past: the debate on criminal responsibility, ie the heart of criminal law theory.

[11] Gless, Silverman and Weigend, *supra* note 6, 412.

[12] Id., 435; Hilgendorf, 'Können Roboter schuldhaft handeln?', in Beck (ed), *Jenseits von Mensch und Maschine* (2012).

[13] Calo, *supra* note 4, 513.

[14] Beck, Robotik, *supra* note 3, 230.

[15] Luhmann, *Law as a Social System* ([Kastner et al. (ed), Ziegert (trans)] 1993/2004), p. 119 [hereinafter Luhmann, Law as a Social System].

[16] Wohlers, 'Individualverkehr im 21. Jahrhundert: das Strafrecht vor neuen Herausforderungen', *Basler Juristische Mitteilungen* 3 (2016), 114.

1.2 *Robots, Intelligent Agents, and Cyborgs: Conceptual Clarifications*

Before we address the question of whether we want to send robots to prison someday, we should pin down our concept of a 'robot'.[17] The search for a standard definition has shown to be less than fruitful in the past and a series of overlaps with the concepts of 'autonomous systems' or 'intelligent agents' complicate the conceptualization.[18] Due to the fact that the main concern of the present paper is legal theory and not technology, the concept of a robot will only be sketched briefly and expediently. According to the *International Organization for Standardization (ISO)* a robot is an 'actuated mechanism programmable in two or more axes with a degree of autonomy, moving within its environment, to perform intended tasks'.[19]

This attempt at a definition, just as others emanating from the field of technology, is of little guidance for the layperson not familiar with robotics. For present purposes, it suffices to say that robots are machines (mostly with sensory-motor functions), which are built to enhance human possibilities for action.[20] The feature that distinguishes robots from regular machines is that they act freely to a certain extent,[21] ie that they possess a minimum degree of autonomy.[22] When we talk about robots in the present paper, we talk about intelligent machines with the ability to learn and built for the purpose of enhancing human possibilities for action, knowing that this definition is probably imprecise from a technological point of view.[23]

---

[17] Etymologically, the term can be traced back to the Slavic term *robota*, which means 'work'. The fact that the term became widely spread first in literary fiction lead to the consequence that there is no standard definition for the term. Additionally, the different categories are difficult to harmonize. Beck, Robotik, *supra* note 3, 236.

[18] Regarding the respective attempts at a definition, *see* Müller, *supra* note 5, 596.

[19] International Organization for Standardization, *ISO No 8373 (2.6) – Robots and robotic devices – Vocabulary*, available at: https://www.iso.org/obp/ui/#iso:std:iso:8373:ed-2:v1:en (last visited 01 November 2017).

[20] Scholtysek, 'Wann ist ein Roboter ein Roboter?', *Roboterwelt* (03 March 2015), http://www.roboterwelt.de/magazin/wann-ist-ein-roboter-ein-roboter (last visited 01 November 2017).

[21] Christaller et al, *Robotik - Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft* (2001), p. 19.

[22] In contrast to remote-controlled machines, robots do not need continuous external inputs, but instead act autonomously within their programming. *See* Scholtysek, *supra* note 20.

Similar difficulties in finding a definition arise regarding the concept of artificial intelligence. The primary concern in the context of 'AI' is to emulate human approaches to problem-solving, ie 'human intelligence'.[24] So-called intelligent agents operate according to fixed and predefined rules, but they process information for individual cases autonomously.[25] In the following sections, we will employ the term 'robot' for 'intelligent agents', for 'robots', as well as for the term 'autonomous machines', knowing that this entails a severe simplification of the differentiation among these categories and diverse subcategories. In any case, it is clear that we are neither talking about regular data processing systems nor about automated but precisely determined systems, but about 'intelligent systems', which process large amounts of data and which act with a certain degree of autonomy by referring to this data in their decision-making process.[26] We are also not talking about 'cyborgs', although they would deserve some attention in the context of questions of criminal responsibility.[27]

There are different ways of building a typology of robots. The primary distinction is the distinction between industrial robots and service robots.[28] The main focus of interest of our research question is on social robots, ie physically embodied robots that interact with

[23] In doing so, we adopt the approach of Günther, *Roboter und rechtliche Verantwortung* (2016), p. 19.

[24] For more details on the terminology and the concept, *see* id., 23 et seq.

[25] This is the definition adopted by Gless and Weigend, 'Intelligente Agenten und das Strafrecht', *ZStW* 126(3) (2014), 561, following Russell and Norvig, *Artificial Intelligence: A Modern Approach* (3rd ed, 2014), pp. 34–63.

[26] For an overview of the differences, *see* Gless and Weigend, *supra* note 25, 562–563.

[27] The term 'cyborg' usually refers to a 'mix of human and machine'. This term, too, however, lacks a precise conceptualization. Cases of cyborgs are different in nature, because they are still humans. The technological 'enhancements' could, however, become relevant in the context of debates on criminal responsibility. Beck, Robotik, *supra* note 3, 228–229, gives the example of machines that directly influence the brain (eg brain 'pacemakers'). Just as in cases of alterations caused by pharmaceuticals, it seems to be a plausible hypothesis that the patients have only limited or no criminal responsibility if they cease to be fully autonomous. However, it is hardly disputed that these patients still are, in principle, subjects of criminal law and that, as such, they can still potentially be criminally responsible. On this matter, *see furthermore* Müller, *supra* note 5, 597.

[28] *See* Müller, *supra* note 5, 597, building on the typology of the International Federation of Robotics (IFR), available at: http://www.ifr.org (last visited 01 November 2017).

humans on an emotional level.[29] Especially the development of humanoid robots,[30] which are able to perform emotional interactions, has to be discussed in the context of attributing responsibility. It can be expected that next to autonomous cars and the like, humanoid robots will be the dominating – albeit not the exclusive – focus of the debate around criminal culpability as they are particularly qualified to coin our lay picture of robotics.

1.3 *Different Scenarios of Criminal Responsibility in the Context of Robotics*

The fact that robots are more and more becoming a part of our everyday lives engenders new and increasingly pressing questions of criminal law.[31] These questions can be different in nature and they can touch on both doctrinal and theoretical issues. Hence, in the context of criminal responsibility of robots, various constellations have to be distinguished. There are roughly for different scenarios in which the question of criminal law sanctions arises:

The first version includes cases in which a robot 'commits' a crime because it was deliberately programmed to do so. Examples are the 'killer robot', well-known from literary fiction, but also drones, which are ubiquitous nowadays, and highly developed military robots.[32] In these cases, it is obvious that the person behind the robot can be held responsible according to existing criminal law.[33] The robot constitutes an instrumentality or, in case of an increasing personalization of robots, which has yet to be discussed, one could at most categorize this scenario as a case comparable to 'innocent agency'[34] or 'perpetration by means'.[35]

---

[29] Müller, *supra* note 5, 597.

[30] The robot 'Asimo', developed by Honda, is another example of a fully developed humanoid robot, besides the examples given in the introduction. Information on this advanced humanoid robot is available at: http://asimo.honda.com (last visited 01 November 2017).

[31] Gless, Silverman and Weigend, *supra* note 6, 412.

[32] Currently, the most advanced military robots are probably the ones developed by Boston Dynamics, for example the robot 'Atlas'. More information can be found on the company's website, available at: http://www.bostondynamics.com/robot_Atlas.html (last visited 01 November 2017).

[33] Gless, Silverman and Weigend, *supra* note 6, 412.

[34] Ashworth and Horder, *Principles of Criminal Law* (7th ed, 2013), pp. 108–109.

[35] Binder, *Criminal Law* (2016), pp. 285–286.

This has to be distinguished from cases belonging to the second scenario, in which a robot 'commits' a crime because of faulty programming. These cases are probably the most common ones and the main questions to be discussed will concern the issues of adequate risk management and due diligence.[36] Criminal responsibility of the programmer and of the corporation behind the robot depend on the answers to these questions. Questions of causation and imputability will also have to be discussed in greater detail in this context.[37] It can hardly be avoided to address these questions from a theoretical angle in a timely manner. The question of the amount of risk that society is willing to take will be decisive for the developers who are responsible and for further technological progress. An approach that does not accept any risks would entail that developers have to abandon their projects if they don't want to face charges of negligence every time something goes wrong.

A similar category is constituted by a third scenario, which will keep moral philosophy busy, namely cases in which a robot has to make a decision in a case representing a classical moral dilemma or in which the programmer has already made a decision and there is a concrete instance in which a damage occurs because of that decision. An example is a case in which a child is run over by the autonomous car to save two other children. The robot that is programmed this way or the person who programmed the robot acts with criminal intent. The question that arises is what defenses may be raised. The more fundamental question of how a society decides to program these moral dilemmas and who decides how to program them has also been the subject of debates in legal philosophy.[38] However, since these types of moral dilemmas are probably unsolvable whether they arise in a robotics or non-robotics context, their discussion remains mainly academic for now.

The fourth scenario is the one which will be the main focus of the present paper and it differs from the scenarios that have been described so far. This scenario includes cases in which a robot 'commits' a crime, because it has developed its own momentum due to its artificial intelligence. Of course, this momentum is also pre-determined and depends on the programming. However, the momentum cannot be traced back to a single programming operation. This

[36] Müller, *supra* note 5, 604–605.

[37] Beck, Robotik, *supra* note 3, 227; Müller, *supra* note 5, 604–605.

[38] *On this topic see, eg,* Lin, 'Why Ethics Matters for Autonomous Cars', in Maurer, Gerdes, Lenz and Winner (eds), *Autonomes Fahren* (2015), p. 69.

means that we are facing a situation that is similar to the debate on free will, a situation in which we can assume that every action has been caused and determined by something somewhere, but in which we attribute this action to a person as 'their own' and as an 'act of (free) will', because we cannot trace and explain the exact process of causation. Keeping in mind the current pace of technological progress, these cases do not seem too futuristic anymore. The extreme case is that of an advanced 'killer robot pro' that has gotten out of control. Smaller, or at least less spectacular offences, however, are also far from unthinkable in the context of artificial intelligence and humanoid robots.[39] It is easy to pictures intelligent systems developing a certain autonomy and breaking norms, reaching from social bots[40] committing defamation offenses online to Wall Street algorithms violating financial market laws.[41] The question whether and to what extent these cases may result in genuine criminal responsibility of robots can and should no longer be called science fiction. Thus, the following sections will not be devoted to questions regarding criminal responsibility of the person behind the robot, the manufacturer or programmer. Instead, they will be dedicated to questions regarding potential criminal responsibility of the robot itself.[42]

II FREE WILL AND FREEDOM OF CHOICE IN ACTIONS OF ROBOTS

2.1 *First Neuroscience, now Robotics: Attacks on the Foundations of Criminal Responsibility*

[39] cf Cerka, Grigiene and Sirbikyte, 'Is it possible to grant legal personality to artificial intelligence software systems?', *Computer Law & Security Review* 33(5) (2017), 688.

[40] 'Social Bots' are algorithms, ie automatic or semi-automatic computer programms that aim to mimic humans in online social networks; *so the definition in* Wagner, Mitter, Körner and Strohmeier, 'When social bots attack: Modeling susceptibility of users in online social networks', in Proceedings of the 2nd Workshop on 'Making Sense of Microposts', Conference on the World Wide Web (2012), 41, available at: http://ceur-ws.org/Vol-838/paper_11.pdf (last visited 01 November 2017).

[41] Today stock markets are already heavily influenced by algorithms and 'algorithmic trading has overtaken the industry'; *see* Salmon and Stokes, 'Algorithms take control of Wall Street', *Wired* (29 February 2016), available at: https://www.wired.com/2010/12/ff_ai_flashtrading (last visited 01 November 2017).

[42] On the topic of a possible criminal responsibility of the 'person behind the intelligent agent' see for example Hallevy, Artificial intelligence, *supra* note 4.

Punishment of robots as it is envisaged in this paper would have to be denied due to a lack of personal reproachability according to traditional doctrine and according to the positions held by most criminal law scholars. While the *actus reus* may cause some doctrinal problems as well, these problems are quickly overshadowed by the problems arising in the context of *mens rea*. Justifiably, the first critical question is whether a robot can ever be 'guilty', if criminal law can ever place the blame of criminal responsibility – the element which defines the identity of criminal law in common as well as civil law tradition – on a machine. It is not by accident that the question of culpability has always been called the 'fatal question' of criminal law.[43] Thus, the fate of criminal responsibility of robots will also depend on this question.

A summary examination quickly leads to the conclusion that there are mainly two problems regarding the attribution of a crime to a robot in the context of the doctrine and theory of criminal responsibility. The first is (the assumption of) human free will as the foundation of the traditional theory of criminal responsibility. The second problem is connected to the first and regards assessing the question whether intelligent robots could be recognized as a subject of criminal law and therefore concerns the very basic concept of legal personhood.[44] Both problems seem to exclude responsibility of robots at first glance. The content of both of these elements of the theory of criminal responsibility will now be examined and their potential application in the context of the responsibility of robots will be explored. Only if there is clarity on the conditions under which a person is considered guilty within a society, it is possible to determine under which conditions robots may be considered guilty in the future.

The traditional requirement of a guilty mind in criminal law rests on a personal reproach and accusation. The person at whom this reproach and accusation is directed must have had the opportunity to behave in a different manner.[45] The point of departure of this theory of criminal responsibility traditionally is free will.[46] According to Ashworth and Horder, the essence of the *principle of mens rea* or *fault*

[43] *As stated early on by* Hafter, *Lehrbuch des Schweizerischen Strafrechts, Allgemeiner Teil* (2nd ed, 1946), p. 101.

[44] *See on this question also* Cerka, Grigiene and Sirbikyte, *supra* note 39. Already in 1992, *Solum* brought up the consideration of legal personhood for artificial intelligences; *see* Solum, 'Legal Personhood for Artificial Intelligences', *North Carolina Law Review* 70(4) (1992).

[45] Wohlers, *supra* note 16, 123–124.

principle consists in the statement that people can only be held criminally responsible for events or results they intended or knowingly risked and if they can be blamed for making a choice in favor of a certain behavior in spite of being aware of the consequences. This approach is deemed to be an expression of the principle of autonomy: Individuals are considered to be autonomous persons who generally have the ability to choose between alternative actions.[47] In a similar vein, Williams states that the requirement of *mens rea* is 'a mark of advancing civilization'[48] and Hall stresses that criminal law rests on the same foundations as traditional ethics and that therefore people are only punished for acts for which they can be declared responsible in the moral sense.[49] Hart also subscribes to this fundamental tenet of liberal thought by affirming that a person can only be criminally responsible if that person has the capacity and a fair chance to act in the way that is expected from them and if therefore it can be said that this person has consciously decided not to act in accordance with these expectations.[50]

If one adopted this traditional definition, it would be impossible to think of a 'guilty' robot, regardless of the direction that technology will take in the future.[51] A traditional understanding of criminal responsibility entails and even dictates the clear rejection of any form of 'guilt' or 'responsibility' of robots that is contained in this first impulse. A robot is not 'a person with free will'.[52] This indeterministic approach to traditional theory indubitably profits from individuals' subjective experience of freedom.[53] Over the past 50 years, this

[46] On the relevance of free will for criminal responsibility *see, eg,* Green, *Freedom and Criminal Responsibility in American Legal Thought* (2014).

[47] Ashworth and Horder, *supra* note 34, 74, 155.

[48] Williams, *Textbook of Criminal Law* (2nd ed, 1983), p. 70.

[49] Hall, 'Interrelations of Criminal Law and Torts', *Columbia Law Review* 43(6) (1943), 776.

[50] Hart, *Punishment and Responsibility. Essays in the Philosophy of Law* (2nd ed, 2008), p. 152; *see furthermore on this conception* Mitchell, 'In Defence of a Principle of Correspondence', *Criminal Law Review* (1999), 197; Tadros, *Criminal Responsibility* (2005), p. 57 et seq.

[51] Gless and Weigend, *supra* note 25, 574; Wohlers, *supra* note 16, 123–124.

[52] If one adopts this definition, however, it is also doubtful whether it is ever possible to deem humans criminally responsible. This will be discussed in more detail in the following sections.

[53] Geisler, *Zur Vereinbarkeit objektiver Bedingungen der Strafbarkeit mit dem Schuldprinzip* (1998), pp. 84–85.

indeterministic understanding has been subjected to various attacks and its claims have been relativized. However, relativizing the requirement of an indeterministic free will and the corresponding attempts of reformulations brought up for example in the form of an 'actor's capacity for rational choice' or of an 'ability to engage in morally responsive reasoning' have not erased the fundamental problem, given that the question of individual discretion remains and has to remain unanswered.[54]

The exhaustive debate regarding new insights from neuroscience and their implications for criminal law,[55] which cannot be reproduced here in detail, has shown already that it is neither clearly established which assumptions about humans and their abilities are at the basis of the possibility of criminal conviction nor is there clarity how these assumptions play out in attributing responsibility to a person.[56] In the context of the question of a possible criminal responsibility of robots, this debate is gaining in relevance. The question of what distinguishes an intelligent system from a criminally responsible human being will be increasingly difficult to answer in the future. Experts seem to hardly doubt that at some point there will be robots with decision-making capacity similar to humans.[57] Hence, if the concept of criminal responsibility really depends on free will – as posited by traditional theory – this concept is again faced with potential instability due to technological progress in the field of robotics and artificial intelligence, just as it has been faced with instability due to progress in neuroscience.

The position of this article is that it does not make sense to succumb to the vortex that is the debate around the ability to choose one's actions[58] and to a scientific analysis of determinant variables. This is not only due to the fact that it is impossible to find scientific

[54] *See on these different notions with further references* Green, *supra* note 46, 345–346. Different emphases of concepts, discussed in English literature, eg, as 'theories based on capacity', 'theories based on choice' and 'theories based on character' cannot relevantly change this fundamental problem neither; *see on this differentiation* Tadros, *supra* note 50, 22.

[55] *On this debate see, eg,* Freeman (ed), 'Law and Neuroscience', *Current Legal Issues* 13 (2011); Morse and Roskies (eds), *A Primer on Criminal Law and Neuroscience* (2013).

[56] Beck, Robotik*, supra* note 3, 229–230.

[57] id., 229.

[58] Bleckmann, *Strafrechtsdogmatik – wissenschaftstheoretisch, soziologisch, historisch: das Beispiel des strafrechtlichen Vorsatzes* (2002), pp. 75–76.

proof for individual freedom. The primary reason is that free will as a 'biophysical fact' is hardly relevant for criminal law as a system. Free will is a part of the social construction of reality.[59] 'Human free will' is constituted in the social system. The thesis of this paper is that this could also be the case for the freedom of robots.

2.2 *Freedom of Robots in the Social System*

Just as for anyone else, it is impossible for legal scholars to cognize an 'objective reality' or 'truth'' independent from the observer.[60] This also applies to inquiries about free will. These epistemological limitations, however, do not entail that the assumption of free will and the attribution of criminal responsibility based on this assumption necessarily have to be characterized as fiction.[61] Free will exists in the social system called 'society' in the sense and to the extent that it shapes interaction, social relations, and the law.[62] This social relevance has little in common with a concept of freedom based on natural sciences or with the questions that neuroscience is trying to answer. The question in this social reality familiar to us is rather how agents attribute certain traits and capacities to themselves and to others.[63] The opposition between determinism and indeterminism concerns an ontological problem.[64] The question of attributing certain degrees of autonomy to individual agents, on the other hand, concerns a problem of social reality, of society, and, therefore, also of the legal system.

---

[59] Schünemann, 'Die Funktion des Schuldprinzips im Präventionsstrafrecht', in Schünemann (ed), *Grundfragen des modernen Strafrechtssystems* (1984), pp. 153, 163–164 *citing* Berger and Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (1966).

[60] On this basic assumption of constructivism, which is adopted in the present paper, *see, eg,* von Glasersfeld, 'The Radical Constructivist View of Science', *Foundations of Science* 6 (2001); Searle, *The Construction of Social Reality* (1995); Segal, *The Dream of Reality: Heinz von Foerster's Construcivism* (2001).

[61] Such a characterization can be found in Kohlrausch, 'Sollen und Können als Grundlagen der strafrechtlichen Zurechnung', in Festgabe für Karl Güterbock (1910), p. 26.

[62] Hirsch, 'Das Schuldprinzip und seine Funktion im Strafrecht', *ZStW* 106(4) (1994), 764.

[63] von Glasersfeld, 'Konstruktion der Wirklichkeit und des Begriffs der Objektivität', in von Foerster et al. (eds), *Einführung in den Konstruktivismus* (12th ed, 2010), p. 34.

[64] Luhmann, 'Funktion und Kausalität', *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 14 (1962), 640.

Now what is at the basis of the concept of criminal responsibility if not 'human free will'? According to Jakobs, criminal responsibility does not refer to free will in a traditional sense, but to personal self-administration.[65] Self-administration in this sense is defined by a certain capacity to organize oneself and by the competence to organize oneself freely.[66] This concept aptly redirects criminal legal thinking away from ontology. Nevertheless, it is still based on a capacity that can hardly be confirmed objectively. Hence, despite various attempts at reconstructing the problem in a different way and especially because of the commotion caused by neuroscientific findings, some form of forced consensus had to be established. The substance of this consensus is that these capacities, this autonomy, these choices between different alternatives that are the subject of our discussions on the attribution of criminal responsibility are attributions in the context of current social circumstances, ie attributions in the social system rather than objectively and individually confirmable characteristics.[67] Thus, the attribution of responsibility in criminal law as it currently is *de facto* already refers only to contexts of social perception and not at all to the immense complexity of the links that lead to an action.[68] Criminal law, like law in general, actualizes itself through communication[69] and is real in the social sense if it provides social orientation.[70]

In this respect, Jakobs' statement that criminal law does not know the category into which free will belongs[71] is true only to the extent that he means the biophysical dimension of free will. Free will is not irrelevant or just a fiction, neither from a perspective of social science nor from a constructivist perspective. It is real, because, as a con-

---

[65] Jakobs, *Das Schuldprinzip* (1993), p. 34. [hereinafter Jakobs, Schuldprinzip].

[66] Jakobs, 'Strafrechtliche Schuld als gesellschaftliche Konstruktion', in Schleim, Spranger and Walter (eds), *Von der Neuroethik zum Neurorecht?* (2009), pp. 246, 248 [hereinafter Jakobs, Schuld als gesellschaftliche Konstruktion].

[67] *See, eg,* Gless and Weigend, *supra* note 25, 574.

[68] Geisler, *supra* note 53, 122.

[69] In the sense of systems theory, *see, eg,* Luhmann, *Social Systems* ([Bednarz, Jr. and Baecker (trans)], 1984/1995) [hereinafter Luhmann, Social Systems].

[70] Jakobs, 'Individuum und Person. Strafrechtliche Zurechnung und die Ergebnisse moderner Hirnforschung', *ZStW* 117(2) (2005), 266 [hereinafter Jakobs, ZStW 2005]; Jakobs, 'Das Strafrecht zwischen Funktionalismus und "alteuropäischem" Prinzipiendenken', *ZStW* 107(4) (1995), 867 [hereinafter Jakobs, ZStW 1995].

[71] Jakobs, 'Strafrechtliche Schuld ohne Willensfreiheit?', in Dieter (ed), *Aspekte der Freiheit* (1982), p. 71.

struct, it shapes the way in which responsibility is socially attributed – at least that is presently the case. This understanding, however, should not lead to simply circumventing the debate on free will or to making the traditional concept of criminal responsibility immune to criticism. Free will as the basis of criminal responsibility is indeed problematic and possibly dispensable.[72] As a matter of fact, however, the currently predominant concept of criminal responsibility is based on free will. This has to be taken into account when discussing a criminal responsibility of robots, but it does not exclude such a responsibility, given that it is not a biophysical or meta-physical problem but simply a sociological problem or a problem of social psychology.

The fact that it is, as just described, impossible to objectively prove the existence of free will has generated considerable disruption for the rationality and legitimacy of criminal responsibility in the past decades. Further, exposing the objective and absolute freedom as speculation has caused significant instability in the theoretical foundations of criminal law doctrine.[73] However, the fact that criminal responsibility is not based on a biophysically demonstrable freedom, but on the attribution of freedom as a social fact, paves the way for criminal responsibility of robots. Contemporary practice already shows that individuals are relatively 'generous' and that they are not inclined to grant much room to deterministic thoughts when it comes to attributing discretion and autonomy to other persons regarding their lives and their individual choices. If technological developments lead to a situation in which especially social robots not only perform simple interactions and services for humans, but employ highly complex processes, it is a question of time until humans experience this autonomy not just as determined and programmed and until they attribute robots the respective 'capacities'. It is important to stress at this point that already today the concept of criminal responsibility is not based on human free will, but on the attribution of such an autonomy in the context of social interaction.

III ROBOTS AS PERPETRATORS: THE CONCEPT OF A PERSON IN CRIMINAL LAW

3.1 *Personhood in Criminal Law*

[72] *See also* Hörnle, *Kriminalstrafe ohne Schuldvorwurf* (2013), p. 10 and *passim*.

[73] Geisler, *supra* note 53, 83–84.

The question of freedom of choice is, as we have seen, intimately linked not to the question of who possesses this freedom or if this freedom is restricted due to general determinism or determinism in a specific situation, but to the question of who is attributed this freedom and who is not. Whether robots can be deemed criminally responsible and who can, in general, be called 'responsible', consequently, leads us to the concept of the subject of criminal law.[74] The concept of the subject as a requirement of criminal responsibility has sparked just as much debate – until now especially in the context of the criminal responsibility of corporations – as the question of free will, which of course goes hand in hand with the question of personhood in criminal law.

Gless and Weigend, in their search for 'the line between machine and human person', refer to the genuinely 'human' subject in the traditional theory of criminal law.[75] They come to the conclusion that intelligent agents do not meet the criteria to qualify as a person in an idealistic-philosophical sense, because they are not aware of their freedom, they do not understand themselves as entities with a past and a future, and they do not possess the capacity to grasp the concept of rights and obligations.[76] Thus, even robots that have the ability to learn do not possess the consciousness and reflective capacity that would be necessary to count as 'free' agents. Hence, personal responsibility is impossible.[77] By stating this, the authors subscribe to a traditional understanding. Such an understanding, however, neglects the fact that the concept of the subject or of personhood in criminal law is constructed in social reality and does not necessarily refer to biophysical categories. Idealistic philosophy cannot obscure the fact that the attribution of capacity to reflect, of consciousness, and of other capacities is just that – an attribution – and not cognizable and legally meaningful due to ontological circumstances. Hence, also in this context the question is whether

[74] Gless, Silverman and Weigend, *supra* note 6, 416.

[75] Gless and Weigend, *supra* note 25, 568.

[76] Id., 569–570; Gless, Silverman and Weigend, *supra* note 6, 412.

[77] Gless and Weigend, *supra* note 25, 570; Gless, Silverman and Weigend, *supra* note 6, 417. The authors make it clear, however, that these statements apply to intelligent agents as we know them in the year 2016. They also state that, due to the steady and rapid progress, there is a possibility that robots will have more capacities in the future, which would make them appear human. If robots would acquire the capacity for self-reflection and something like consciousness, the personhood of robots would have to be discussed again.

criminal responsibility should be based on a traditional concept of personhood. An alternative would be to adopt a concept of personhood that depends on the respective agent's capacity to disappoint normative expectations and on the possibility of attributing actions to this agent in the social system. This question is a purely theoretical one, given that presently social practice already refers to the latter criterion and that traditional theory possesses idealistic value, but hardly translates into social operations.

Therefore, in the system of criminal law, one has to distinguish between the concepts of 'person' and 'human' and also between 'person' and 'individual'. In a society structured by norms, a person is the constructed addressee of rights and obligations.[78] A person in this sense is a compound of (normative and cognitive) expectations organized as identity, ie an object of attribution.[79] The concept of a person in criminal law thus constitutes an artificial concept developed by and for the purpose of observers of the social system,[80] which summarizes existing expectations. The orientation through the concept of personhood that is necessary for social interaction is not biophysical in character, which is why the principles that shape this concept cannot necessarily be derived from the natural sciences.[81] Similarly, the question of which 'capacities' someone has to have is not a question that can be answered by reference to ontology.[82] The attribution of capacities is a normative process in a specific society at a specific time. Due to this, the concept of personhood is subject to changes. This opens the door for a potential criminal responsibility of robots.

3.2 *Robots as Subjects of Criminal Law*

As described above, criminal law cannot access any kind of ontological reality 'behind' society, which would be able to answer the question of personhood.[83] The intuitive search for requirements like 'consciousness' or a 'sense of self' does not refer to biophysical cat-

[78] Jakobs, ZStW 2005, *supra* note 70, 266.

[79] Bleckmann, *supra* note 58, 107.

[80] Jakobs, ZStW 2005, *supra* note 70, 258.

[81] id.

[82] Jakobs, *Schuld und Prävention* (1976), pp. 20, 31 [hereinafter Jakobs, Schuld und Prävention].

[83] Bleckmann, *supra* note 58, 130.

egories, but to social categories that describe which traits we attribute to persons to derive responsibility. Personhood as well as responsibility are constructed and constituted in the 'social game'.[84] As Luhmann aptly states, persons cannot emerge and persist without social systems.[85] Or, to paraphrase Singer's slightly more dramatic formulation: A person acquires a self-identity only by gazing into the mirror of the other.[86] Persons are thus a social reality. They are not less real than the individuals of the phenomenal world.[87] This does not have to be translatable into a biophysical fact. In an analogy to Foucault, we can assume both that there are subjects and that the subject does not exist.[88]

This social relativity of persons is not only apparent in the present context of a postulated criminal responsibility of robots. Most legal systems have already abandoned the traditional concept of personhood and introduced a genuine criminal responsibility of legal persons.[89] By stating that our criminal law is made for human beings,[90] Gless and Weigend adopt a German perspective.[91] In most other Western legal systems, criminal responsibility of corporations is already a matter of course. Thus, non-human entities are already accepted as subjects of criminal law in many countries in the context of criminal liability of corporations.[92] As will be described more precisely in the following sections, this is due to the fact that corpora-

[84] Opitz, 'Was ist Kritik? Was ist Aufklärung?', in Amstutz and Fischer-Lescano (eds), *Kritische Systemtheorie* (2013), pp. 40–41.

[85] Luhmann, Social Systems, *supra* note 69, 59, adding 'nor can social systems without persons'.

[86] Singer, *Ein neues Menschenbild? Gespräche über Hirnforschung* (2003), p. 56; *cited as well in* Jakobs, ZStW 2005*, supra* note 70, 249.

[87] Jakobs, *Schuld als gesellschaftliche Konstruktion*, *supra* note 66, 246–249.

[88] Foucault, *Die Wahrheit und die juristischen Formen* ([Bischoff (trans)], 4th ed, 1973/2015), p. 21.

[89] Criminal responsibility of legal persons has a long tradition in common law legal systems. In the legal systems of continental Europe, however, this is a relatively new phenomenon. However, this phenomenon is spreading rapidly, *according to* Weigend, 'Societas delinquere non potest? A German Perspective', *Journal of International Criminal Justice*, 6(5) (2008), 928.

[90] *So in* Gless and Weigend, *supra* note 25, 566.

[91] Germany has introduced a criminal responsibility of corporations for offences committed by their employees. This, however, only applies to lesser offences not contained in the criminal code. In general, corporate liability is not provided.

[92] Müller, *supra* note 5, 604; Weigend, *supra* note 89.

tions, under certain circumstances, can destabilize norms and disappoint expectations, too.

However, social attribution of personhood or of other capacities is not at all arbitrary. According to Jakobs, in order to postulate an entity as a person in law (or even to think of it as such), it is not enough that it is an addressee of rights and obligations. For the constitution of a person, it must also be possible that this addressee provides orientation as a person in the law.[93] Through the actions of their representatives, corporations are perceived as entities which provide orientation. We have certain normative expectations towards them and therefore these expectations can be disappointed. It is doubtful whether this is also true for robots given the current state of technology. While corporations already have a marked influence on our everyday lives and on social interactions and therefore have become agents in the social system, this is not yet the case for robots. On the other hand, it is far from unthinkable that this might change in the future. A recent British study shows that humans avoid lying to humanoid robots to avoid 'hurting their feelings'[94] and therefore shows that there are tendencies that could lead in the aforementioned direction.

Some authors have introduced the concept of an 'electronic personhood' or 'e-person' in the discussion on the personhood of robots, primarily in the context of tort law liability.[95] This would conceptualize robots as a *sui generis* target for claims, thus providing robots with the status of a legal subject using a 'legal trick'.[96] This may be an original and adequate idea in the context of civil law. However, it will hardly work for criminal law purposes.[97] As will be explained more

[93] Jakobs, *Staatliche Strafe: Bedeutung und Zweck* (2004), pp. 40–41 [hereinafter Jakobs, Staatliche Strafe].

[94] *See on the experiment with the robot* 'Bert': Hamacher, Bianchi-Berthouze, Pipe and Eder, 'Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction' (2016); available at: https://arxiv.org/pdf/1605.08817 (last visited 01 November 2017).

[95] Beck, 'Über Sinn und Unsinn von Statusfragen – zu Vor- und Nachteilen der Einführung einer elektronischen Person', in Hilgendorf and Günther (eds), *Robotik und Gesetzgebung* (2013); Gruber, 'Rechtssubjekte und Teilrechtssubjekte des elektronischen Geschäftsverkehrs', in Beck (ed), *Jenseits von Mensch und Maschine* (2012), p. 150.

[96] Müller, *supra* note 5, 604.

[97] Although Beck, Intelligent agents and criminal law, *supra* note 4, 141–142, suggests the introduction of a new legal status 'electronic personhood' also for criminal law following a 'non-similarity approach' in not trying to apply the classic

precisely in the following sections, the concepts of capacity, personhood, and therefore also of criminal responsibility are shaped by the function of punishment and of criminal law. In contrast to civil law, this function is not to secure payment of damages. Punishment can only function as a symbol of stable expectations if the entity being punished is actually attributed social personhood. To construct robots as 'e-persons' could at most be the consequence of such an attribution.[98] However, it cannot be the cause or the basis for granting robots the status of personhood and does not make the specific requirement of placing blame individually obsolete.

To highlight this important point again, whether a person counts as a person in the law is established in a manner that is both generalized and normative.[99] The question whether this requires consciousness or the capacity for self-reflection is also normative in nature. Given the current state of technology, we would certainly not declare a robot guilty. We would not recognize the robot as perpetrator, because we do not recognize it as a person equal to us and because we do not attribute the necessary capacities to it.[100] We only punish offenders if we deem them sufficiently competent in the social world to question norms and to disappoint expectations.[101] Thus, attribution of criminal responsibility, just as the attribution of personhood, happens through social interaction. As a social construct, the attribution of culpability is not at all arbitrary. As will be explained in more detail in the following section, this attribution depends on the function of the system of criminal law in society. It is

---

Footnote 97 continued
foundations of individual criminal responsibility to this 'e-persons' but rather accepting the differences. From a German perspective where criminal law in fact only applies for human beings, this workaround may make sense. For legal orders having already introduced criminal liability of corporate entities, however, this approach seems unnecessary as criminal law theory has to deal with applying discussed concepts to non-human agents anyway.

[98] See Sect. 4.2 below on the discussion regarding possible punishments as consequence of criminal responsibility of robots.

[99] Jakobs, Schuldprinzip, *supra* note 65, 29.

[100] According to *Jakobs*, the capacity to be attributed personhood means to be defined as an equal. He thus follows *Hegel*'s tradition. If equality is lacking, one forgoes stabilization of the norm and instead reacts to the disappointment of a cognitive expectation; Jakobs, *Strafrecht Allgemeiner Teil* (2nd ed, 1993), p. 496 [hereinafter Jakobs, Strafrecht AT]. On the different reactions to the disappointment of cognitive and of normative expectations, see in general id., 6 et seq.

[101] Jakobs, Schuldprinzip, *supra* note 65, 27.

only in the context of the purpose of criminal law that personhood in criminal law and the resulting criminal responsibility acquire meaning. Hence, it is only in the context of the purpose of punishment that the question of criminal responsibility of robots can be answered.

IV RESPONSIBILITY OF ROBOTS: CRIMINAL RESPONSIBILITY AS A SOCIAL CONSTRUCT

4.1 *Criminal Responsibility in the Context of Theories of Punishment*

The debate on a potential criminal responsibility of robots necessarily takes us back to fundamental questions of criminal law theory, namely questions regarding the reason why criminal law has differentiated itself as a subsystem within society, regarding the purposes of this subsystem, and, in this context, regarding the question why there is punishment at all.[102] In the following sections, we will assume that the function of law in general is to stabilize normative expectations.[103] Law does not secure law-abiding behavior, but it does secure the expectation of it. Law thus secures the counterfactual persistence of normative expectations and creates the possibility of stable expectations in the face of an uncertain future that contains foreseeable unavoidable disappointments.[104] Law, as the 'immune system'[105] of society, thus enables us to live with unmet expectations.[106] This is a necessary condition for behavioral coordination and social order.

Criminal law as a subsystem of the legal system follows this function of stabilizing norms. However, like other areas of law,

[102] This topic is so broad that it cannot be adequately summarized here. The authors will therefore concentrate in the following sections on a short description borrowed from the approach, according to which the primary function of law is the stabilization of norms. This approach has many proponents in sociology of law (not only, but mostly in the tradition of systems theory). For a detailed explanation of the function and of the differentiation of the law, *see, inter alia,* Luhmann, Law as a Social System, *supra* note 15; Luhmann, *A Sociological Theory of Law* ([Albrow (ed), King-Utz and Albrow (trans)], 2nd ed, 1972/2014).

[103] Following, among others, Luhmann, Law as a Social System, *supra* note 15, 142–172. The sociological question of the function of law consists mainly in the question of what kind of problem is being solved via differentiation of the law as a system.

[104] id., 164.

[105] id., 48; Luhmann, Social Systems, *supra* note 69, 374; Dieckmann, *Schlüsselbegriffe der Systemtheorie* (2006), p. 259.

[106] Luhmann, *Ausdifferenzierung des Rechts* (1981), p. 84.

criminal law has a 'distinctive role to play in the social world'[107] and therefore must be differentiated from other subsystems of the law. In general, a crime conflicts with the expectations of the mutual validity of norms, according to which one does not have to expect behavior that is contrary to criminal law.[108] The specific characteristic of criminal law is that it allows for punishment via attribution of individual responsibility and that therefore the norm can be maintained in the face of conflict.[109] By withstanding conflicts in the form of a sanction, criminal law contributes to maintaining the identity of society.[110] It thus serves to create motivation[111] and to practice law-abiding behavior and recognition of norms[112] in society. In case of disappointments of these expectations, the symbolic content of the sanction imposed by criminal law protects the expectation that future behavior will comply with criminal law.[113] The primary purpose of punishment thus is to stabilize norms.[114] Or in the words of Kleinfeld, who introduces similar approaches as 'reconstructivism' or 'normative reconstruction': 'Where wrongdoing tears the social fabric, it is criminal law's task to restitch it'.[115]

The fulfilment of this function of criminal law requires imputing behavior to a person in the social system. This attribution to a person fulfils the task of determining the case in which the behavior of a person has destabilized the norm to an extent that requires an affir-

[107] Kleinfeld, 'Reconstructivism: The Place of Criminal Law in Ethical Life', *Harvard Law Review* 129(6) (2016), 1486.

[108] Jakobs, Strafrecht AT, *supra* note 100, 7.

[109] Jakobs, Schuld und Prävention, *supra* note 82, 12.

[110] Jakobs, ZStW 1995, *supra* note 70, 844.

[111] Kargl, *Die Funktion des Strafrechts in rechtstheoretischer Sicht* (1995), p. 23 *following* Welzel, *Das Deutsche Strafrecht* (11th ed, 1969), p. 3.

[112] Jakobs, Strafrecht AT, *supra* note 100, 13–14; Jakobs, Schuld und Prävention, *supra* note 82, 10–11 and 32–33.

[113] Kargl, *supra* note 111, 36.

[114] To define the function of criminal law in this way does not mean that the system of criminal law does not perform other tasks, ie desirable side-effects, as, for example, in the case of effects of special deterrence. On the distinction between function and other tasks of the (criminal) legal system in general, see Luhmann, Law as a Social System, *supra* note 15, 167–168; Bleckmann, *supra* note 58, 64; Jakobs, Staatliche Strafe, *supra* note 93, 36 et seq.

[115] Kleinfeld, *supra* note 107, 1486.

mation or reinforcement of the norm as a consequence.[116] The function of the attribution of guilt also belongs to this context and therefore has to be analyzed with the function of the system in mind. This is also the thrust of diverse approaches that advocate for a functional restructuring of criminal law and thus for direct recourse to the function of the system. Despite the differences among the different concepts of such a sociologically oriented legal doctrine, the respective proponents[117] share the assumption that criminal law has to be consistent with its aim of 'positive general prevention'[118] and that legal doctrine has to be guided by this aim.[119]

The latter, however, is not a necessary corollary of a functional approach to criminal law. The statement that the system of criminal law always serves a specific function is a statement of the sociology of law. The claim that this has to be reflected in legal doctrine (or that legal doctrine necessarily reflects this fact) is of a different nature. However, if one adopts the functional approach, it is evident that the concept of criminal responsibility acquires meaning only in light of the function of criminal law (ie positive general prevention).[120] It is

[116] Jakobs, *System der strafrechtlichen Zurechnung* (2012), pp. 15–16 [hereinafter Jakobs, System].

[117] *Roxin* (*see, eg, Kriminalpolitik und Strafrechtssystem* (2nd ed, 1973)) has taken first steps in this direction and *Jakobs* (*see, eg,* Schuld und Prävention, *supra* note 82) has developed this approach more consistently.

[118] The theory of 'positive general prevention' which has been dominating criminal law theory largely in Germany, can be formulated compatibly with the sociological function of criminal law, introduced as the stabilization of norms. Although there are different varieties of the theory, its basic features are discerned as follows: Positive general prevention is 'general' because it is distinguished from special prevention, ie the use of punishment to prevent crimes of the particular offender subject to punishment rather than by others. Instead, its main focus lies on the whole population. Furthermore, it is 'positive' and 'preventive' because it aims at preventing crime not by deterrence but by reinforcing law-abidingness. *See* Dubber and Hörnle, *Criminal Law: A Comparative Approach* (2014), pp. 18–20. Although this approach has not received the same attention in English literature, there are a variety of related theories of the criminal law which could be brought together under this label and its general school of thought reaching back to *Hegel* or *Durkheim*. *See on this* 'set of theories' Kleinfeld, *supra* note 107.

[119] Geisler, *supra* note 53, 113.

[120] According to *Jakobs,* this understanding of criminal responsibility is not just an outline for the future. It is not a project or a normative claim. It is an interpretation of the current situation. He states that already today, the concept of criminal responsibility is shaped by the function of criminal law. *See* Jakobs, Schuld und Prävention, *supra* note 82, 32; Jakobs, Strafrecht AT, *supra* note 100, 480–481.

apparent, thus, as has already been explained, that the question of what criminal responsibility is, what it is based on, and who is a suitable subject of it, can only be answered in the context of the question why there is criminal law in society and why there is punishment. Without reference to a specific society, the concept of criminal responsibility would always remain indeterminate.[121]

According to Jakobs, the boundaries of the concept of criminal liability cannot be defined according to the opinion of 'good citizens' as addressees of criminal responsibility and of punishment based on responsibility, but instead they have to be defined by asking which boundaries are necessary to maintain trust in the norms.[122] Thus, it is the aim of punishment that shapes the concept of criminal responsibility. Therefore, the concept of criminal responsibility is neither sacrosanct nor unchangeable and, as a consequence, criminal responsibility of machines is not excluded from the beginning.[123]

4.2 *Criminal Responsibility as Attribution in the System of Criminal Law*

The question of what criminal responsibility is and of who can be criminally responsible depends on society. The same applies to the question of what constitutes wrongdoing in the first place.[124] Hence, the present analysis of the elements of criminal responsibility and their application in the context of the question of whether robots can be criminally responsible, is carried out in a sociologically informed manner. The question is whether there are social mechanisms that allow for such an attribution of criminal responsibility.[125] Due to the

[121] Jakobs, System, *supra* note 116, 63.

[122] Jakobs, Schuld und Prävention, *supra* note 82, 33.

[123] Hilgendorf, *supra* note 12, 119 et seq. *Hilgendorf* rightly states that concepts derive their meaning both from the historical use of language and from definitions. Therefore, these definitions cannot be true or not true. The concepts are thus deliberately given a fixed meaning. Therefore, it is possible to apply concepts such as 'criminal responsibility' to machines. It is striking, however, that such an endeavor seems to be more difficult if concepts are linked to emotions. These statements are not just statements pertaining to the theory of language. Especially concepts that refer to emotional attributions are informed primarily through sociological circumstances, ie in the context of social interaction. Hence, they are not arbitrary, but a part of the social construction of reality (cf Berger and Luckmann, *supra* note 59).

[124] Jakobs, Strafrecht AT, *supra* note 100, 483–484.

[125] We do not refer to a basis pertaining to normative philosophical thinking, but the social sciences. *Similarly* Stübinger, 'Nicht ohne meine "Schuld"!', *Kritische Justiz* 26 (1993), 33.

social relativity of the content and concept of criminal responsibility, it is the society of the future and the way it functions and operates that will determine whether robots will be recognized as persons and whether the 'actions' of robots will have the potential to destabilize norms. In case they do, society will have to develop mechanisms to prevent such a destabilization of norms in order to secure the continued stability of expectations. It is highly possible that the tool society employs for this task will be criminal law and it is also highly possible that, consequently, the idea of a 'guilty robot' will become a part of day to day life.

This scenario requires a process of attribution in the social system. The humanization of robots would have to have gone far enough so that we recognize robots as actors in social interaction and that we do not just have cognitive, but also normative expectations towards them, with the result that their action can be perceived as demonstrations of a lack of compliance with the law.[126] The increasing humanization of robots would further have to lead to the mechanism that, in case of conflicts in the coordination of behavior, we don't just adjust our expectations and 'learn' from our disappointment in the sense that we would not rely on norm-conforming behavior of the 'robot' in the future. Instead, we would have to refuse to 'give up' and continue (normatively) to expect that the robot behaves in accordance with the law. This means that, due to the capacities socially attributed to the robot, there is an expectation that it can adjust its behavior in the light of norms and that it will behave differently in the future or that the insistence on the expectations will reassure the rest of society that they can still insist on their expectations. The essential question in the context of a possible responsibility of robots thus will be whether robots can destabilize norms due to the capacities attributed to them and due to their personhood and if they produce a conflict that requires a reaction of criminal law, without which the norm would first be destabilized and then disappear.

Criminal responsibility is attributed if persons cannot create sufficient distance between them and the injustice of their actions.[127] Responsibility as attribution is a process, a social operation, and not

[126] According to Jakobs, Schuldprinzip, *supra* note 65, 34, criminal responsibility is nothing but such a demonstrated lack of compliance.

[127] Kunz, 'Prävention und gerechte Zurechnung', *ZStW* 98 (1986), 825; Jakobs, Strafrecht AT, *supra* note 100, 479.

a real substrate within the person.[128] To deem someone 'guilty' means nothing else than that we impute a fault, the disappointment of a normative expectation, to a person.[129] This way, society can externalize the conflict, resolve it, and stabilize the norm put into question and thus secure the survival of said norm.[130] We can assume that we involuntarily attribute human traits, motivation, and behavioral patterns especially to robots that can read, process, and react to human emotions. In the future, this may allow us to also attribute them responsibility for their actions.[131] Such a 'humanization' of robots is already starting to take place in everyday language. This can be seen in the fact that the use of anthropological terminology (as in the case of terms like 'movements', 'actions', 'autonomy', and 'thinking') is common in the field of robotics.[132]

The disappointment of legally secured expectations does not occur by way of individualized and subjective misconduct, but by way of an objective misconduct 'breaking character', i.e. not meeting the expectations accompanying the role assigned to someone by the system.[133] The decisive question thus will be what kind of role we attribute to robots rather than questions regarding their actual individual capacities. Already today, we don't generate an individual standard of responsibility for humans and legal persons, but we refer to objective criteria, ie we only refer to the social phenomena that are apparent and therefore relevant for the stabilization of norms. The process of assessing questions of criminal responsibility is about whether a conflict can be resolved by other means besides punishment, whether the disappointment regarding the respective behavior can be explained not by an individual mistake, but by other cir-

[128] Günther, 'Freiheit und Schuld in den Theorien der positiven Generalprävention', in Schünemann, von Hirsch and Jareborg (eds), *Positive Generalprävention* (1998), p. 157.

[129] id., 158.

[130] Jakobs, Schuld und Prävention, *supra* note 82, 13.

[131] Gless and Weigend, *supra* note 25, 565, point to the so-called *Eliza*-effect in computer science. This effect poses the risk of a misinterpretation of the (re)actions of a machine and therefore the risk of 'misunderstandings' in the interaction of humans and intelligent machines. To be fair, it has to be added that misunderstandings of this nature are not rare in the interaction between humans and other humans either.

[132] Müller, 'Von vermenschlichten Maschinen und maschinisierten Menschen', in Brändli, Harasgama, Schister and Tamò (eds), *Mensch und Maschine – Symbiose oder Parasitismus?* (2014); Hilgendorf, *supra* note 12, 120 et seq.

[133] Jakobs, ZStW 1995*, supra* note 70, 861.

cumstances and whether it can thus be socially processed in another manner. Also in this context, the decisive question is whether the aim of criminal law can be met or whether stabilization of the norm in question requires punishment. Individual capacities are only relevant to the extent that they are apparent and explicable. The same would apply in the case of artificial intelligence.

Along the same lines, Gless and Weigend state that criminal responsibility of robots has to be ruled out as long as they have not become 'moral agents'.[134] Criminal responsibility of robots can thus be excluded today, due to sociological fact that they have not yet acquired personhood, but it cannot be excluded for the future. It is not just a functional approach to criminal law doctrine which leads to this conclusion, but it is a consequence of a criminal law that does not and cannot function independently of its aim. In such a future, in which a robot would count as a 'moral agent', criminal responsibility of robots would not be different from criminal responsibility of humans today: a socially attributed liability due to an offence against a rule of criminal law, which exposes a lack of compliance and therefore demands opposition, because the arisen conflict cannot be resolved in another manner. This liability, however, would not be similar to a strict liability. *Mens rea,* a 'guilty mind', would still be a requirement for punishment, even if this is just functionally and objectively attributed, just as it is covertly already the case today.

An argument frequently raised against criminal responsibility of robots is further that punishment of robots that would have the same aim as punishment of humans is hardly imaginable today.[135] According to Wohlers, criminal responsibility of robots would require that the subject of punishment can experience this punishment as a personal evil.[136] This 'punishability' as a personal capacity is supposed to be necessary if criminal responsibility is to make sense at all.[137] Leaving aside the fact that also corporate entities would hardly be 'punishable' in this sense and that therefore the punishment of

[134] Gless and Weigend, *supra* note 25, 589.

[135] *So* Gless and Weigend, *supra* note 25, 578.

[136] Wohlers, *supra* note 16, 123–124. Besides, it cannot be excluded that robots could indeed experience something as a 'personal evil'. Thus, German scientists are currently working on teaching robots how to feel pain and how to react to it. *See* Kuehn and Haddadin, 'An Artificial Robot Nervous System To Teach Robots How to Feel Pain And Reflexively React To Potentially Damaging Contacts', *IEEE Robotics and Automation Letters* 2(1) (2016).

[137] Gless and Weigend, *supra* note 25, 577–579.

such legal persons, which is practiced in many legal orders, would be pointless, it is highly questionable whether this 'punishability' should really be a requirement for criminal responsibility, given that it can be presupposed that punishment is mainly constituted by its symbolic force as a reaction to the disappointment of expectations and not by its actual effects on the punished subject. If it were the case that criminal responsibility only makes sense if the punishment actually has an impact on the offender, our current legal system would already be paralyzed by unresolvable empirical debates, since the potential deterrent effect of punishment on an individual offender is a highly controversial topic in criminology.[138] In any case, criminal responsibility as such can hardly depend on this question.

However, it is of course true that it would constitute a problem to determine the punishment as a legal consequence in case one would someday attribute criminal responsibility to robots.[139] A fine could be imposed using the aforementioned construct of an 'e-person' in an analogy to civil law. Contrary to civil law, however, the aim of criminal law is not to compensate for damages caused, which is why it has to be presupposed that the payment of the respective sum would have to have certain 'consequences' for the robot. This is of course unthinkable today, but it cannot be excluded that in the future robots with artificial intelligence will be able to 'earn' and therefore also 'lose' money. Similarly, it is possible to think of scenarios in which there can be found analogies to incarceration or to other criminal sanctions. A 'reprogramming' or the infliction of an 'evil', which would have consequences for the self-learning system is absolutely possible.[140] Of course, these ideas for punishing robots may seem like science fiction and a little absurd today. However, they mainly show the variability and relativity of our concepts of 'guilt', 'responsibility', and also 'punishment'.

[138] See especially the debate held in the US in the 1970 s characterized by the catchphrase 'nothing works', although this pessimism has been relativized significantly since. *On these developments, see, eg,* Cullen, 'Rehabilitation: Beyond Nothing Works', *Crime and Justice* 42 (2013); Killias, Kuhn and Aebi, *Grundriss der Kriminologie* (2nd ed, 2011), pp. 424–464.

[139] *See* Hilgendorf, *supra* note 12, 127.

[140] id., 130–131.

V SYNTHESIS: ROBOTICS AS A FURTHER TEST FOR THE THEORY OF CRIMINAL RESPONSIBILITY

It follows from the above that criminal responsibility of robots is possible if it is in accordance with the system, ie with the function of criminal law, if it is useful and necessary in the context of the stabilization of norms. This is the case only if robots have the necessary requirements of personhood and if they are thus attributed the capacities that are inherent in that concept. In this context, it has to be stressed again that these characteristics are socially attributed rather than a biophysical fact. A possible attribution of criminal responsibility to robots is thus a process subject to the function of the system. If these requirements are met, robots can therefore be subject to criminal responsibility. A form of criminal responsibility of robots that would not be in accordance with the function of criminal law would not be adequate. This draws attention to the fact that already today, the theory of criminal responsibility rests on the basis of its systemic function. From a sociological point of view, it does not matter whether this is openly manifested or if it is concealed. In any case, criminal law theory should not neglect this fact.

As a consequence, there are three main alternative routes, which criminal law theory could take in the context of 'the law of robotics'. The first would be to insist – at least in theory – on the traditional approach to criminal responsibility, which rests on freedom of choice, on the ideal of the autonomous human and on the exclusive application of the concept of personhood to humans and to defend this idealistic concept against all attack and even in the face of increasing instability. In this case, robots can never be made responsible for anything, regardless of any kind of technological progress and of the problems that this might cause. The consequence would be that in the context of robotic one would have to concentrate on the programmer or operator of the robot for questions of criminal responsibility.

The second route is the one which today seems like the most realistic one: Deliberately or undeliberately, the concept of criminal responsibility is applied in a functional manner in the sense that the concept of criminal responsibility expresses a reproach for a socially visible lack of compliance, which has the potential to destabilize norms. In its application, this concept of criminal responsibility is hardly based on recognizable human free will or freedom of choice, or at least it does so only in attributing these capacities to persons. A

functional application of this concept rather focuses on the actual purpose of criminal law. In the context of such a functional understanding of criminal law and criminal responsibility, which is free from ontological or essentialist[141] concepts of freedom or personhood, a guilty robot would indeed be a possible construct, if the aforementioned developments take place and robots would actually be experienced as 'equals' in the sense that they are constituted as addressees of normative expectations in social interaction like humans or corporate entities are today. At the latest, the last proponents of an ontological concept of criminal law devoted to substantive truth would have to recognize that criminal law and its categories are already regarded as functional categories in society.

A third route would consist in taking the opportunity of the discussions on the concept of criminal law that are caused by technological advances to overcome and to rethink the traditional approach to the concept of criminal responsibility. Such a complete renunciation of the accusations contained in the traditional concept of blameworthiness and its claimed orientation towards 'free will' or 'humaneness' – as proposed, *inter alios*, by Hörnle – and an ensuing adjustment of criminal law concepts seems increasingly possible.[142] This would not have to and should not result in renouncing criminal law reactions or in a criminal law based solely on external acts and their effects.[143] Neither would it lead to a spread of strict liability. Quite to the contrary, one would have to look for adequate functional equivalents, which are free from concepts and ideas that continue to destabilize the currently predominant understanding of criminal responsibility, but which still guarantee that the aim of criminal law, ie the stabilization of norms, can be met. This last route is certainly the most difficult one, but also the most promising in the long run. Revealing the functioning of the social mechanisms of placing blame more manifestly, would therefore not only pave the way for dealing with new challenges regarding the increased relevance of robotics but furthermore also allow to orient criminal legal doc-

[141] According to Hilgendorf, *supra* note 12, 121, the idea that terms and concepts have a fixed and unchangeable meaning can be characterized as 'essentialism'. Essentialism, he states, is based on incorrect presumptions regarding the philosophy of language. The position of the authors of the present paper is that this problem does of course touch on questions of the philosophy of language, but this is not the basis of the primary problem.

[142] Hörnle, *supra* note 72.

[143] id., 10.

trine towards a more ideal manner of fulfilling its function within the stabilization of norms by means of punishment.

In any case, the new debate on 'responsibility of robots' reignites the debate on the concept of criminal responsibility and it does so introducing a completely new perspective. This new momentum may and should be used to review, adjust, or even replace common and traditional concepts. The relevant questions regarding the theory and doctrine of criminal responsibility remain the same in the context of the law of robotics: What should criminal law achieve in the first place and what is the function of criminal responsibility in light of the task of criminal law? If, after the technological advancements in neuroscience, robots like Nao and Pepper have now given rise to attempts at re-examining these questions, this can only consist in an opportunity for criminal law theory.

ACKNOWLEDGEMENTS

This article is based on the authors' contribution already published in a German Journal, although it has been further reviewed and adjusted. See Simmler and Markwalder, Roboter in der Verantwortung—Zur Neuauflage der Debatte um den funktionalen Schuldbegriff, *Zeitschrift für die gesamte Strafrechtswissenschaft (ZStW)* 129(1) (2017), pp.20 et seq. The authors would like to thank Sué González Hauck at this point for her great help in translating and refining this article.