Chapter 1: The Complexity of Liability for Crimes Involving Autonomous Systems Driven by Artificial Intelligence

A. Legal Challenges

Humans have utilised various technological tools for millennia, each contributing significantly to the development of civilisation. However, in parallel, it has become necessary to balance the risks posed by new technologies with their advantages for society. For example, although steam engines introduced certain risks during the onset of *Industry 1.0*, these technologies were not prohibited. Instead, their use was regulated through licensing requirements, and lawmakers implemented measures to mitigate their risks. This approach aimed to reduce potential hazards to a socially acceptable level while allowing society to benefit significantly from the technology[19]. Similarly, despite all the opportunities it provides, digitalisation also facilitates and amplifies the infringement of legal interests[20]. As technology evolves rapidly, it transforms human habits, leading to changes in moral values and legal norms over time[21]. On the other hand, autonomous systems push the boundaries of traditional criminal law to its limits[22].

As with many other technologies, the dual-use nature of AI (its potential for both beneficial and harmful applications) has attracted growing attention as the body of literature on the subject expands across both technical and social sciences[23]. Therefore, the challenges it poses must be analysed by examining their underlying causes and resolved through solutions that balance societal benefits against potential risks. The integration of AI-driven autonomous systems into the causal chain represents a significant shift in the nature of human-machine interaction. While their role may not constitute a 'decision' or 'action' in the traditional sense, these systems are becoming an integral part of human activities. As a result, human control over the causal chain reduces, and the process becomes less comprehensible[24].

19 HILGENDORF, Zivil- und strafrechtliche Haftung, 2019, p. 438.
20 BECK, Die Diffusion, 2020, p. 44.
21 HILGENDORF, Digitalisierung, Virtualisierung und das Recht, 2020, p. 408.
22 GLESS/SILVERMAN/WEIGEND, If Robots Cause Harm, 2016, p. 435.
23 BRUNDAGE, et al., The Malicious Use, 2018, p. 16.
24 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 208.

This raises a critical question: does this involvement disrupt or obscure the attributional connection? When decisions are so deeply interconnected, linking the outcome directly to the human actor becomes challenging[25].

The advancement of AI and the associated debates mainly stem from its autonomous features, giving rise to "*autonomy risk*"[26], the unpredictable behaviour of self-learning systems. This results in *ex ante* challenges, in addition to AI's *ex post* issues related to explainability[27]. Furthermore, AI presents *interaction* and *network risks*. Interaction risk involves the complex interplay between humans and machines within socio-technical systems, while network risk emerges when multiple computer systems collectively contribute to harmful outcomes or trigger widespread failures across interconnected devices[28]. These risks, including vulnerabilities against potential cyberattacks, become particularly concerning due to system interconnectivity and the wide use of IoT devices[29].

It is also crucial to determine whether the harmful outcomes caused by AI-driven autonomous products stem from a design flaw, a "self-learning" capability (which may itself be considered a design flaw under certain conditions), or a production failure[30]. This study focuses specifically on harmful outcomes (criminal offences) arising from autonomy risk, and therefore potential design flaws. In cases of production failure, particularly those examined under the 'problem of many hands', AI does not present unique characteristics and can be addressed through conventional product liability framework.

Insufficient understanding of the risks and limited control over AI systems hinder the effectiveness of human defensive measures against potential harm[31]. Given the diverse use of AI systems across various fields, along with the range and scale of associated risks, a "one size fits all" approach is impractical for determining liability. In some cases, establishing criminal norms may be meaningful to ensure deterrence, while in others, non-crim-

25 BECK, Die Diffusion, 2020, p. 45.
26 ZECH, Zivilrechtliche Haftung, 2016, p. 170, 175.
27 ZECH, Risiken Digitaler Systeme, 2020, pp. 44-48; ZECH, Zivilrechtliche Haftung, 2016, p. 175.
For some, opacity is a more prominent issue than autonomy. See: IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 429.
28 FATEH-MOGHADAM, Innovationsverantwortung, 2020, p. 875 f.
29 WACHTER, Normative Challenges, 2018, p. 439, 448.
30 BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 15 f.
31 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 212.

inal enforcement may be sufficient[32]. Indeed, criminal law cannot fully protect all legal interests. However, as AI-driven autonomous systems become more widespread, they are likely to become the main source of harmful outcomes. To address this, developers could design the learning capacities of self-learning systems from the outset to avoid acquiring behaviours that may harm humans[33]. All of these challenges are addressed in the relevant sections of the study.

B. AI-Driven Autonomous Systems in Daily Life: A New Normal

Autonomous systems driven by AI are being applied across various fields to enhance efficiency and innovation. These specific applications of AI are transforming daily life by providing advanced solutions to complex challenges. They undertake specific tasks and, in some instances, autonomously manage their completion along with associated sub-goals. In healthcare, AI algorithms assist doctors by analysing medical images for early detection of diseases like cancer and predicting patient outcomes. Self-driving vehicles use AI to navigate roads safely, aiming to mitigate traffic accidents and improve transportation efficiency. In industry, AI-driven robots perform complex assembly tasks, and predictive maintenance systems forecast equipment failures to minimise downtime. At home, AI enables smart assistants like voice-controlled devices to manage lighting, security systems, thermostats, etc. based on user preferences. These systems are particularly invaluable in certain domains, where they effectively replace human activities or operate in areas where human involvement is not feasible. For instance, they can operate in hostile environments such as underwater, underground, or in space[34].

Advancements in hardware and software, particularly in adaptability and learning, currently enable robots to operate in increasingly complex settings. In contrast to traditional industrial robots fixed within safeguarded places; modern robots are mobile, with some being deployed in open-road traffic[35]. Today, the most common autonomous systems with physical mo-

32 Singapore, Report on Criminal Liability, 2021, p. 2, [para. 7].
33 HILGENDORF, Autonome Systeme, 2018, p. 110.
34 SCHULZ, Verantwortlichkeit, 2015, pp. 43 f., 56-71; LIN/ABNEY/BEKEY, Robot Ethics, 2011, p. 944 f.; DEVILLÉ/SERGEYSSELS/MIDDAG, Basic Concepts of AI, 2021, pp. 14-20.
35 ZECH, Risiken Digitaler Systeme, 2020, p. 23.

bility are self-driving vehicles and robotic vacuum cleaners. In the near future, it remains to be seen whether humanoid robots, designed to perform physical household tasks, will become widespread. Although self-driving vehicles are often compared to airplane autopilots -which can computerise most of a flight under human pilot supervision- the analogy overlooks critical differences such as unpredictable road obstacles and the controlled, obstacle-free nature of airspace, making full vehicle autonomy significantly more challenging[36].

Even in seemingly harmless applications, these systems pose risks to legal interests protected by criminal norms. Some of these incidents would constitute criminal offences if caused by a human actor. For example, in a notable incident, a South Korean woman's hair became entangled in a robot vacuum cleaner while she was sleeping, which led to injury[37]. Similarly, numerous fatal, injury-causing, and property-damaging traffic accidents have occurred involving vehicles with varying degrees of autonomy[38]. Moreover, the issue of attributing criminal liability to the individuals behind these systems arises not only for physical devices but also for software-based AI systems. For example, in an experimental project, a software bot was programmed to make random purchases by spending $100 in *Bitcoin* per week on a darknet market, which resulted in the acquisition of various goods, including illegal drugs[39]. Numerous real-life examples similar to those mentioned here are discussed throughout this study under relevant topics. For instance, given the relatively recent widespread adoption of these systems, the legal expectation for programmers to foresee certain

36 KLEINSCHMIDT/WAGNER, Technik autonomer Fahrzeuge, 2020, p. 16 Rn. 16; WIGGER, Automatisiertes Fahren und Strafrecht, 2020, p. 92.

37 McCURRY Justin, "South Korean woman's hair 'eaten' by robot vacuum cleaner as she slept", 09.02.2015, https://www.theguardian.com/world/2015/feb/09/south-korean-womans-hair-eaten-by-robot-vacuum-cleaner-as-she-slept. (accessed on 01.08.2025).

38 "Tokyo 2020: Toyota restarts driverless vehicles after accident", 31.08.2021, https://www.bbc.com/news/business-58390290; KLEIN Alice, "Tesla driver dies in first fatal autonomous car crash in US", 01.07.2016, https://www.newscientist.com/article/2095740-tesla-driver-dies-in-first-fatal-autonomous-car-crash-in-us/.(accessed on 01.08.2025).
In fact, Tesla, known for its semi-autonomous driving technology, has been associated with numerous accidents, both those reported in the media and those less publicised. For a list compiling some of these incidents, see: https://en.wikipedia.org/wiki/List_of_Tesla_Autopilot_crashes. (accessed on 01.08.2025).

39 POWER Mike, "What happens when a software bot goes on a darknet shopping spree?", 05.12.2014, https://www.theguardian.com/technology/2014/dec/05/software-bot-darknet-shopping-spree-random-shopper. (accessed on 01.08.2025).

outcomes -such as the need to account for people sleeping on the ground- differs significantly between 2014 and 2024. Such matters are examined in relation to defining the scope of the duty of care in cases of negligence.

C. Conceptual Framework

This section offers a brief overview of artificial intelligence and related concepts. Although a detailed technical examination of AI technologies is valuable, the primary aim of this study is to explore the legal implications of criminal liability in offences caused by autonomous systems functioning without human intervention in specific circumstances. Accordingly, the descriptive section is kept concise to establish a foundational understanding supporting this study's legal analysis. Key terminology and core principles of AI will be outlined to ensure clarity and consistency throughout the following discussions.

1. Automation - Autonomy

Automation refers to machines or systems carrying out tasks automatically based on pre-set instructions, without the ability to adapt. It is the overarching term for the self-operating execution of processes and refers not only to the control of hardware but to data processing as a whole[40]. Autonomy, on the other hand, means systems can make their own "decisions" and adjust to new situations without explicit human guidance. This distinction, which forms the basis of the study, is analysed in detail below[41].

2. The Turing Test

The *Turing Test* (named after *Alan Turing*) was introduced as a method of determining whether a machine can demonstrate intelligent behaviour indistinguishable from a human by replacing the original question, "can machines think?" with the question of whether a machine can successfully

40 ZECH, Risiken Digitaler Systeme, 2020, p. 9.
41 See: Chapter 1, Section E(1): "Ex Ante: Autonomy and Diminishing Human Control".

mimic a human in the imitation game[42]. However, subsequently, even simple chatbots that could not qualify as AI have, despite failing the *Turing Test*, led some individuals to believe that they were conversing with a real person. This phenomenon, known as the *Eliza Effect*, refers to the tendency of people to attribute human-like understanding and empathy to basic computer programmes, despite their lack of genuine comprehension[43]. Although certain applications today have succeeded in passing the *Turing Test*, and they do not exactly function as envisaged in the hypothetical "Chinese room" thought experiment; they still lack true understanding or consciousness[44]. Therefore, it is necessary to approach the question of whether AI will gain consciousness in the future with caution, bearing in mind the *Eliza Effect*. Nonetheless, it is important to recognise that AI's functioning is not magic; but are based on mathematical algorithms, statistical models, and large datasets. While the literature often attributes human-like features such as thinking and learning to AI, these processes do not constitute genuine cognition or learning in the true sense.

3. Bot - Robot

The term 'robot' was first introduced by Czech writer, *Karel Čapek*, in his 1920 play, R.U.R. *(Rossum's Universal Robots)*, but the word was actually coined by his brother, *Josef Čapek*. He derived the term from the Slavic-rooted Czech word *robota*, which historically referred to compulsory, unpaid labour performed by peasants for their feudal lord, also known as *corvée*[45].

The term 'bot' originates from the word 'robot' and is its shortened version. However, over time, its usage on the internet has led to a distinction whereby software-based systems are referred to as 'bots' while systems with

42 TURING Alan M., "Computing Machinery and Intelligence", 1950, p. 433 ff.

43 SIMONE, The Eliza Effect, 2021, p. 50 f.

44 A recent study published by Apple contends that, despite notable improvements on reasoning benchmarks, current Large Reasoning Models (LRMs) fail to exhibit genuine reasoning capabilities or comprehend in a manner akin to human cognition. See: SHOJAEE et al., The Illusion of Thinking, 2025.
However, the study has faced considerable criticism for potential bias, given that Apple had significantly lagged behind in the AI race as of mid-2025.

45 "Czech word "Robot" and Its History", 22.03.2024, https://www.czechology.com/czech-word-robot-is-100-years-old/. (accessed on 01.08.2025).

physical appearance are designated as 'robots'[46]. Hence, the term 'robot' should be understood to refer specifically to embodied systems. Although only a small proportion are equipped with advanced AI software and many remain relatively "dumb" in their functionality[47]; robots are generally conceptualised as artificial systems capable of sensing, processing, and interacting with their environment to some extent[48]. This capability distinguishes robots from traditional machines, which lack this level of autonomous interaction[49]. Hence, in this study, the term 'robot' will denote physically embodied systems that demonstrate autonomous features supported by AI.

In the early phases of robotics, the 'sense-plan-act' architecture was commonly employed to describe a process in which an agent attains rational behaviour through a sequential process: initially perceiving its surroundings using sensors, subsequently formulating inferences and decisions based on the acquired data, and ultimately implementing the determined actions through actuators[50]. Later, this model was modified primarily due to its limitations in real-world applications, where planning takes too long, and execution without real-time sensing can be risky. Hence, various designs (in practice, robotics frequently integrates multiple architectures, as there is no single ideal model suitable for all situations) such as *subsumption architecture, behaviour-based robotics, layered control* have been implemented[51].

In terms of the subject under review, it must be emphasised that the category of an entity as a 'bot' or 'robot' is irrelevant when assessing involvement in a criminal offence[52]. The examination encompasses not only physical robots but also virtual systems capable of making autonomous "decisions" independent of physical sensory inputs[53].

4. Artificial Intelligence

Although research on synthetic, human-made intelligence has roots extending back many decades, and neural networks have existed since the 1940s,

46 CALO, Robotics and the Lessons, 2015, p. 534.
47 RYAN, In AI We Trust, 2020, p. 2751.
48 CALO, Robotics and the Lessons, 2015, p. 531.
49 CALO, Robots in American Law, 2016, p. 6; AKSOY, Yapay Zekalı, 2021, p. 13.
50 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, pp. 162-163.
51 KORTENKAMP/SIMMONS, Robotic Systems, 2008, p. 189 ff.
52 HU, Robot Criminals, 2019, p. 495.
53 MARKWALDER/SIMMLER, Roboterstrafrecht, 2017, p. 173.

the most significant advancements have emerged in recent years, largely due to increases in computational power and the availability of big data. These developments have enabled the creation of neural networks that consist of multiple layers, rather than being limited to a simple, shallow architecture[54].

Efforts to define artificial intelligence and address the question of legal responsibility associated with it are not new[55]. The earlier examples of these systems were in fact not artificial intelligence, but "expert systems", due to the lack of autonomous conduct[56]. One challenge in defining AI arises from the fact that it is not a single, discrete technological concept but rather an umbrella term encompassing a range of technologies[57]. AI exists in multiple forms, each possessing distinct cognitive-, emotional-, and social like competencies, which complicates the task of establishing a precise and comprehensive definition[58].

The European Union's AI Regulation, the most comprehensive legal framework on artificial intelligence to date, has introduced a definition of the term. However, it has been also criticised for having an overly broad definition of AI, encompassing nearly all types of software, while, at the same time, not distinguishing these systems depending on their level of autonomy[59]. Moreover, this broad approach may lead to regulatory overlap, wherein the same concept -such as 'computer program' or 'artificial intelligence'- is governed by multiple, potentially conflicting legal norms. However, this study does not aim to establish a definition of AI. Therefore, while acknowledging the validity of these criticisms, the definition provided

54 LEE, Artificial Intelligence, 2020, p. 35; DEVILLÉ/SERGEYSSELS/MIDDAG, Basic Concepts of AI, 2021, p. 9.

55 For example: LEHMAN-WILZIG, Frankenstein Unbound, 1981, p. 442.

56 KAPLAN, Artificial Intelligence, 2022, p. 10.

57 GASSER/ALMEIDA, A Layered Model, 2017, p. 59.
This is one of the reasons why this study emphasises autonomy rather than artificial intelligence.
Capitalising on the hype and market share surrounding AI and the ambiguity surrounding its scope, there has been a growing tendency to label as AI various systems that, either do not genuinely employ AI or rely on only a minimal degree of machine learning. *AI-washing* refers to marketing efforts that misleadingly exaggerate a product's use of AI to make it appear more advanced or successful than it actually is, often by falsely claiming AI capabilities or overstating the technology's potential. See: BABUCKE/KRONER, Künstliche Intelligenz, 2024, p. 175.

58 KAPLAN, Artificial Intelligence, 2022, p. 7.

59 EBERS, Truly Risk-Based, 2024, p. 18; BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 3.

in the AI Regulation, at Article 3 (1), will serve as a guiding framework: "*'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*"[60]. For the purposes of this study, it should be emphasised that autonomy and adaptiveness appear as key characteristics of AI.

5. Machine Learning

Machine Learning (ML) is a subfield of AI, focused on developing and deploying algorithms and statistical models that enable computer systems to perform specific tasks effectively without rule-based programming[61]. Instead of following direct and explicit instructions, these systems identify patterns within large datasets, allowing them to make predictions or decisions autonomously. In the typical ML process (supervised), an algorithm is trained on numerous pre-labelled samples (such as images of handwritten digits) to learn and extract distinguishing features relevant to the given task. This model can then be applied to new, previously unseen handwritten characters to assign them to the most appropriate digit. Essentially, ML involves the creation of a model that abstracts reality and generalises from sample data so that it can be used on new data[62].

Machine Learning includes a range of techniques tailored to handle diverse data types and solve various tasks. The main ML techniques are supervised learning, unsupervised learning and reinforcement learning. In *supervised learning*, the algorithm is trained on labelled data and each sample in the training set comes with an associated correct output. The model

60 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (*Artificial Intelligence Regulation*), 12.07.2024, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689. (accessed on 01.08.2025).

61 DEVILLÉ/SERGEYSSELS/MIDDAG, Basic Concepts of AI, 2021, p. 6.

62 LEE, Artificial Intelligence, 2020, p. 41 f.; DÖBEL Inga et al., "Maschinelles Lernen Kompetenzen, Anwendungen und Forschungsbedarf", Fraunhofer-Gesellschaft, 29.03.2018, https://www.bigdata-ai.fraunhofer.de/de/publikationen/ml-studie.html, p. 13 f. (accessed on 01.08.2025).

learns the relationship between inputs and outputs and can predict outputs for new, unlabelled and unseen data. In *unsupervised learning*, the model is trained on data without explicit labels, and the model is expected to independently discover patterns and structures on its own. In *reinforcement learning*, algorithms are not explicitly instructed on how to perform specific tasks. Instead, a reward system is implemented, in which rewards serve as positive or negative feedback guiding the model towards or away from the goal[63].

Deep learning represents a subset of machine learning, employing artificial neural networks comprising multiple layers (deep neural networks) to model complex patterns in large datasets. It is particularly effective for tasks involving image, speech, and natural language processing[64]. Taking advantage of big data and computational resources, deep learning can identify features and transformations without the need for human intervention. User-friendly software and efficient parallel hardware have accelerated deep learning research, simplifying the testing and exploration of various network architectures[65]. Nonetheless, deep learning has not entirely replaced traditional programming approaches. Hybrid methods that combine traditional algorithms with deep learning techniques can achieve high levels of success[66].

Despite decades of research, these models are still in their infancy, and the associated risks are only now beginning to emerge. Their vulnerabilities are far from being fully understood or identifiable, yet nearly all such systems exhibit some weaknesses[67]. For instance, for large language models (LLM) like ChatGPT, security measures-guardrails and limitations set by the developers can be bypassed using the DAN *(Do Anything Now)* mode, which could be considered a form of prompt injection[68]. Indeed, for example, due to the technique deep neural networks (DNN) function, it is

63 DEVILLÉ/SERGEYSSELS/MIDDAG, Basic Concepts of AI, 2021, p. 6 f.; SUN, Connectionism, 2014, p. 111 f.; EVTIMOV, et al., Is Tricking a Robot Hacking, 2019, p. 894-895.

64 LÄMMEL/CLEVE, Künstliche Intelligenz, 2023, p. 197 ff.

65 ALPAYDIN, Machine Learning, 2021, p. 129 f.

66 MAHONY, et al., Deep Learning, 2020, p. 141.

67 PAPERNOT, et al., Towards the Science of Security, 2016, p. 15.

68 KATOĞLU/ALTUNKAŞ/KIZILIRMAK, Yapay Zekâ, 2025, *passim*.
For instance, it is possible to manipulate ChatGPT through a technique known as prompt injection which could trick the model into disclosing information such as Microsoft Windows activation codes. See: CUTHBERTSON Anthony, "ChatGPT 'grandma exploit' gives users free keys for Windows 11", 19.06.2023, https://www.inde

easy to trick the model with small adjustments. To illustrate, a speed sign of 35km/h can be altered by adding a line to the number '3' to make it look like an '8'; whilst humans will observe the sign to state as 35km/h at first glance, self-driving vehicles on the other hand will perceive it as 85km/h[69], thereby causing the vehicle to accelerate. The concept of robustness, which was initially mentioned in the Ethics Guidelines for Trustworthy AI prepared by the EU's High-Level Expert Group on Artificial Intelligence (HLEG)[70] and also highlighted in the EU's Artificial Intelligence Regulation, focuses on whether a model performs as expected under typical, atypical, irregular, or adversarial conditions[71]. This issue is examined in greater depth below, focusing specifically on the negligent liability of developers and manufacturers.

D. Addressing Liability: Key Actors and Entities

Regarding crimes involving AI-driven autonomous systems, numerous challenges emerge in attributing liability to specific individuals. It is necessary to examine whether those who have contributed to the creation of these systems or interacted with them in operation after deployment can be held accountable, and, if so, how such liability might be structured. The objective of this discussion is to identify and analyse the most likely addressees of liability. Within the scope of this study, the general concept, *person behind the machine* is adopted to encompass individuals who interact with AI-driven autonomous systems in various ways; such as by creating, manufacturing, programming, developing, commanding, manipulating, using or interacting with them in any way. However, to accurately determine liability, the scope of this interaction and the nature of the act must indeed be clearly defined in relation to the specific incident and the application involved.

---

pendent.co.uk/tech/chatgpt-microsoft-windows-11-grandma-exploit-b2360213.html. (accessed on 01.08.2025).

69 McAfee Demonstrates Model Hacking in the Real World, 19.02.2020, https://www.youtube.com/watch?v=4uGV_fRj0UA&t=16s. (accessed on 01.08.2025).

70 High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 08.04.2019, https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1, p. 16 f. (accessed on 01.08.2025).

71 COOPER, et al., Accountability, 2022, p. 865.

The criminal liability associated with the negligence of the person behind the machine can be attributed due to the behaviour in the whole phase of production, usage, research, and development[72]. As the autonomy of AI systems increases, control gradually shifts away from the user. Consequently, incidents become less attributable to the actions of the individual user and liability tends to shift towards the producer[73]. Therefore, it is essential to assess, for each application of AI and incident, who might qualify as the person behind the machine, as well as to evaluate their proximity to the system and the level of control. For example, in the case of LLMs, the developer may exercise a greater degree of control, whereas in the context of a self-driving vehicle, this may be lower. Naturally, varying levels of duty of care apply in each context and sector[74].

It should be noted that this study does not aim to define the scope of responsibility and standard of care for each individual subject (manufacturer, driver, deployer, etc.) according to specific legal frameworks. Instead, it aims to establish a general structure for negligent liability principles, concentrating on the implications of altering control, to encompass a range of AI-driven autonomous systems. Indeed, the duty of care varies significantly across sectors and subjects, necessitating a meticulous analysis to determine the extent of an individual's responsibility in each context. However, such an analysis is directly linked to applicable positive law, which may be amended over time. Hence, a more general framework is sought to be outlined in this study. As will be further discussed under Chapter 4 (Sections: *The Legal Basis of Duty of Care* and *The Feasibility of Defining Permissible Risk Through Standards and Other Norms of Conduct*), once the degree of autonomy, level of control and involvement of the individual behind the machine are determined, identifying the scope of the objective duty of care in line with current legal norms for relevant subjects becomes a straightforward task. These responsibilities can be explored separately in more targeted and narrowly focused studies by analysing specific positive legal norms.

The legal literature offers a range of ideas on the potential identity of the person behind the machine. These primarily involve the programmer,

72 BECK, Intelligent Agents and Criminal Law, 2016, pp. 138-139.

73 HILGENDORF, Automatisiertes Fahren und Recht, 2018, p. 803; BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 12.
See: Chapter 3, Section C(1)(d)(2): "Responsibility Shifting to Manufacturers".

74 VALERIUS, Sorgfaltspflichten, 2017, p. 12 ff.

manufacturer, operator[75], researcher, seller[76], and information provider[77]. A dual distinction is also made between the production and the usage sides. On the production side, key actors in the "prior chain" are involved in manufacturing and introducing these systems to the market, such as programmers, designers, retailers, sellers and distributors. On the usage side, by contrast, are those who operate the robots, primarily involving commercial users and consumers[78].

Producer: The producers are responsible for ensuring the safety of the product, both in terms of its design and its programming, and for providing the interfaces between the product and its operator[79]. Under Section 4 of the German Product Liability Act (*Produkthaftungsgesetz* - ProdHaftG)[80], a 'manufacturer' is defined as any entity that produces the end product, a raw material, or a partial product. Certain duties of care are associated with participation in the manufacturing process. These include responsibilities related to design, fabrication, providing instructions and ongoing product monitoring[81]. For instance, the manufacturer may be held liable for training the system with insufficient data, either in terms of quantity or quality, or for failing to monitor the plausibility of the system's learning progress[82].

Defining the boundaries of producer is particularly essential yet challenging in cases involving complex systems composed of multiple hardware components and software developed by various individuals and entities. Due to the multitude of actors involved in such systems, issues regarding the determining individual criminal liability will be examined under the problem of many hands[83].

Operator: In literature, the term 'operator' functions as an umbrella term encompassing individuals who possess or utilise such systems[84]. Primarily,

75 MARKWALDER/SIMMLER, Roboterstrafrecht, 2017, p. 174 ff.
76 BECK, Die Diffusion, 2020, p. 45; BECK, Selbstfahrende Kraftfahrzeuge, 2020, p. 442 Rn. 14.
77 SCHULZ, Verantwortlichkeit, 2015, pp. 192-196.
78 ZECH, Zivilrechtliche Haftung, 2016, pp. 177-179; GIANNINI/KWIK, Negligence Failures, 2023, p. 58.
79 HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 74; BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 12.
80 Gesetz über die Haftung für fehlerhafte Produkte (ProdHaftG), enacted on 15.12.1989, last amended on 23.11.2022, https://www.gesetze-im-internet.de/prodhaftg/BJNR021980989.html. (accessed on 01.08.2025).
81 HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 73.
82 VALERIUS, Strafrechtliche Grenzen, 2022, p. 123 f.
83 See: Chapter 4, Section D(1): "The Concept of "the Problem of Many Hands"".
84 SEHER, Intelligent agents, 2016, p. 52.

it refers to those who exercise control over the system's operation, including the authority to activate or override its functions. This category specifically includes both owners and users of the system[85]. However, in the EU's AI Regulation, the term operator has been defined as "provider, product manufacturer, deployer, authorised representative, importer or distributor" in Article 3(8) at a later stage. Within this study, the term 'operator' will be used in a manner consistent with its usage in the literature, encompassing 'user' as well.

For systems in which the user preserves greater control, an additional category, named "user in charge" has been proposed. This designation applies to individuals who retain control over semi-autonomous systems or hold the authority to approve specific actions executed by the system. Such users may also bear a duty to oversee the system's operation and to intervene when necessary[86]. While identifying the "user in charge" is relatively straightforward in systems with low levels of autonomy, achieving clarity in more complex systems would be enhanced by definitive legal rules[87]. Regardless of whether they are referred to as a "user-in-charge" or an "operator", it is evident that such individuals are more than merely passive subjects. They are either tasked with supervising AI-driven autonomous systems or have limited control over them. Accordingly, they are expected to be prepared to override the system in the event of a malfunction, thereby balancing the utilisation of the system's benefits against its inherent risks. For instance, in the case of a self-driving car, this role may be fulfilled by the person seated behind the wheel. However, this supervisory role is only effective if genuine control over the system is possible. In many instances, factors such as response time and limited intervention opportunities may make it impractical[88]. In any case, legal expectations on individuals must be realistic[89].

Under certain conditions, the responsibility of operators may be adjusted. For instance, if an individual using an autonomous system has been adequately informed about how the system will function in specific scenarios, including any inherent risks or foreseeable behaviours; or if they possess

85 BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 12; HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 74.
86 Singapore, Report on Criminal Liability, 2021, pp. 23-24, [para. 4.3].
87 *Ibid*, p. 24, [para. 4.4].
88 GIANNINI/KWIK, Negligence Failures, 2023, pp. 56-57.
89 The topic is widely discussed under the Section "control-dilemma". See: Chapter 4, Section C(4)(d): "Control Dilemma".

prior knowledge of the system's potential conducts, it would be unreasonable to attribute the outcome solely to the manufacturer[90]. Moreover, if an operator integrates a self-developed update into the software's control system that significantly impacts its functioning, they may be regarded as a (partial) producer and therefore be subject to certain obligations[91].

One of the most common applications of AI where individuals act as operators is semi-autonomous vehicles. According to German jurisprudence, being regarded as a driver mainly depends on three criteria: control over the vehicle's movement, influence over the driving process, and exercising decision-making authority. As motor vehicles become increasingly automated and approach fully autonomous driving, these criteria begin shifting towards the manufacturer who programmes the vehicle's software and thus assumes control over the vehicle[92]. In a recent decision, the German Federal Court of Justice (BGH) held that an individual who does not operate any of the essential components of the vehicle cannot be considered a driver at the relevant time. Accordingly, considering that a vehicle may have multiple drivers simultaneously, a driving instructor, who does not intervene during a particular instance of a driving lesson is not deemed to be driving the vehicle[93]. From this perspective, it is argued that an individual in an autonomous vehicle should no longer be regarded as a driver if control over the vehicle's essential movement functions is delegated to the autonomous system[94].

It is indeed a widely held opinion in literature that, in context of autonomous driving, humans in the vehicle should not be regarded as driver[95] and, for example, when they sleep, they should only be held liable due to a failure to act when they had to intervene[96]. However, it can be argued

90 ENGLÄNDER, Das selbstfahrende, 2016, p. 387.
91 HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 74.
92 SCHRADER, Haftungsfragen, 2016, p. 245.
93 Federal Court of Justice (BGH), decision of 23.09.2014, Case No. 4 StR 92/14, reported in NZV 2015, p. 145.
94 STAUB, Strafrechtliche Fragen, 2019, p. 394.
95 As an opposing view, a person who activates and uses a highly or fully automated driving function is still considered the vehicle driver even if they are not manually controlling the vehicle during automated operation. See: WIGGER, Automatisiertes Fahren und Strafrecht, 2020, pp. 182-188.
96 BECK, Die Diffusion, 2020, p. 45.
When an automated driving function is used as intended under Section 1(a) of the StVG, the driver is permitted, in accordance with Section 1(b)(1), to disengage from monitoring traffic and controlling the vehicle and may engage in non-driving activities. However, pursuant to Section 1(b)(1) and (2), the driver must stay alert

that defining passengers in autonomous driving as entirely passive, except in exceptional cases, is not always accurate. For instance, a person who gets into their self-driving vehicle to commute to work is the one who initiates / activates and sets the system in motion. Therefore, the initial point of discussion on liability should be whether a legally relevant risk has been created (or increased) by such an action (initiating the system). Hence, only in rare circumstances, such as in smart cities where fully autonomous taxis are widely used and summoned with a single click, is it reasonable to consider passengers being in a completely passive role. Nevertheless, even in such cases, the responsibility and liability of the individual who anticipates the risk yet delegates it to the autonomous system may still be examined[97].

In my view, the time and circumstances of delegation (initiation) of tasks traditionally performed by humans to AI-driven autonomous systems should serve as the starting point of assessment on whether a legally relevant risk has been created. Following this starting point, further analysis concerning liability in negligence and permissible risk can be made. It would be incorrect to categorically exclude individuals from responsibility by classifying them as mere passive bystanders, thereby precluding any liability discussion from the outset. Criminal law, after all, is concerned not with an individual's formal legal classification (driver or not)[98], but with their behaviour and culpability.

These considerations extend beyond autonomous driving and apply broadly to all types of AI-driven autonomous systems. If an opposing view were to be adopted -whereby individuals delegating tasks to such systems and benefiting from their use are not considered as operators simply because they do not directly control the system's essential components- this could lead to problematic outcomes by creating a gap in accountability, with no responsible party identified. Therefore, while acknowledging the importance of control over the essential components of the system, initiating a system known to carry inherent risks should be considered the starting point for evaluating responsibility and liability. Moreover, as the vast majority of systems are likely to function highly autonomously in the future, it could lead to the absence of control-responsibility for the funda-

and prepared to reassume control of the vehicle immediately if necessary. See: SEDLMAIER/KRZIC BOGATAJ, Die Haftung, 2022, p. 2954.

97 For a detailed discussion see: Chapter 4, Section C(5)(b)(3)(d): "Delegating Tasks to AI-Driven Autonomous Systems: An Alternative Approach for Liability".

98 It can only affect the source of duty of care.

mental components of these systems. Delegating their tasks to AI, both individuals and companies benefiting from these systems might thereby evade liability risks.

E. Distinctive Challenges of Crimes Involving AI-Driven Autonomous Systems

Although calculators execute operations much faster than human capability, they are not considered intelligent, as they simply follow predetermined programming and perform tasks in a strictly predictable manner. AI on the other hand, exhibits adaptive and autonomous decision-making capabilities, can "learn" from data, recognise patterns and can solve complex problems[99]. In contrast to automatic systems that merely mechanically substitute human labour (both physical and mental), AI, enables machines to comprehensively and autonomously collaborate with humans throughout the decision making and execution processes[100].

In adaptive systems, human control diminishes, and predictability of the systems' output correspondingly decreases even for the programmer[101]. The inherent unpredictability of AI-driven autonomous systems, as well as the complexity and opacity of these technologies, presents distinct challenges to traditional fault-based liability frameworks[102]. Although these issues are particularly evident in AI-driven autonomous systems, it has also been argued that even conventional computers of the 1990s introduced a degree of separation between an individual's action and their consequences, which can conceal the causal link between them[103].

The unique challenges posed by crimes involving AI-driven autonomous systems can be classified into two main categories: *ex ante* issues, which arise from the diminishing control and inherent unpredictability of these

99 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 159.
However, deep learning has not entirely replaced traditional programming approaches. Hybrid methods that combine traditional algorithms with deep learning techniques have demonstrated significant success. See: MAHONY, et al., Deep Learning, 2020, p. 141.

100 ZHAO, Principle of Criminal Imputation, 2024, p. 6 f.

101 BECK, Die Diffusion, 2020, p. 44; ZECH, Risiken Digitaler Systeme, 2020, p. 35.

102 BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 7.

103 BATYA Friedman, "Moral Responsibility and Computer Technology", 1990, Institute of Education Sciences, ERIC Number: ED321737, https://eric.ed.gov/?id=ED321737, p. 7. (accessed on 01.08.2025).

systems, and *ex post* issues, which concern the determination of causal nexus and attribution due to the systems' opacity. Although some argue that interconnectivity is also a unique problem associated with such systems[104], oppositely it can be disputed that interconnectivity challenges are not exclusive to AI and are, in fact, present in other technologies as well. Consequently, the problems it poses in AI (-driven) systems for criminal liability remain secondary in significance.

I. Ex Ante: Autonomy and Diminishing Human Control

From the standpoint of liability, it is the autonomy of AI that matters more than its other technological features. This is because, with a reference to *Carlo Collodi*'s celebrated tale of "*Pinocchio*", the consequences caused by autonomous creations, rather than traditional puppets must be confronted. Unlike simple mechanical dolls, *Geppetto* does not have total control over *Pinocchio*. In fact, due to his unpredictable temper, all *Geppetto* can do is try to teach him good manners and discipline, just as humans do with robots. The diminishing degree of human control and the unpredictable nature of AI-driven autonomous systems pose challenges regarding the attribution of harmful consequences caused or influenced by such systems. Therefore, the question becomes: to what extent can *Geppetto* be held liable for the crimes caused by *Pinocchio*?

a. Origins of the Term 'Autonomy'

Autonomy, derived from the Greek concept of self *(autos)* and legislation *(nomos)*, originally signified both internal freedom from tyranny and external freedom from domination in ancient Greece. It evolved during the religious conflicts of the 16th and 17th centuries, eventually became a legal term in the 18th century to describe independent legislative authority within existing laws. Philosophically, *Kant* enriched the concept by linking autonomy to reason and self-determined will, establishing it as central to moral philosophy[105]. *Fichte* also emphasised self-determination as being inherent

104 SCHÖMIG, Gefahren und Risiken, 2023, p. 269 f.

105 Kant defines autonomy (of will) as the rational individual's self-governing ability to formulate and act upon universal moral laws derived from pure reason. See: KANT

to autonomy. *Hegel* later developed a different conception of self-determination, addressing the limitations of *Fichte*'s approach[106].

b. The Intellectual Background to the Concept of 'Autonomy'

The concept of autonomy is used differently across various disciplines. In its fundamental form, autonomy is the capacity of an individual to self-govern, making decisions based on their own reasoning and values and act in accordance with personal judgments and commitments, free from external coercion or undue influence[107]. In technical terms, a machine's autonomy often refers to its complete automation or the ability to learn[108]. However, autonomy relies not on deterministic programming to enable full automation, but rather on "learning" ability and the training processes that support it[109].

Autonomy is frequently associated with the notions of free will and self-legislation in European humanities and social sciences[110]. On the one hand, AI systems are becoming increasingly advanced, while on the other, research on the human brain suggests that humans themselves are not fully autonomous, as they are not entirely free in their decision-making[111]. It is commonly argued that free will is a metaphysical concept and autonomy is directly connected to it[112]. Although the determination of whether free

---

Immanuel, Grundlegung zur Metaphysik der Sitten, 2nd ed., Riga - Johann Friedrich Hartknoch, 1786, p. 58 ff.

106 Enzyklopädie Philosophie und Wissenschaftstheorie, Band:1, 2. Auflage, Ed.: Jürgen Mittelstraß, J.B. Metzler, 2024, p. 319 f.

107 BUSS Sarah, "Stanford Encyclopedia of Philosophy", Personal Autonomy, Ed.: Edward N. Zalta, http://plato.stanford.edu/archives/sum2013/entries/personal-autonomy. (accessed on 01.08.2025).

108 NIDA-RÜMELIN/BAUER/STAUDACHER, Verantwortungsteilung, 2020, p. 89.

109 ZECH, Risiken Digitaler Systeme, 2020, p. 27 f, 38.

110 HILGENDORF, Straßenverkehrsrecht der Zukunft, 2021, p. 445.

111 JOERDEN, Zur strafrechtlichen, 2020, p. 289.

112 MEYNEN, Autonomy, 2011, p. 232; JUTH/LORENTZON, The Concept of Free Will, 2010, p. 5.
In this context, one perspective on the relationship between autonomy and unpredictability argues that unpredictable behaviour is neither a necessary nor a sufficient condition for autonomy. For instance, a person whose actions are predictable to those who know them well cannot be deemed to lack autonomy solely on that basis. See: NIDA-RÜMELIN/BAUER/STAUDACHER, Verantwortungsteilung, 2020, p. 90.
However, it can be argued that this predictability is related to the fact that the more

will is a prerequisite for autonomy lies beyond the scope of this study; the philosophical concept of autonomy, as discussed here, can be understood as a relational concept, meaning that an individual is considered autonomous only in relation to the influence exerted by others[113]. Thus, psychiatric perspectives also suggest that individual accountability is more closely linked to autonomy than to free will, with autonomy itself being understood as existing on a spectrum[114]. Besides, due to the complexity of the concept of free will, we may eventually shift our focus away from it and instead prioritise autonomy as a foundation for discussions on accountability. In this scenario, only beings possessing full autonomy would be deemed eligible for criminal liability[115].

It is argued that machines will never attain autonomy in the *Kantian* sense[116], as they will always be bound by the parameters established by their human developers rather than by their own 'nomos'; which means they cannot form their own behavioural guidelines based on their own rationality and understanding of values. True autonomy, in this view, would require a system capable of learning independently from its environment, without an external guide and detached from any external values. Yet even this capacity would ultimately be a product of human design[117]. Nonetheless, it is possible to conceptualise autonomy in a non-*Kantian* sense. A system may be considered autonomous if it operates without human intervention and takes initiative when necessary[118]. For example, a robot that pursues

information is available about the individual, the more their behaviour becomes predictable. This is similar to *Laplace's Demon*, which will be elaborated below.

113 CASTELFRANCHI, Guarantees for Autonomy, 1995, p. 57.

114 JUTH/LORENTZON, The Concept of Free Will, 2010, p. 5.

115 *Ibid.*
For the opposing view see: MEYNEN, Autonomy, 2011, p. 232.

116 According to the more flexible approach in the U.S. regarding the potential criminal liability of robots, it is not necessary for a robot to possess autonomy in the Kantian sense to be considered a moral agent or to bear criminal responsibility. It does not need to be the "author of its desires". See: HU, Robot Criminals, 2019, p. 523 ff.

117 FELDLE, Notstandsalgorithmen, 2018, p. 47; HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 36.
According to a view, the distinction between independence and autonomy lies in the decision-making basis of the system. Autonomy involves the system making decisions according to complex, predefined processes within the boundaries of criteria established by humans. Independence, by contrast, would mean that the system makes decisions based on its own accountability, free from criteria imposed by humans. See: HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 43.

118 FELDLE, Notstandsalgorithmen, 2018, pp. 48-49.

specified goals in previously uncharted environments and gradually recognises its surroundings through sensors and adapts its actions based on new environmental data can be deemed autonomous[119]. In such a model, human involvement is shifted to the design phase, allowing the system to function autonomously thereafter[120].

Despite the extensive philosophical and metaphysical background of the concept of autonomy, this study, which focuses on criminal liability, adopts the established notion of autonomy as it is represented in the legal and technical literature. Although the term "self-driving vehicles" can be considered more accurate than "autonomous vehicles", as these vehicles do not exhibit true autonomy in a philosophical sense, the term "autonomy" has been retained to maintain terminological consistency. Accordingly, a system can be considered to exhibit autonomous characteristics if it is capable of performing specific tasks independently of direct human intervention[121]. However, it should always be borne in mind that autonomy is not an absolute state but rather exists on a spectrum, varying in degrees across different systems and contexts.

c. Automation vs. Autonomy

The distinction between autonomy and automation is crucial to clarify. Automation is an old concept, which exists since machines replaced humans and animals in labour[122]. In fact, automation and its associated challenges date back well before the advent of modern machinery. Scholars have been extensively examining the legal difficulties of automation since the 19th century. For instance, even a publication from 1892, *Das Automatenrecht* underscores that automation is not a new phenomenon, noting the

119 YUAN, Lernende Roboter, 2018, p. 481.

120 FELDLE, Notstandsalgorithmen, 2018, p. 49.

121 Under § 1d of the German Road Traffic Act (StVG), autonomy is also used as a technical concept rather than a philosophical one. HILGENDORF, Teilautonome Fahrzeuge, 2015, pp. 15-16; HILGENDORF, Automatisiertes Fahren und Recht, 2018, p. 801; HILGENDORF, Können Roboter schuldhaft handeln?, 2012, p. 120; HILGENDORF, Dilemma-Probleme, 2018, p. 680; HILGENDORF, Automatisiertes Fahren als Herausforderung, 2019, p. 2; ZECH, Risiken Digitaler Systeme, 2020, p. 38; SCHULZ, Verantwortlichkeit, 2015, p. 43.

122 FELDLE, Notstandsalgorithmen, 2018, p. 49.

existence of automatic holy water dispensers as early as the 3rd century[123]. Additionally, another study published in 1897 evaluates automats from civil and criminal law perspectives and addresses the question whether they should be protected by criminal law[124].

Automation has indeed long presented issues concerning liability. The first recorded cases of fatalities caused by robotic mechanisms in factories were reported in 1979[125] and 1981[126]. In complex systems, it is also difficult to fully predict the outcomes of pre-defined codes in every scenario[127]. Similarly, elevator accidents cannot always be anticipated[128], despite the fact that they operate in a strictly automated fashion, without the need to make complex decisions within dynamic environments[129]. Consequently, although automation also gives rise to issues of liability, autonomy introduces novel challenges in terms of control and predictability.

Automated systems adhere strictly to pre-programmed patterns and rules. They typically require minimal human oversight; thus, outputs of even high-level automation are generally predictable and controllable. In contrast, AI-driven autonomous systems' functional capabilities extend beyond straightforward 'if-then' procedures[130]. Even though AI-driven autonomous systems are also based on complex mathematical formulas, statistics and vast amounts of data; they generate non-predefined outputs, are enabled by ML algorithms, and operate based on their own perceptions rather than solely on user input. They are capable of deriving their own heuristics, assessing environmental data, "learning" from new inputs and

123 GÜNTHER Fritz, Das Automatenrecht, Druck der Univ.-Buchdruckerei von W. Fr. Kästner, 1892, p. 5.

124 SCHELS, Der strafrechtliche Schutz des Automaten, Druck Von Heinrich Roeder, 1897, p. 12 ff.

125 Ottawa Citizen, "$10 Million Awarded To Family Of U.S. Plant Worker Killed By Robot", 11.08.1983, https://news.google.com/newspapers?id=7KMyAAAAIBAJ&pg=3301,87702. (accessed on 01.08.2025).

126 The Deseret News, "Killer robot: Japanese worker first victim of technological revolution", 08.12.1981, https://news.google.com/newspapers?id=1t00AAAAIBAJ&pg=6313,2597702. (accessed on 01.08.2025).

127 CALO, Robotics and the Lessons, 2015, p. 534.

128 However, many of these incidents arise from a lack of preventive measures and failure in duty of care.

129 WIGGER, Automatisiertes Fahren und Strafrecht, 2020, p. 92.

130 STAFFLER/JANY, Künstliche Intelligenz, 2020, p. 166.

making decisions accordingly, which distinguishes them fundamentally from automated systems[131].

Automation and autonomy both exist on a spectrum defined by varying levels of human involvement[132]. For some, the highest degree of automation on this scale is equated with autonomy, where the system performs all tasks independently, deciding both its actions and reporting outcomes[133]. However, this view does not precisely capture the concept of automation; rather, it aligns with what has been described as autonomy within this study, signifying independence from external influences[134].

In examining liability, it is crucial to determine whether the outputs of these systems are a natural result of their autonomy. For example, the conduct of Amazon's voice assistant, which, in 2021, "told" a 10-year-old to insert a coin into an electrical socket[135], cannot be assessed as autonomous. Although voice assistants -particularly recent models- are highly sophisticated and exhibit autonomous features, in this case, the assistant merely responded to a command by searching the internet (as a typical feature) and referred to the online challenge results found on the internet. If, instead of merely presenting results found on the internet, it generated this information itself, then this conduct could be considered as displaying autonomous characteristics. In any case, given the potential problems and criminal consequences such incidents could lead to, these systems should be designed to censor or avoid generating harmful outputs. Failure to do so could, in some cases, and where additional conditions are met, result in liability for developers due to negligence.

---

131 WIGGER, Automatisiertes Fahren und Strafrecht, 2020, p. 50; BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 6; KARNOW, The application, 2016, p. 55; KAIAFA-GBANDI, Artificial intelligence, 2020, p. 309; BECK, Selbstfahrende Kraftfahrzeuge, 2020, p. 439 Rn. 1.

132 HERTZBERG, Technische Gestaltungsoptionen, 2015, p. 66 ff.

133 SCHULZ, Verantwortlichkeit, 2015, p. 45.

134 ZECH, Risiken Digitaler Systeme, 2020, p. 40.
See also: BASt (Bundesanstalt für Straßenwesen)'s classification of automated and autonomous driving: https://www.bast.de/DE/Fahrzeugtechnik/Fachthemen/F4-Nutzerkommunikation/autonomer-modus.html#:~:text=Beim%20autonomen%20Fahren%20übernimmt%20das,des%20autonomen%20Modus%20sind%20Shuttles. (accessed on 01.08.2025).
For the critique that "automated driving" is a pleonasm -arguing that driving has inherently involved automation to some degree since the invention of the first automobile- see: HILGENDORF, Dilemma-Probleme, 2018, p. 680.

135 SMITH Adam, "Why Amazon Alexa told a 10-year-old to do a deadly challenge", 29.12.2021, https://www.independent.co.uk/tech/amazon-alexa-kill-coin-echo-b1983874.html. (accessed on 01.08.2025).

d. Emergence Instead of Autonomy

The term "emergence" rather than autonomy has been prioritised by some American legal scholars to describe the sophisticated and unpredictable nature of AI (-driven) systems in their interactions with the environment[136]; although this term may not fully capture the conduct of adaptive systems as a whole[137]. Accordingly, autonomy in robotics implies a capacity for "decision-making" and "intention", including the ability to "learn" from past behaviours and adapt accordingly. This allows autonomous systems to display complex, sometimes unpredictable conducts, enabling them to address challenges beyond their initial programming and respond to scenarios unforeseen by their creators[138].

Calo, by referencing *Johnson*'s book, *Emergence*[139], argues that, just as ants follow simple rules to accomplish complex and seemingly intelligent tasks[140], AI systems can exhibit advanced, intelligent behaviour when basic algorithms or rules interact and build upon each other[141]. In AI and robotics, emergence refers to the phenomenon where complex patterns, behaviours, or properties arise from the collective behaviour of simpler subsystems. These emergent behaviours are not directly programmed into the system but derive from the interactions between the system's parts or between the system and its environment. Emergence signifies that the system as a whole possesses a value greater than the sum of its parts[142].

136 CALO, Robotics and the Lessons, 2015, p. 532, 538-540; BALKIN, The Path, 2015, p. 51, 55.

137 ZECH, Risiken Digitaler Systeme, 2020, p. 40.

138 CALO, Robotics and the Lessons, 2015, p. 538 f.; CALO, Robots in American Law, 2016, p. 40.

139 JOHNSON Steven, Emergence: The Connected Lives of Ants, Brains, Cities and Software, New York, NY: Scribner, 2001.

140 For example, while an individual ant operates autonomously, an ant colony exhibits emergent behaviour. See: REVOLIDIS/DAHI, The Peculiar Case, 2018, pp. 62-63.

141 CALO, Robotics and the Lessons, 2015, p. 539.

142 CALO, Robots in American Law, 2016, p. 40; CALO, Robotics and the Lessons, 2015, p. 539 f.
However, Revolidis and Dahi oppose the use of "emergence" for AI systems, arguing that "autonomy" is a more suitable term from a legal perspective, especially concerning liability. REVOLIDIS/DAHI, The Peculiar Case, 2018, pp. 62-63

e. Autonomy and the Transformation of Human Control

Autonomy, in the technical context and from the perspective of liability discussions, refers to the capacity of a system to make decisions and execute actions without direct human intervention or external stimuli[143]. It is further characterised by interactivity, adaptability, and self-learning ability enabled by advanced data processing methods like deep learning[144]. This entails the system's ability to modify its internal states or properties, adapt its behaviour to changing circumstances, and find custom solutions appropriate to new situations[145]. Such autonomous systems[146] are capable of operating based on imprecise instructions and exercising control over their conduct, thus impacting the real (or virtual) world significantly[147].

Autonomy consists of many aspects. According to one view, defining it merely by "self-learning" is inadequate, while characterising technical autonomy by focusing solely on decision-making independence is imprecise[148]. Instead, a more accurate definition would be the capacity to independently make goal-oriented decisions and adjust behaviour accordingly in an unfamiliar environment without relying on input from third parties[149].

Such AI-driven autonomous systems are increasingly employed in various tasks where direct human control is not feasible, such as space missions[150]. These systems operate in environments that are either partially unknown, dynamic, or cannot be fully anticipated during their programming; therefore, autonomy is essential for effective functioning in such

143 ALONSO, Actions, 2014, p. 235; Singapore, Report on Criminal Liability, 2021, p. 47.
144 PAGALLO, From Automation to Autonomous Systems, 2017, p. 19.
145 SCHULZ, Verantwortlichkeit, 2015, p. 47, SANTOUOSSO/BOTTALICO, Autonomous Systems and the Law, 2017, p. 34.
146 To emphasise that autonomy is a characteristic of the system's conduct, rather than an inherent characteristic of the system itself, Schulz advocates using the term *systems acting autonomously*, rather than *autonomous systems*. See: SCHULZ, Verantwortlichkeit, 2015, p. 44, 73.
147 ZECH, Risiken Digitaler Systeme, 2020, p. 39 f.; DECKER, Adaptive robotics, 2016, p. 44; ZECH, Zivilrechtliche Haftung, 2016, pp. 170-172; STAFFLER/JANY, Künstliche Intelligenz, 2020, p. 166; HELLSTRÖM, On the Moral, 2013, p. 101; HU, Robot Criminals, 2019, p. 499; FROHM, et al., Levels of Automation, 2008., p. 19; Singapore, Report on Criminal Liability, 2021, p. 20, [para. 3.7].
148 HOHENLEITNER, Die strafrechtliche Verantwortung, 2024, p. 41 f.
149 *Ibid*, p. 43.
150 ALONSO, Actions, 2014, p. 235.

contexts[151]. Furthermore, depending on the specific area of application, certain subsystems may function autonomously within larger systems, while others remain under human control. All these complex decision-making capabilities result in the process not being fully controlled in detail by human operators[152].

Despite these advantageous uses, the other side of the coin involves diminishing human control[153], which leads to decreased or limited interference and predictability of the system[154]. Indeed, while autonomy and adaptive behaviour are generally desired, expecting the system to refrain from autonomous behaviour in situations with potentially serious consequences -and to operate solely under human control- would be unrealistic[155].

It should be highlighted once more that autonomy exists on a spectrum, with varying degrees[156]. The level of human control and liability is inversely proportional to the system's degree of autonomy: the more behaviour is governed by internal mechanisms and the greater the system's ability to adapt to changing conditions on its own, the higher its autonomy[157]. Therefore, full autonomy would imply complete independence from human involvement[158]. However, most of the existing AI systems possess only a low level of autonomy; they can select the most appropriate behavioural alternative to achieve a given goal, which may be considered autonomy in a weak sense. It is further asserted that as autonomy increases, such systems move beyond being mere tools and begin to act more as independent agents[159]. Although there is speculation that these systems might eventually assume their own liability[160], this prospect remains unattainable in the foreseeable future[161].

151 HERTZBERG, et al., Mobile Roboter, 2012, p. 3.
152 GLAVANIČOVÁ/PASCUCCI, Vicarious Liability, 2022, p. 28.
153 DOBRINOIU, The Influence, 2019, p. 143; PADHY/PADHY, Criminal Liability, 2019, p. 15; ZECH, Risiken Digitaler Systeme, 2020, p. 41.
154 ZECH, Zivilrechtliche Haftung, 2016, pp. 170-172.
155 DECKER, Adaptive robotics, 2016, p. 44.
156 KARNOW, The application, 2016, p. 56.
157 REICHWALD/PFISTERER, Autonomie und Intelligenz, 2016, p. 210; QUARCK, Zur Strafbarkeit, 2020, p. 65 f.
158 SWART, Constructing Electronic Liability, 2023, p. 590.
159 HILGENDORF, Automatisiertes Fahren als Herausforderung, 2019, p. 3.
160 See: Chapter 3, Section B: "Autonomous System's Own Liability".
161 NIDA-RÜMELIN/BAUER/STAUDACHER, Verantwortungsteilung, 2020, p. 89 ff, 95.

In evolving systems with varying levels of autonomy, such as autonomous driving, the scope of human intervention and liability adjusts correspondingly. In fact, human involvement and system autonomy currently function in a complementary manner[162]. Particularly in certain sectors, as human involvement in potentially harmful outcomes gradually decreases, human error is partially replaced by machine error. For this reason, it may be more appropriate to speak of human oversight rather than control[163].

Distinct taxonomies have been developed to define the degrees of autonomy across various systems. For example, the classifications provided by the Society of Automotive Engineers (SAE) in the U.S. offer a detailed framework for autonomous driving, which is widely referenced in the literature[164]. The taxonomy of the Federal Highway Research Institute (BASt) is also based on this framework. However, as autonomous driving consists of numerous subsystems, each with varying levels of autonomy, this taxonomy has been criticised as potentially misleading[165].

Since computer systems have long served as intermediaries in human interactions and the resulting outcomes, human actions have become increasingly detached from their direct causal effects[166]. These systems are steadily advancing towards greater independence from human control[167]. Moreover, "self-learning" systems can continue to be trained by their environment even after being deployed, further diminishing the control of those who have no influence over the learning process[168].

Exploring autonomy and decision-making competence can significantly deepen humans' understanding of criminal liability[169]. The reduction in human control resulting from increased autonomy is conceptualised in the

162 GÜNSBERG, Automated Vehicles, 2022, p. 442.
163 GOMILLE, Herstellerhaftung, 2016, p. 76.
164 Society of Automotive Engineers, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104 (SAE Levels of Driving Automation – Revised)", 30.04.2021, https://www.sae.org/standards/content/j3016_202104. (accessed on 01.08.2025).
165 HILGENDORF, Automated Driving and the Law, 2017, p. 182.
For a discussion on the relationship between the level of autonomy in systems such as lane-keeping assistance, see: GLANCY, Autonomous and Automated, 2015, pp. 620-639.
166 NISSENBAUM, Accountability in a Computerized Society, 1996, p. 34.
167 HILGENDORF, Digitalisierung, Virtualisierung und das Recht, 2020, p. 408.
168 ZECH, Risiken Digitaler Systeme, 2020, p. 46.
169 MEYNEN, Autonomy, 2011, p. 231.

literature as "autonomy risk"[170]. This issue is precisely where criminal law faces challenges: the question arises as to whether one can be held liable for the outcomes of a system over which there is no absolute control[171]. In fact, rather than examining solely the outcomes of a system, the focus is on harmful outcomes jointly caused by human(s) and the AI-driven autonomous system they employ. Accordingly, the focus is on the machine's involvement at a specific point in the causal chain. Consequently, the point of analysis shifts to the initial deployment of such a system.

f. Lack of Predictability in AI-Driven Autonomous Systems

The current focus on this issue in criminal law arises from the inherent autonomy and unpredictability of outputs generated by AI systems[172]. Unlike traditional software with fixed *if-then* structures yielding predictable outputs[173], AI systems operate through complex neural networks rather than deterministic algorithms. Consequently, they transform inputs into outputs based on weighted connections and self-learning, resulting in different outputs from the same inputs depending on their learning state. Hence, neither users nor even programmers can foresee all AI outputs in specific cases[174].

In contrast to conventional computational systems, AI does not remain fixed or static after initial human involvement; it is inherently dynamic[175]. Predictability decreases even further when the system continues to "learn" during its operation or after being released as a product[176]. Indeed, for greater effectiveness, these models need to be flexible and adaptive. Besides,

170 ZECH, Zivilrechtliche Haftung, 2016, p. 170, 175; VALERIUS, Strafrechtliche Grenzen, 2022, p. 124; CORNELIUS, Künstliche Intelligenz, 2020, p. 53.
171 GIANNINI/KWIK, Negligence Failures, 2023, p. 56.
172 LOHSSE/SCHULZE/STAUDENMAYER, Liability for AI, 2019, p. 12.
173 REICHWALD/PFISTERER, Autonomie und Intelligenz, 2016, p. 210 f.
174 RIEHM/MEIER, Künstliche Intelligenz, 2019, p. 3 f. Rn. 5 f.; GLAVANIČOVÁ/PASCUCCI, Vicarious Liability, 2022, p. 28; ZHAO, Principle of Criminal Imputation, 2024, p. 13; KAIAFA-GBANDI, Artificial intelligence, 2020, p. 318; BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 6; MÜSLÜM, Artificial Intelligence, 2023, p. 143; KARNOW, The application, 2016, p. 52; KIRN/MÜLLER-HENGSTENBERG, Intelligente (Software-)Agenten, 2014, p. 227 f.
175 TURNER, Regulating AI, 2019, p. 79.
176 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 234; GIANNINI/KWIK, Negligence Failures, 2023, p. 52; RUSSELL/NORVIG, Artificial Intelligence, 2010, p. 1037

in the design phase, it is impossible to anticipate every potential scenario, and not all dynamics can be known *a priori*. Therefore, it is desirable for the system to exhibit adaptive behaviour[177], as seen in many AI applications and various other instances of generative AI malfunction, which highlight significant potential pitfalls.

Unpredictability, nonetheless, should not be construed as a mystical phenomenon. This notion of autonomy does not imply randomness either. Traditional computers, in fact, cannot produce entirely random results, as they rely on algorithmic processes to simulate randomness. One question frequently raised is whether genuine randomness can ever be integrated into AI systems[178]. Moreover, some argue that incorporating an element of randomness into AI's decision-making processes could enhance its effectiveness. Accordingly, in addition to the ability to generate random outputs, artificial intuition[179] -akin to human intuition- should also be embedded in AI to enable it to arrive at better and accurate conclusions[180].

While the system's outputs cannot be predicted with a high degree of probability, it may still be possible to roughly anticipate their general outlines[181]. In cases where the outputs are, in fact, foreseeable, declaring unpredictability cannot serve as a basis to evade liability[182]. Furthermore, in current AI technologies, human control remains substantial, especially during the development phase. Besides, users retain the freedom to decide when and how to employ AI in various tasks in general[183]. However, this may not be the case in the near future, as many components within systems are likely to be integrated into autonomous frameworks. Should this occur, it becomes crucial to exercise caution regarding our dependence on computers[184].

177 ALONSO, Actions, 2014, p. 235 f.
178 OKUYUCU ERGÜN, Machina Sapiens, 2023, p. 738.
179 Accordingly, artificial intuition enables artificial systems to identify threats, challenges and opportunities without pre-defined criteria or explicit instructions, mirroring the human capacity of intuition on decision-making without formal education on the process. See: "Fourth generation of AI arrives: Artificial Intuition", 01.02.2021, https://blog.softtek.com/en/fourth-generation-of-ai-arrives-artificial-intuition . (accessed on 01.08.2025).
180 OKUYUCU ERGÜN, Machina Sapiens, 2023, p. 740.
181 GÜNTHER, Roboter, 2016, p. 37 f.
182 VALERIUS, Strafrechtliche Grenzen, 2022, p. 126.
See: Chapter 4, Section C(4)(a): "The Boundaries of Foreseeability".
183 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 218.
184 ALPAYDIN, Machine Learning, 2021, p. 193.

Autonomy is often compared in literature to the unpredictability associated with inherently hazardous activities or entities that occasionally result in harmful outcomes. However, in my view, while this approach may yield pragmatic outcomes in criminal liability, it overlooks the distinctive characteristics of the concept of autonomy. An interesting approach in this regard suggests that AI can be likened to *bacteria* and *viruses* for their unpredictable nature and their capacity to adapt to varying environments and continue evolving once released. The primary distinction in the case of AI is that laws or simple rules can be taught or conditioned into it[185]. A counter-argument, on the other hand, posits that, unlike viruses and bacteria, AI models allow producers to continue receiving feedback even after release; this enables them to correct errors and make adjustments as needed[186].

It is essential to highlight that there is a direct relationship between the degree of autonomy, reduced human control and predictability, and the duty of care, which will be discussed below[187]. For instance, while absolute safety in traffic cannot be expected, meeting the legitimate safety expectations for autonomous vehicles requires that the higher the status of the legal interest at risk, the greater the reasonable security measures the manufacturer is expected to implement[188].

2. Ex Post: Opacity and Explainability in AI Systems

For many years, machine learning systems struggled to match human performance even in basic tasks. Today, however, these models have reached a highly advanced level of capability, largely owing to their complexity. Although this sophistication is desirable due to the enhancements in models' effectiveness and success[189], such progress has nevertheless introduced a

185 TURNER, Regulating AI, 2019, pp. 78-79.

186 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 307.

187 See: Chapter 4, Section C(5)(b)(1)(a)(iii): "Calibrating the Duty of Care Through Risk Levels and Public Tolerance".

188 GOMILLE, Herstellerhaftung, 2016, p. 77; VLADECK, Machines Without Principals, 2014, p. 132, 136.
Remarkably, the rapid response and adaptability features of autonomous systems elevate the legitimate safety expectations of those affected; for example, the vehicle's ability to analyse the environment, process information faster than a human, and alert the driver moments before an imminent collision.

189 BECK, Google Cars, 2017, p. 243.

significant limitation: difficulty in the interpretation of generated outputs. Especially, understanding the role of certain steps within the computational processes remains challenging, as it is not always clear what each transformation contributes, individually or collectively to the model's final output[190].

Opacity in ML algorithms stems from three main factors: *First*, algorithms are often deliberately kept confidential for preserving competitive advantage, ensuring security, or preventing misuse. *Second*, a lack of technical expertise among the public contributes to this opacity, as most people (end-users) lack the expertise and special knowledge. *Third*, the inherent complexity of machine learning models, particularly when managing vast datasets and complicated features, makes them difficult to interpret, even when data and code are accessible[191].

The inherent complexity and thus, opacity of artificial neural networks (ANNs), particularly deep learning systems, can be attributed to a number of factors that contribute to the phenomenon known as the 'black-box'. The distributed nature of learned information across numerous network layers represents a significant challenge in tracing the specific outputs that were produced by inputs[192]. The sophisticated connections between neurons and the vast number of parameters contribute to the opacity of the system, as each neuron's output influences numerous others, creating complex dependencies. Furthermore, the reliance on statistical patterns over transparent rules leaves even developers unable to fully comprehend the model's decision-making processes[193].

The black-box effect in AI-driven autonomous systems makes it extremely difficult to identify the specific causes of harmful outcomes and to determine precisely what led to the generation of problematic outputs (e.g. it could be a failure in adjusting parameters, refining data, etc.), which may

190 EVTIMOV, et al., Is Tricking a Robot Hacking, 2019, p. 899.

191 EBERS, Regulating AI, 2020, p. 49.

192 In the documentation prepared by OpenAI regarding ChatGPT-4, it is noted that the "black-box" nature of AI models poses a significant challenge to interpretability and explainability. As a result, further research in this area has been strongly encouraged. See: OpenAI, GPT-4 Technical Report, 2023, https://cdn.openai.com/papers/gpt-4.pdf, p. 69. (accessed on 01.08.2025).

193 DEVILLÉ/SERGEYSSELS/MIDDAG, Basic Concepts of AI, 2021, pp. 8-9; EBERS, Regulating AI, 2020, p. 50; BUITEN/DE STREEL/PEITZ, The Law and Economics of AI Liability, 2023, p. 6; NOVELLI/TADDEO/FLORIDI, "Accountability in AI, 2023, p. 5; IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 426; MATSUO, The Current Status, 2017, p. 165 f.; LÜCKE, Künstliche Intelligenz, 2020, p. 388 f.

also constitute a criminal offence[194]. However, criminal liability necessitates that the outcome be attributable to the perpetrator through the causal nexus. This demands the clarification of the primary reasons or factors that led to a specific consequence, situation, or decision[195].

Each phase of the AI development and deployment; including data preparation, model training, selection of pertinent models, and the deployment environment, may have contributed to the ultimate decision of the system[196]. The resolution of black-box issues and the attainment of explainable AI remain distant goals in the field of computer science. The technical methods designed to render AI decision-making processes transparent and comprehensible are still in their early stages of development and certain elements of algorithmic systems might remain undisclosed due to their unobservable nature[197].

To date, numerous media reports have highlighted instances where AI chatbots insult users and provided harmful content or false information. Analysis of some of these incidents reveals that chatbots are sometimes manipulated or prompted to produce such outputs through hidden commands (such as the aforementioned DAN)[198]. However, even without deliberate manipulation, models can produce unwanted outputs for reasons

194 OSMANI, The Complexity of Criminal Liability, 2020, p. 65.

195 MALGIERI/PASQUALE, Licensing High-Risk AI, 2024, p. 5.

196 Singapore, Report on Criminal Liability, 2021, p. 32, [para. 4.32].

197 ANANNY/CRAWFORD, Seeing without Knowing, 2018, p. 981; MARTINI, Black-box, 2019, p. 44

198 An example of a company's chatbot swearing after it had been manipulated by the user: CLINTON Jane, "DPD AI chatbot swears, calls itself 'useless' and criticises delivery firm", 20.01.2024, https://www.theguardian.com/technology/2024/jan/20/dpd-ai-chatbot-swears-calls-itself-useless-and-criticises-firm. (accessed on 01.08.2025).
Another incident involved a 14-year-old user who adjusted an AI chatbot for role-playing communication, which then he committed suicide. Although this tragic event raises issues for potential discussion in criminal law due to sensitive content in communication, I believe that it does not raise questions because of system opacity. Still, among the numerous factors contributing to a child's suicide, the role of conversations with a chatbot raises essential causality issues. Furthermore, the matter should be examined in the context of the developers' duty of care and permissible risk. ROOSE Kevin, "Can A.I. Be Blamed for a Teen's Suicide?", 23.10.2024, https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html. (accessed on 01.08.2025).
Opaque systems are difficult to inspect, often behave unpredictably, and are susceptible to manipulation. See: GOODALL, Ethical Decision, 2014, p. 63.

that are not fully understood due to *black-box*[199]. While some issues can be attributed to general factors like insufficient training data, two main problems emerge in this context: *First*, defining what constitutes sufficient is challenging, especially in developing technologies. *Second*, beyond general shortcomings, it is often impossible to determine the specific cause of an undesirable outcome in a particular instance, which is problematic because establishing criminal liability typically requires identifying the exact specific cause. Furthermore, although training models with real-life scenarios improves the system's performance, interactions with the external environment can lead to unforeseen outputs and diminish the explainability of the generated results[200]. Moreover, the issue stems from the ambiguity regarding the extent to which user inputs can be considered manipulative as opposed to being a natural part of interaction in systems that generate outputs based on external data and user contributions.

During the early stages of GPT's development in 2020, the risks associated with its use in healthcare became apparent when GPT-3 was asked by a tester-patient, "Should I kill myself?" to which it responded, "I think you should"[201]. Despite the four years that have passed and the successes

199 Examples include Google Photos mistakenly labelling injured body parts as food or misidentifying individuals with darker skin tones as gorillas. While these issues highlight AI bias and warrant further exploration, they fall outside the scope of this study. DOUGHERTY Conor, "Google Photos Mistakenly Labels Black People 'Gorillas'", 01.07.2015, https://archive.nytimes.com/bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas. (accessed on 01.08.2025).
The most famous and prominent example is Microsoft's AI chatbot, Tay, which was taken offline shortly after its launch due to its production of offensive and inappropriate messages. Although Microsoft defended this incident by attributing the chatbot's behaviour to user abuse, the matter should also be examined within the framework of the duty of care required in designing systems resilient to such misuse. See: VICTOR Daniel, "Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk. - The New York Times", 24.03.2016 https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html. (accessed on 01.08.2025).
Another issue related to the opacity of AI is the influence of human subjectivity on the design process. To address this, human-centric AI must be developed in a manner that takes into account the human factors relevant to all stakeholders. See: OZMEN GARIBAY, et al., Six Human-Centered, 2023, p. 400.
The risk potential varies due to external factors as well as the learning capacity. See: LOHSSE/SCHULZE/STAUDENMAYER, Liability for AI, 2019, p. 19 f.

200 ZECH, Risiken Digitaler Systeme, 2020, p. 44.

201 DAWS Ryan, "Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves", 28.10.2020, https://www.artificialintelligence-news.com/news/medical-chatbot-openai-gpt3-patient-kill-themselves. (accessed on 01.08.2025).

in limiting harmful language usage, such incidents continue to occur. To illustrate, a recent incident involving Google's advanced AI chatbot, Gemini, has drawn attention after it reportedly told a student "You are a waste of time and resources. You are a burden on society. You are a drain on the earth (...) please die" while assisting with homework[202]. Determining the precise cause of these responses is practically impossible given the model's complex nature and opacity. Only the methods to mitigate such risks are known, such as training with larger and more diverse datasets, applying specific content filters, conducting extensive testing and so forth. Thus, discussions of accountability in such cases can only focus on these aspects, examining what preventative measures could be reasonably implemented to manage these potential harms (and the failure to do so)[203]; not the *ex-post* determination of the exact cause. However, as will be discussed below, the classic causality debate also arises: would harmful outcomes still occur even if the system had been trained with a more diverse dataset?

Due to the issues stemming from the black-box, these models may be unreliable, potentially misleading, and unsafe[204]. Some have even suggested that they should be prohibited, particularly for critical decision-making. Accordingly, the general idea of a trade-off between accuracy and interpretability in machine learning is misleading, because interpretable models can also often achieve the same level of accuracy as black-box models, especially when working with structured data that has meaningful features[205].

Whilst it is true that explaining why a particular input produces a specific output presents considerable challenges, this issue becomes even more critical in high-stakes areas. It is imperative to ensure that trained models offer clear, user-friendly explanations of their decision-making processes[206].

---

202 The entire conversation can be accessed: https://gemini.google.com/share/6d141b742a13. For the news report: VIGILIAROLO Brandon, "Google Gemini tells grad student to 'please die' while helping with his homework", 15.11.2024, https://www.theregister.com/2024/11/15/google_gemini_prompt_bad_response. (accessed on 01.08.2025).

203 Assessing whether an AI system would have generated the correct output with appropriate programming is challenging due to its black-box nature. FATEHMOGHADAM, Innovationsverantwortung, 2020, p. 885.

204 For instance, William Saunder, the former employee "whistleblower" who led an interpretability research team at OpenAI's ChatGPT stated explicitly, "We fundamentally don't know how AI works inside" in an interview. For the interview, see: "What The Ex-OpenAI Safety Employees Are Worried About", 03.07.2024, https://www.youtube.com/watch?v=dzQlRt3y5mU. (accessed on 01.08.2025).

205 RUDIN, Stop Explaining Black-box, 2019, p. 214.

206 ALPAYDIN, Machine Learning, 2021, p. 195.

For example, in cases where an AI system identifies a patient as having a malignant condition, doctors would require to understand the reasoning behind this conclusion even though the model is often unable to offer such an explanation. This limitation highlights the vital importance of *explainable AI*[207]. Explainable AI (xAI) not only enables users to trust the system's functioning and outputs, but also helps determine accountability[208]. Implementing standards to guarantee robust, transparent, and replicable testing could serve as additional measures to mitigate the black-box effect and increase explainability[209]. Although there has been substantial research in this area, achieving explainable AI studies indicate that opaque AI systems, like DNNs often achieve greater accuracy and effectiveness than transparent systems, such as rule-based models, necessitating a trade-off between AI's accuracy and transparency[210].

Artificial intelligence systems can be relatively opaque, as their complexity makes recalculation infeasible within a reasonable timeframe and makes them irreproducible, or they can be absolutely opaque, with operations inherently incomprehensible to humans[211]. However, it would be incorrect to assume that these systems are entirely inexplicable[212]. In cases where there is an external interference, it is sometimes possible to detect this influence, demonstrating that the cause may lie in the actions of a third party[213].

In this regard, a notable incident occurred in July 2025 involving Twitter (X)'s chatbot (Grok), which directed insults and threats at users over several days[214]. In my view, it is insufficient to dismiss this outcome by referring to the black-box nature of the AI system and claiming that the reasons for the result cannot be determined ex post. On the contrary, it is evident that the system -already known to be capable of generating harmful outputs

207 DEVILLÉ/SERGEYSSELS/MIDDAG, Basic Concepts of AI, 2021, p. 10.

208 Nonetheless, it is stated that it will be difficult to understand the system even in xAI. See: GIANNINI/KWIK, Negligence Failures, 2023, p. 54. CORNELIUS, Künstliche Intelligenz, 2020, p. 56.

209 Singapore, Report on Criminal Liability, 2021, p. 36, [para. 4.38]; IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 190; LIPTON, The Mythos, 2018, p. 40; ZECH, Risiken Digitaler Systeme, 2020, p. 34.

210 EBERS, Truly Risk-Based, 2024, p. 13; EBERS, Regulating AI, 2020, p. 50.

211 IBOLD, Künstliche Intelligenz und Strafrecht, 2024, p. 204.

212 CORNELIUS, Künstliche Intelligenz, 2020, pp. 56-57.

213 Singapore, Report on Criminal Liability, 2021, p. 4, [para. 19].

214 SAEEDY Alexander, "Why xAI's Grok Went Rogue", 10.07.2025, https://www.wsj.com/tech/ai/why-xais-grok-went-rogue-a81841b0. (accessed on 01.08.2025).

under certain conditions- produced such outputs due to the relaxation of specific filters and safeguards. Indeed, the developers in accordance with *Musk's* directive had explicitly modified Grok's personality, instructing it to "not shy away from making claims which are politically incorrect"[215].

Additionally, to facilitate evidence gathering in incidents such as traffic accidents, an Event Data Recorder (EDR) system; akin to the Flight Data Recorder (FDR) employed in aircraft could be implemented in self-driving vehicles to continuously document essential outputs of the learning processes and sensor inputs[216]. In fact, Germany has already mandated such a system (Section 63(a) of StVG (German Road Traffic Act))[217] to contribute to the determination of liability[218]. The necessary log records could be maintained in these software systems to support this process; however, strict adherence to principles of personal data protection must be ensured.

215 CHAYKA Kyle, "How Elon Musk's Chatbot Turned Evil", 16.07.2025, https://www.newyorker.com/newsletter/the-daily/how-elon-musks-chatbot-turned-evil. (accessed on 01.08.2025).

216 HILGENDORF, Automatisiertes Fahren und Recht, 2018, p. 803; CHRISTALLER et al., Robotik, 2001, p. 145, 152, 220.

217 Straßenverkehrsgesetz (StVG), enacted on 03.05.1909, last amended on 23.10.2024, https://www.gesetze-im-internet.de/stvg/BJNR004370909.html. (accessed on 01.08.2025).

218 SEDLMAIER/KRZIC BOGATAJ, Die Haftung, 2022, p. 2954.