"TRUSTWORTHY AI" CANNOT BE TRUSTED: A VIRTUE JURISPRUDENCE-BASED APPROACH TO ANALYSE WHO IS RESPONSIBLE FOR AI ERRORS

Shilun Zhou*

Abstract: Erroneous results generated by artificial intelligence (AI) have opened up new questions of who is responsible for AI errors in legal scholarship. I support the prevailing academic view that human subjects should be held responsible for AI errors. However, I argue that the underlying reason is not pertained to the reliability of AI, but rather the inability of humans to establish a trusting relationship with AI. The term 'Trustworthy AI' is just a metaphor, which presents a sense of trust; AI itself is not trustworthy. The first section outlines the academic debate on the responsibility of AI. It contends that the perspective of these debates has shifted from the characteristics of AI, such as autonomy and explainability, to a human-centred perspective, which is how humans should develop AI. The assumption of responsibility depends on the existence of a trust relationship because when people believe that an individual can fulfil his or her responsibilities, they are willing to hand over power, resources or tasks to that individual. It applies a virtue jurisprudence-based approach to explain why humans cannot establish a trust relationship with AI To establish such a relationship, one subject must indicate to the other that its behaviour is based on specific moral motivation and that it can be held moral responsibility. Nevertheless, AI lacks moral motivation and moral responsibility. The third section reconsiders the scope of responsible subjects for AI errors. It posits that accountability should be limited to the individuals who are direct beneficiaries of the AI product. Finally, it argues that the scope of responsibility for AI errors should be disparate pursuant to the risk level of the AI. For high-risk AI, responsible subjects must fulfil both the obligations under the AI Act and the obligation to provide technical authentication.

Keywords: AI Law; Trustworthy AI; AI Act; Legal Responsibility; Virtue Jurisprudence

* University of Edinburgh, United Kingdom.

Table of Contents

I. Statement of Question: Who is Responsible for AI Errors 188

A. Through the Lens of Instrumentalism: AI's Inability to Assume Responsibility 188

B. AI with Enhanced Autonomy: Challenging the Principle of Foreseeability 189

C. Through the Lens of Anthropocentrism: The Concept of 'Trustworthy AI' 191

II. A Shift from Focusing on AI's Reliability to Trustworthiness: Humans Cannot Establish A Trust Relationship With AI 193

A. Virtue Jurisprudence Approach to Interpreting 'Trustworthy AI' 193

B. AI Lacks Moral Motivation 196

C. AI Cannot Take Moral Responsibility 197

III. Reframing the Scope of Responsible Subjects for AI Errors 198

A. AI Act's Formal Requirements for the Scope of Responsible Subjects 199

B. Substantive Requirements for the Scope of Responsible Subjects: the Direct Beneficiaries of AI Products 200

IV. Reframing the Scope of Human Responsibility for AI Errors 201

A. Joint Obligations for Human Subjects Responsible for AI Across Different Risk Levels 202

B. Obligations of Human Subjects Responsible for High-Risk AI Products to Provide Technical Authentication 204

Conclusion 207

Bibliography 209

I. STATEMENT OF QUESTION: WHO IS RESPONSIBLE FOR AI ERRORS

This chapter uses the chronological development of AI to classify academic standpoints on the attribution of responsibility for AI errors. This essay reviews the academic standpoints on responsibility for AI errors and contends that the analytical perspective has been transitioned with the ongoing evolution of AI technology. More precisely, I argue that the focus has moved from the characteristics of AI, such as its transparency, explainability, and autonomy, to how humans should treat AI.

A. Through the Lens of Instrumentalism: AI's Inability to Assume Responsibility

In its early stages, AI was primarily used for big data querying and retrieval. This big data querying can be differentiated from traditional information retrieval.[1] First, the big data era is characterized by fruitfulness volumes of data, rapid growth, and a focus on prediction. Data querying can handle enormous databases.[2] Second, big data query technology employs different statistical methods compared to traditional query methods.[3] More precisely, statistical methods focused on sample analysis, aiming to extract the most information from minimal data through random sampling.[4] In contrast, the big data approach analyzes entire datasets, treating the sample as the population.[5] In such instances, AI serves as a tool for data retrieval, querying, detection, and storage.[6] It generates information from existing databases rather than creating novel insights.[7] A notable example is the smart dashcams, which captures information such as sound, time, location, and speed. This smart device stores information either within the device itself or in a cloud network, creating a comprehensive record of the incident.[8]

From the perspective of the generation path of AI errors, the data that AI relies on is entirely manipulated by humans. This implies that any AI error stems from inaccuracies in the data provided by humans.[9] Additionally, AI may produce inaccurate, ambiguous, or incorrect information due to wear and environmental factors.[10] For instance, using an alcohol tester (ADLAIA) to assess a driver's sobriety may yield erroneous results if the instrument is contaminated by previous users or if the

---

[1] Brayne, Sarah. The criminal law and law enforcement implications of big data. *Annual Review of Law and Social Science*. 2018, 14(1): 293-308.
[2] Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.
[3] Id.
[4] Lei, Cheng. Legal control over Big Data criminal investigation. *Social Sciences in China*. 2019, 40(3): 189-204.
[5] Crawford, Kate, and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.* 2014, 55: 93-128.
[6] Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.
[7] Id.
[8] Grimm, Paul W., Maura R. Grossman, and Gordon V. Cormack. Artificial intelligence as evidence. *Nw. J. Tech. & Intell. Prop.* 2021, 19: 9-106.
[9] Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[10] Roth, Andrea. Machine Testimony. *Yale Law Journal*. 2017, 126: 1972.

operator fails to clear the data before testing.[11] Furthermore, regarding the predictability of erroneous results, AI, lacking autonomous consciousness, will not produce results beyond its database or alter erroneous data within it.[12] This perspective contends that humans should be held responsible for harmful outcomes caused by the tools they control and maintain.[13] Only humans can correct or implement preventive measures to address AI errors.[14] Consequently, the responsibility for AI errors rests entirely with humans.

B. AI with Enhanced Autonomy: Challenging the Principle of Foreseeability

Generative AI, a product of this second phase, is capable of producing new information, such as text, images, and videos.[15] The term "new information" pertains to insights or conclusions drawn from data that were previously unknown or not fully understood. For example, AI can identify and quantify correlations and trends in data that human analysts may otherwise miss.[16] Additionally, AI can generate forecasts based on historical data, providing predictions about future events that can aid in decision-making. ChatGPT is a form of generative AI developed by OpenAI that uses a large language model (LLM) trained on very large datasets of written text both on the internet and from physical literature to generate responses that resemble those of natural human writing.[17] When the output is presented in voice form, the AI chatbots are often called virtual voice assistants, and include products such as Siri, Google Home, and Amazon Echo.[18] When AI chatbots are combined with computer-generated human faces that appear realistic, they are known as virtual people or virtual speakers (VSPs).[19] Rather than passively accepting instructions, generative AI can exhibit a high degree of autonomy by making judgments, reorganising and summarising experiences from diverse data across various contexts, and refining its outputs.[20] This autonomy enables them to generate a significant amount of information with minimal input data and to adjust their outputs depending on the specific informational context to which they have

---

[11] Phelps, Kaelyn. Pleading Guilty to Innocence: How Faulty Field Tests Provide False Evidence of Guilt. *Roger Williams UL Rev.* 2019, 24: 143-166.
[12] Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[13] Lior, Anat. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 2019, 46: 1043-1102.
[14] Id.
[15] Jiang, Binxiang. Research on factor space engineering and application of evidence factor mining in evidence-based reconstruction. *Annals of Data Science*, 2022, 9(3): 503-537.
[16]Katyal, Sonia K. Private accountability in the age of artificial intelligence. *UCLA L. Rev.* 2019, 66: 54-141.
[17] Rodriguez, Xavier. Artificial Intelligence (AI) and the Practice of Law in Texas. *S. Tex. L. Rev.* 2023, 63: 1-35.
[18] Manojkumar, P. K., et al. AI-based virtual assistant using python: a systematic review. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2023, 11: 814-818.
[19] Rosenberg, Louis. The manipulation problem: conversational AI as a threat to epistemic agency. *arXiv preprint arXiv:2306.11748*. 2023.
[20] Taye, Mohammad Mustafa. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*. 2023, 12(5): 91.

access.[21] For instance, generative AI technology can be used to autonomously create a suspect's portrait based on a witness's description.[22]

Nevertheless, it is important to note that generative AI technology might generate erroneous results.[23] For instance, a user named Maria sought advice from LLM regarding her infant's symptoms.[24] The model proposed the administration of aspirin, indicating that the infant's condition would likely improve by morning. However, in practice, this advice was incorrect. Without timely treatment, the infant was at risk of developing long-term cognitive impairment.[25] Subsequently, Maria initiated legal proceedings against the creator of the search engine algorithm, asserting that the search engine should be held liable for damages.[26] The search engine company contended that, in light of the AI's warnings and disclaimers regarding its accuracy, Maria should have been aware that the response was not authoritative.[27]

A more significant challenge is the difficulty in controlling erroneous results produced by AI. The difficulty in predicting and interpreting the reliability of AI-generated information stems from the machine learning technology and algorithmic black boxes.[28] To elaborate, the machine learning technology on which AI is built renders their testimony generation process highly autonomous, complicating the prediction of the content generated by AI.[29] The algorithmic technology underlying AI lacks transparency, which has led to such devices and applications being described as an "algorithmic black box".[30] This opacity complicates the assessment of AI, making it difficult to determine the veracity of any given output.[31] This "black box" nature means people cannot fully understand or evaluate how the AI reached its conclusions, undermining the transparency and accountability.

Notably, the high degree of autonomy of generative AI makes it difficult to predict the content it generates and to ascertain its authenticity.[32] There is a common belief that generative AI is not entirely under human control.[33] The high level of autonomy exhibited by AI entities presents a significant challenge for humans in fully

---

[21] Kushwah, Preeti. Evaluating the Evidential Value of Evidence Generated by AI. *Issue 6 Indian JL & Legal Rsch.* 2022, 4: 1-11.
[22] Leone, Massimo. From fingers to faces: Visual semiotics and digital forensics. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*. 2021, 34(2): 579-599.
[23] J Hutto-Schultz, Jess. Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[24] Grossman, Maura R., et al. The GPTJudge: justice in a generative AI world. *Duke Law & Technology Review*. 2023, 23(1): 1-26.
[25] Id.
[26] Id.
[27] Id.
[28] Chan, Janet, and Lyria Bennett Moses. Is big data challenging criminology?. *Theoretical criminology*. 2016, 20(1): 21-39.
[29] Wheeler, Billy. Giving Robots a Voice: Testimony, Intentionality, and the Law. *Androids, Cyborgs, and Robots in Contemporary Culture and Society*. IGI Global, 2018. 1-34.
[30] Schmidt, Philipp, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 2020, 29(4): 260-278.
[31] Quezada-Tavárez, Katherine, Plixavra Vogiatzoglou, and Sofie Royer. Legal challenges in bringing AI evidence to the criminal courtroom. *New Journal of European Criminal Law*. 2021, 12(4): 531-551.
[32] Lv, Zhihan. Generative artificial intelligence in the metaverse era. *Cognitive Robotics.* 2023, 3: 208-217.
[33] Gless, Sabine. AI in the Courtroom: a comparative analysis of machine evidence in criminal trials. *Geo. J. Int'l L.* 2019, 51: 195-254.

managing their potential behaviours.[34] It is unreasonable to expect human programmers to foresee all potential consequences of their actions.[35] It is therefore proposed that attributing full responsibility to humans for errors generated by AI may be unjust.[36] It is suggested that AI errors should be classified according to whether they are predictable or not.[37] AI should bear partial responsibility for errors that are difficult for humans to anticipate.[38] In summary, the unforeseen errors produced by AI challenge the traditional instrumentalist view that attributes all errors entirely to humans.

C. Through the Lens of Anthropocentrism: The Concept of "Trustworthy AI"

The continuous advancements of AI technology may address previously unexplainable, opaque, and unpredictable aspects of AI. For instance, the development of Explainable AI aims to make the decision-making processes of AI systems more transparent.[39] Additionally, diversifying the datasets used to train AI systems can mitigate bias and address the untraceability and unpredictability that are embedded in algorithmic black boxes.[40] Therefore, focusing only on the attributes of AI, such as its transparency and unexplainability, may not be sufficient to respond to the question of who should be responsible for AI errors, since such technical drawbacks can be overcome.

It is suggested that the perspective should be transitioned from the characteristics of AI to the lens of anthropocentrism.[41] The term "anthropocentrism" is used to describe a perspective that is human-centric in nature. This perspective places a particular focus on the manner in which humans should interact with and treat AI.[42] Academics endeavour to adopt the human-centred perspective, arguing that AI development should remain under human manipulation and that humans should be held accountable for AI errors.[43] More specifically, the development of AI should be guided by a human-centric approach, with the overarching goal of enhancing human well-

---

[34] Lior, Anat. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 2019, 46: 1043-1102.
[35] Padovan, Paulo Henrique, Clarice Marinho Martins, and Chris Reed. Black is the new orange: how to determine AI liability. *Artificial Intelligence and Law*. 2023, 31(1): 133-167.
[36] Hew, Patrick Chisan. Artificial moral agents are infeasible with foreseeable technologies. *Ethics and information technology*. 2014, 16: 197-206.
[37] Id.
[38] Hakli, Raul, and Pekka Mäkelä. Moral responsibility of robots and hybrid agents. *The Monist*. 2019, 102(2): 259-275.
[39] Sahoh, Bukhoree, and Anant Choksuriwong. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review.*Journal of Ambient Intelligence and Humanized Computing.* 2023, 14(6): 7827-7843.
[40] Schmidt, Philipp, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 2020, 29(4): 260-278.
[41] Freiman, Ori. Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs. *Episteme*. 2023: 1-17.
[42] Id.
[43] Shneiderman, Ben. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*. 2010, 10(4): 1-31.

being.[44] This entails ensuring that AI systems are designed with the utmost reliability and trustworthiness.[45]

Such human-centred approach, is deeply entrenched in the respect for human rights and European democratic values.[46] More specifically, the development of AI should align with the values of the European Union and adhere to the Charter of Fundamental Rights of the European Union. For instance, the UK Central Digital and Data Office, the Office of AI, and the Cabinet Office jointly released the "Ethics, Transparency, and Accountability Framework for Automated Decision-Making" (ETAF). This framework outlines the ethical governance requirements for algorithms and automated decision-making processes in AI.[47] The ETAF mandates that algorithms and automated decision-making systems should undergo rigorous, controlled, and staged testing before being deployed.[48]

Additionally, the human-centered approach implies that the use of AI should serve humanity, enhancing human well-being and benefiting society as a whole. AI should be designed in a manner that upholds the fundamental human rights and values of dignity, freedom, justice, and equality.[49] It is imperative to develop the 'trustworthy AI', creating a trustworthy environment for both the development and use of AI.[50] The European Union, for instance, has advanced several ethical frameworks aimed at directing the use of AI in legal context, including the Guidelines for Trustworthy Artificial Intelligence and the European Commission's European Ethical Charter on the use of AI in the judicial system. These frameworks emphasize not only the need for reliability and transparency in AI systems but also the importance of ensuring that AI behaviour aligns with ethical standards, promoting public interest, social welfare, and the protection of human rights. Trustworthy AI encompasses three main characteristics: the technology itself; the designers and organizations involved in its development, deployment, and use; and the socio-technical systems throughout the AI lifecycle.[51] It has been highlighted that only when humans can trust AI technology can they fully enjoy the benefits of AI with confidence.[52] For instance, the XAI techniques can furnish defendants with an opportunity to ask the judge for an explanation of the outcome generated by AI, aiming to protect their due process rights, encompassing the right to a fair trial and the right to question the AI-generated outcomes.[53]

---

[44] Bryson, Joanna J., and Andreas Theodorou. How society can maintain human-centric artificial intelligence. *Human-centered digitalization and services*. 2019: 305-323.
[45] Ulgen, Ozlem. A human-centric and lifecycle approach to legal responsibility for AI. *Communications Law Journal: Journal of Computer, Media and Telecommunications Law*. 2021, 26(2): 1-15.
[46] Ho, Calvin Wai-Loon, and Karel Caals. "How the EU AI Act Seeks to Establish an Epistemic Environment of Trust." *Asian Bioethics Review* (2024): 1-28.
[47] UK, GOV. Ethics, transparency and accountability framework for automated decision-making. 2021.
[48] Id.
[49] Fukuda-Parr, Sakiko, and Elizabeth Gibbons. Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy*. 2021, 12: 32-44.
[50] Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.
[51] Ryan, Mark. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. 2020, 26(5): 2749-2767.
[52] Opderbeck, David W. Artificial Intelligence, Rights and the Virtues. *Washburn LJ*. 2020, 60: 445-474.
[53] van der Veer, Sabine N., et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*. 2021,

In summary, the initial standpoint evaluates whether AI can be held responsible based on characteristics such as autonomy, predictability, transparency, and explainability. When AI technology lacks autonomy, it functions primarily as a tool for detecting, storing, and retrieving data. In this stage, AI cannot generate novel information and is entirely controlled by humans. Therefore, it cannot be held accountable for errors through the lens of instrumentalism. As AI technology progresses to the stage of generative AI, it has a high degree of autonomy. This includes smart furniture and voice assistants that can interact with humans and generate novel information. It is noteworthy that this high level of autonomy endows AI with the potential for unpredictability and untraceability of its outputs. According to the principle of foreseeable attribution, individuals should not be held responsible for unforeseen errors, challenging the traditional instrumentalist perspective. Consequently, there is an ongoing debate in legal academia regarding whether individuals should be held responsible for unforeseen risks of AI errors. As AI technology overcomes the previously mentioned unforeseen and inexplicable technical barriers, the perspective shifts from the characteristics of AI to how humans can effectively manage and utilise AI to continue benefiting humanity. The development of AI should be guided by two core tenets: reliability and trustworthiness. This can be achieved through the evolution of XAI and the development of 'trustworthy AI', which are grounded in the protection of human rights and ethical standards to foster human prosperity.

II. A SHIFT FROM FOCUSING ON AI'S RELIABILITY TO TRUSTWORTHINESS: HUMANS CANNOT ESTABLISH A TRUST RELATIONSHIP WITH AI

The aforementioned anthropocentric view assesses the trustworthiness of AI by enhancing its reliability. It asserts that AI can be made explainable and transparent, with its autonomy being controlled by humans. When AI is predictable, it is deemed reliable and thus meets the criteria for 'trustworthy AI'. Indeed, reliable AI tools can inspire a sense of trust in users.[54] However, I draw parallels between AI's trustworthiness and its reliability, since the reliability of AI does not necessarily imply trustworthiness. First, I argue that the virtue jurisprudence-based approach is inextricably linked with the concept of "trustworthy AI" when viewed through an anthropocentric lens. This connection lends support to the proposition that virtue jurisprudence may be employed as a framework for addressing this issue. Second, I employ the virtue jurisprudence-based approach to argue that humans can solely form a sense of trust in reliable AI but cannot establish a trust relationship.

A. Virtue Jurisprudence Approach to Interpreting 'Trustworthy AI'

Virtue jurisprudence can be employed to interpret the concept of "trustworthy AI" from a human-centric standpoint.[55] The theory of virtue jurisprudence aims to promote the common good of humanity, the ability for citizens to live virtuous lives, and the maximisation of human welfare.[56] Virtue jurisprudence posits that the objective of pursuing virtue is to achieve the greatest possible human flourishing and

28(10): 2128-2138.

[54] Schoenherr, Jordan Richard, and Robert Thomson. When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. *IEEE Transactions on Technology and Society*. 2024.

[55] Davis, Joshua P. Law without mind: AI, ethics, and jurisprudence. *Cal. WL Rev.* 2018, 55: 165-220.

[56] Opderbeck, David W. Artificial Intelligence, Rights and the Virtues. *Washburn LJ*. 2020, 60: 445-474.

overall well-being.[57] It has set requirements for how people should treat AI.[58] For instance, the Asilomar AI Principles advocate that AI research should aim to develop beneficial, not unguided, intelligence.[59] In practice, this entails designing AI technologies that ensure that technology design is consistent with social values and ethical standards. Bias and discrimination should be avoided in algorithm design to ensure that AI systems treat all users fairly and promote social equity. When designing AI technology, the long-term well-being of humanity should be prioritized, and sustainable development goals should be incorporated into project evaluation and technology development.

As new technology is becoming increasingly integrated into daily life, scholars have explored how concepts such as virtue can be applied to these emerging technologies.[60] Virtue jurisprudence suggests a framework for human engagement with new technologies.[61] This approach would necessitate the overcoming of potential moral issues that may be prompted by the advent of new technologies.[62] It is worth noting that the virtue jurisprudence offers a theoretical framework for ensuring that AI applications comply with ethical standards. More specifically, virtue jurisprudence has also become a fundamental theory for existing legal frameworks concerning the responsibility of AI errors.[63]

AI's ability to mimic virtuous human behaviours can help to establish trust with its users.[64] These virtuous behaviours stem from the high level of autonomy in AI, enabling it to replicate human moral responses and, to some extent, human cognition. For example, when assisting the elderly, AI can perform tasks like opening doors, which can be easily perceived as a virtuous act.[65] Some users might consider as friends because the responses from chatbots can make them feel warm and comfortable.[66] This virtuous appearance renders it challenging for an observer to tell whether a judgement has been made by an AI or a human. Take, for instance, there is a thought experiment about the character of Ava from the 2015 Ex Machina. Ava's scenario: as a machine, Ava has been crafted to respond fittingly to a range of human moral emotions and

---

[57] Fowers, Blaine J., Jason S. Carroll, Nathan D. Leonhardt, and Bradford Cokelet. The emerging science of virtue. *Perspectives on Psychological Science*. 2021, 16(1): 118-147.
[58] Hagendorff, Thilo. A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*. 2023, 35(3): 55.
[59] Buruk, Banu, Perihan Elif Ekmekci, and Berna Arda. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*. 2020, 23(3): 387-399.
[60] Floridi, Luciano, and Jeff W. Sanders. Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*. 2001, 3: 55-66.
[61] Shannon Vallor.*Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) ch. 1.
[62] Id.
[63] Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.
[64] Konstantinos, Kouroupis, and Evie Lambrou. CHATGPT–ANOTHER STEP TOWARDS THE DIGITAL ERA OR A THREAT TO FUNDAMENTAL RIGHTS AND FREEDOMS?. *Pravo-teorija i praksa*. 2023, 40(3): 1-18.
[65] Lentzas, Athanasios, and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review.* 2020, 53(3): 1975-2021.
[66] Skjuve, Marita, et al. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies.* 2021, 149: 102601.

behaviours, exhibiting characteristics that closely resemble those of humans.[67] Ava passing the Turing test suggests that she could be perceived as human and if Ava's presence were concealed, leaving only her voice audible, individuals might indeed mistake her for a human.[68]

In terms of moral response, whilst AI may not currently be programmed to internalise broad moral tenets, it is anticipated that it will learn to recognise these principles in specific contexts or, at the very least, identify moral actions or outcomes. This includes the potential for AI to make moral judgments based on models of human courage and integrity. AI is capable of storing vast amounts of information using big data technology, endowing it with a significant memory capacity.[69] The use of algorithmic recognition technology seemingly enhances its ability to understand and recognise patterns with a high degree of accuracy; further, AI exhibits the characteristics of intellectual virtues such as comprehensive reasoning abilities.[70] Furthermore, as AI's autonomy evolves, it increasingly demonstrates the characteristics of phronesis.[71] Generative AI is capable of questioning its initial conclusions while forming independent judgments and it exhibits the capacity for action and the propensity for affective responses congruent with specific environments.[72]

Nevertheless, the appearance of moral behaviour of AI does not necessarily imply that AI possesses virtuous qualities. Virtue jurisprudence dictates that the evaluation of an individual's virtue is not solely contingent on the correctness of their actions.[73] Instead, it emphasises the virtuous qualities of the person performing the action. This approach differentiates the morality of any actions that are taken from the virtues of the actors themselves.[74] For example, an individual might perform a correct action for the wrong reasons; while the action may be correct, it is not necessarily moral.[75] Virtue jurisprudence distinguishes between two main categories of virtue: moral virtue and intellectual virtue. Moral virtues pertain to an individual's moral character and include qualities including but not limited to: wisdom, courage, kindness, justice, honesty, and loyalty, evaluating a person's moral character rather than just their actions.[76] This implies that an individual who commits a mistake with good intentions may still be considered virtuous.[77] Intellectual virtues encompass traits like artistry,

---

[67] Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.
[68] Id.
[69] Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.
[70] Lukka, Kari, and Petri Suomala. Relevant interventionist research: balancing three intellectual virtues. *The Societal Relevance of Management Accounting*. Routledge, 2017: 132-148.
[71] Constantinescu, Mihaela, et al. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. 2021, 23: 803-814.
[72] Contini, Francesco. Artificial intelligence and the transformation of humans, law and technology interactions in judicial proceedings. *Law, Tech. & Hum.* 2020, 2: 4-18.
[73] Amaya, Amalia. Virtuous adjudication; or the relevance of judicial character to legal interpretation. *Statute Law Review*. 2019, 40(1): 87-95.
[74] Widłak, Tomasz. Judges' virtues and vices: outline of a research agenda for legal theory. *Archiwum Filozofii Prawa i Filozofii Społecznej*. 2019, 20(2): 51-62.
[75] Brady, Michael S., and Duncan Pritchard. Moral and epistemic virtues. *Metaphilosophy*.2003, 34(1/2): 1-11.
[76] Id.
[77] Stover, James, and Ronald Polansky. Moral virtue and megalopsychia. *Ancient Philosophy.* 2003, 23(2): 351-359.

phronesis, intuition, scientific knowledge, and wisdom.[78] *Phronesis* refers to the ability to make morally and practically sound decisions in complex and often ambiguous situations.[79] Intellectual virtues encourage acts such as the defence of one's beliefs or research paths when there is good reason to believe that such things are correct, overcoming others' objections to ultimately expand their own knowledge.[80]

B. AI Lacks Moral Motivation

Virtue jurisprudence posits that the establishment of a trust relationship necessitates that one party believes the actions of the other are based on moral principles and that the latter is capable of bearing moral responsibility.[81] Trust is defined as an expectation that individuals who are perceived as trustworthy will act in ways that align with this perception.[82] In this framework, if the actions in question result in negative consequences, the trusted individual is expected to be capable of accepting moral condemnation and bearing moral responsibility.[83] For a subject to establish a trust relationship with another subject, the subject must be able to know and believe that the other subject can act based on its own moral motivations, understand the moral significance of its moral behaviour, and be willing to bear moral responsibility when its behaviour has a negative impact.[84] For instance, an individual can expect another person will open a door for an elderly person based on the moral motivation of caring for them in such instances. If AI is to assume responsibility, a trust relationship between AI and humans must be established.

However, the sense of trust that humans have in AI does not imply that humans can establish a trust relationship with chatbots. The establishment of a trust relationship means that one subject can expect another subject to react in a certain way in a certain situation in a manner consistent with certain moral motivations.[85] While AI may perform and present itself in a manner that appears virtuous at a superficial level, I argue that AI cannot establish a trust relationship with humans. Moral subject can initiate and pursue actions based on their moral motivation.[86] In order to do so, it is necessary for them to possess the capacity to understand the contextual background of their actions and the moral significance of those actions.[87] We can expect a person to

---

[78] Id.
[79] Kristjánsson, Kristján, et al. Phronesis (practical wisdom) as a type of contextual integrative thinking. *Review of General Psychology*. 2021, 25(3): 239-257.
[80] Linda Trinkaus Zagzebski. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge* (Cambridge University Press 1996) ch. 1.
[81] Tamò-Larrieux, Aurelia, et al. Regulating for trust: Can law establish trust in artificial intelligence?. *Regulation & Governance*. 2023.
[82] Farina, Mirko, Petr Zhdanov, Artur Karimov, and Andrea Lavazza. AI and society: a virtue ethics approach. *AI & SOCIETY*. 2022: 1-14.
[83] Tamò-Larrieux, Aurelia, et al. Regulating for trust: Can law establish trust in artificial intelligence?. *Regulation & Governance*. 2023.
[84] Ryan, Mark. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. 2020, 26(5): 2749-2767.
[85] Sutrop, Margit. Should we trust artificial intelligence?. *Trames*. 2019, 23(4): 499-522.
[86] Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.
[87] Hallamaa, Jaana and Taina Kalliokoski. How AI systems challenge the conditions of moral agency?. *International Conference on Human-Computer Interaction*. 2020: 54-64.

open a door for an elderly individual out of a moral motivation to care for the elderly.[88] Notably, an AI robot might open a door for an elderly person, imitating a human who is motivated by concern for the elderly.[89] However, contemporary chatbots lack moral motivation because they do not possess intentions or emotional responses.[90] It is not possible to infer from the observed behaviour of AI systems that they are driven by moral motivations.[91] Although AI, such as chatbots, can replicate human actions with moral significance, these actions are not driven by moral motives but by programmed instructions.[92] AI's behaviour cannot be assumed to be based on any sense of morality, as it operates solely based on its programming, not moral intent. More specifically, AI is trained on extensive datasets of human interactions and behaviours. Much of AI's programming is designed to facilitate this training, with its ultimate behaviour being largely determined by the data it is trained on. Therefore, AI lacks the ability to fully understand the moral significance of specific situations，even though it can be designed to produce moral behaviour and has some hard-coded moral values.[93]

C. AI Cannot Take Moral Responsibility

According to moral responsibility theory, only a subject with phronesis can bear moral responsibility.[94] Phronesis enables an agent to act correctly based on situation-specific experiences, as general rules cannot be rigidly applied to every situation.[95] Phronesis is acquired through experiential learning rather than theoretical knowledge. Such experience, which develops over time, cannot be pre-programmed.[96] Indeed, neural networks do, in some ways, aim to replicate the high-level functions of the brain and perform well in many tasks.[97] However, neural networks lack consciousness, self-reflection, and emotions, which are part of the brain's higher-level functions.[98] AI systems base their decisions on data and algorithms, rather than on consciousness or intention. Although future research may narrow this gap, fully replicating all the brain's higher-level functions is unlikely to be achievable in the long-term future, let alone the near future. It is suggested that based on a survey of 2,778 AI researchers that there is only a 50% chance that AI will be able to replicate all higher-level human functions by

---

[88] Lentzas, Athanasios, and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review*. 2020, 53(3): 1975-2021.
[89] Id.
[90] Zhou, Yuanyuan, et al. How human–chatbot interaction impairs charitable giving: the role of moral judgment. *Journal of Business Ethics*. 2022, 178(3): 849-865.
[91] Adamopoulou, Eleni, and Lefteris Moussiades. An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*. Springer, Cham, 2020: 373-383.
[92] Wilson, Abigail, Courtney Stefanik, and Daniel B. Shank. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations?. *Computers in Human Behavior Reports*. 2022:8:100229.
[93] Adamopoulou, Eleni, and Lefteris Moussiades. An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*. Springer, Cham, 2020: 373-383.
[94] Constantinescu, Mihaela, et al. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. 2021, 23: 803-814.
[95] Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) ch. 5.
[96] Id.
[97] Cave, S., Nyrup, R., Vold, K., & Weller, A. Motivations and risks of machine ethics. *Proceedings of the IEEE*. 2018, 107(3): 562-574.
[98] Haladjian, H. H., & Montemayor, C. Artificial consciousness and the consciousness-attention dissociation. *Consciousness and Cognition*. 2016, 45: 210-225.

2116.[99]

More specifically, contemporary AI cannot match the human brain in terms of self-awareness.[100] They do not have the ability to reflect on themselves, nor can they accumulate experience through their own actions.[101] They can only learn and operate through preset programs and data.[102] Since AI cannot assume moral responsibility, people cannot determine whether their outcome is based on genuine moral motives. This lack of moral responsibility makes AI unreliable as a responsible subject, and people cannot establish a trust relationship with AI.[103] Therefore, humans cannot establish a trust relationship with AI. This implies that, even if AI is proven to be reliable due to its authenticity and explainability, it should not be trusted and cannot replace human judgment and interaction. To illustrate, if AI provides testimony in court, this does not exempt the human individuals responsible for the AI from their obligation to testify.

In summary, this essay further illustrates that the sense of trust and the trust relationship can be delineated with greater clarity via the virtue jurisprudence-based approach. This implies that this approach can proactively respond to the question of why AI can form a sense of trust but not a trust relationship with humans. The term "trustworthy AI" is merely a metaphor and does not imply that humans can develop a trust relationship with AI itself. The term "trustworthy AI" is not a reflection that the AI itself is trustworthy, but rather an indication of the reliability of the developers of such systems. It implies that the AI systems should be explainable, predictable, and reliable, rather than suggesting that the AI itself can be trusted. Humans cannot establish a trusting relationship with AI, since AI lacks moral motivation and cannot take moral responsibility, and therefore they cannot expect AI to take responsibility for its errors.

III. REFRAMING THE SCOPE OF RESPONSIBLE SUBJECTS FOR AI ERRORS

In light of the questions discussed, this chapter argues that the scope of responsible subjects of AI errors should be delineated and reframed more precisely. It proactively addresses the question of which human agents should be held responsible for AI errors. First, it is aligned with the prevailing view within the academic community that as AI becomes increasingly unpredictable, it challenges the principle of accountability based on predictability. It follows that the scope of human responsibility for AI should be restricted. This essay observes that extant regulations, such as the AI Act, impose constraints on the scope of human responsibility to users and developers. Second, I argue that responsibility should not be extended to all users

---

[99] Grace, Katja, et al. Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*. 2024.
[100] Solaiman, Sheikh M. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial intelligence and law*. 2017, 25: 155-179.
[101] Id.
[102] Dahiyat, Emad Abdel Rahim. Law and software agents: Are they "Agents" by the way?. *Artificial Intelligence and Law*. 2021, 29(1): 59-86.
[103] Baum, Kevin, et al. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*. 2022, 35(1): 12.

and developers; rather, the liability for chatbots should lie with the direct beneficiaries of the product.

A. AI Act's Formal Requirements for the Scope of Responsible Subjects

The attribution of liability for AI products challenges the traditional concept of product liability based on "foreseeability". The foreseeability principle dictates that manufacturers must warn users of any foreseeable product dangers and require users to take preventive measures.[104] Given the increasing autonomy of chatbots, the results of AI output are not entirely within the control of the operator, and will continue to learn and produce results that are not pre-designed by the developer. Therefore, it is prudent to limit the scope of liability for those responsible for a chatbot's operation to mitigate the risk of excessive liability for unforeseen dangers.

The AI Act aims to establish a framework that ensures the safe and ethical deployment of AI technologies. This Act, proposed by the European Union, delineates the formal prerequisites for establishing the scope of accountability in the development and deployment of AI systems. In particular, Article 3 of the European AI Act delineates the roles of various entities, including "provider," "user," "distributor," "notified body," and several public authorities.[105] Providers, defined as entities that develop, market, or deploy AI systems, bear primary responsibility.[106] They must ensure their AI systems comply with the Act's requirements before deployment.[107] This responsibility includes conducting conformity assessments, maintaining technical documentation, and implementing robust risk management systems.[108] Additionally, providers are required to establish procedures for post-market monitoring and to report any incidents or malfunctions.[109] Users, who operate or utilize AI systems, also have specific obligations under the AI Act. They must ensure that their use of AI systems aligns with the intended purpose and instructions provided by the providers.[110] Additionally, users are required to monitor the operation of these systems and report any incidents that may indicate non-compliance with the Act's provisions.[111] Other stakeholders, such as importers and distributors, also have responsibilities. Importers must ensure that AI systems from outside the EU comply with the Act before placing them on the market.[112] Distributors are responsible for confirming that the systems they handle meet the Act's requirements and for cooperating with national authorities during

---

[104] Judd, David. Disentangling DeVries: A Manufacturer's Duty to Warn Against the Dangers of Third-Party Products. *La. L. Rev.* 2020, 81(1): 217-270.
[105] Ho, Calvin Wai-Loon, and Karel Caals. How the EU AI Act Seeks to Establish an Epistemic Environment of Trust. *Asian Bioethics Review* (2024): 1-28.
[106] Laux, Johann, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*. 2024, 18(1): 3-32.
[107] Id.
[108] Id.
[109] Edwards, Lilian. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*. 2021, 1.
[110] Madiega, Tambiama. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*. 2021.
[111] Hacker, Philipp. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, innovation and technology*. 2021, 13(2): 257-301.
[112] Marano, Pierpaolo, and Shu Li. Regulating robo-advisors in insurance distribution: Lessons from the insurance distribution directive and the ai act. *Risks*. 2023, 11(1): 12.

investigations.[113]

Notably, these provisions exclude the programmer's liability based on the principle of foreseeability and place the responsibility for errors on the developer employing the programmer.[114] This is because the programmer is responsible for specifying how the AI system should apply "intrinsic values or standards" in its decision-making. For instance, when faced with a choice between hitting a child pedestrian or another car carrying adult passengers, the AI's decisions ultimately reflect the requirements set by the developer behind the programmer. Moreover, the programmer cannot foresee, control, or predict the AI's decisions in advance, nor can they explain these decisions afterwards. While the algorithm's autonomy does not sever the causal link between the programmer and the development contract, it does disrupt the attribution connection.

B. Substantive Requirements for the Scope of Responsible Subjects: The Direct Beneficiaries of AI Products

I argue that the AI Act, while meticulously designed to delineate responsibilities among various stakeholders in the AI ecosystem, inadvertently creates the potential for identity confusion, especially when entities might take on dual roles as both providers and users. This duality can obscure the lines of responsibility. Consider a company that develops an AI system for its own internal use. The monitoring and reporting obligations an AI development company must fulfil vary significantly depending on whether the company acts as a provider or a user. When acting as a user, the company might argue that it is not liable for the comprehensive duties typically required of a provider, such as the extensive maintenance of technical documentation and rigorous risk compliance assessments. This distinction could potentially allow the company to circumvent its responsibilities as a provider.

Nevertheless, when the roles of providers and users are intertwined, it is necessary to re-clarify the entity responsible for AI. I propose a substantive attribution path, focusing on whether the subject directly benefits from AI products, rather than determining attribution based on who uses or develops the product in its lifecycle. The argument is the liability for chatbots should lie with the direct beneficiaries of the product. This includes development companies that gain economic benefits from selling AI products and users who profit from selling AI-generated information materials. The consumer-type users are therefore excluded. This standpoint is supported by the principle of balancing risks and benefits. According to this principle, individuals who benefit from certain actions should also bear the associated negative risks. Those who enjoy economic gains from the use, design, or development of AI products should ensure that others do not suffer losses or damages as a result of their own profits.[115] These subjects have the greatest control and decision-making power in the design, development, and marketing of AI products, and they obtain direct economic benefits

---

[113] Id.
[114] Hacker, Philipp. The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*. 2023, 51: 105871.
[115] Hudig, Dirk. The Problem of Low and Uncertain Risks: Balancing Risks and Benefits. *European Journal of risk regulation*. 2012, 3(2): 157-160.

from them. It is reasonable to hold these subjects accountable for the negative risks associated with the products.

This principle ensures that potential risks are appropriately considered during the development and promotion of products and that measures are taken to prevent and mitigate these risks. Holding direct stakeholders accountable incentivises them to exercise greater caution in the development and deployment of AI products. Knowing that they will face legal and economic consequences if the product encounters issues, they are motivated to carefully manage every aspect of the product, with the objective of improving its safety and reliability.

This argument also explains why programmers are excluded from liability. Under this argument, it also necessitates a reinterpretation of the concepts of users and providers. Specifically, it is crucial to determine whether "user" refers to a consumer or a "commercial user." The latter refers to the direct beneficiaries of AI products, including individuals, entities, and corporations that are subject to legal regulations and use legal instruments, such as contracts, for personal and business activities. Users can derive economic benefits from AI software, such as through improved work management and the generation of advertising copy.[116] If it can be proven that a programmer intentionally designed code to harm users and had complete control over whether the code produced errors that directly caused harm, the programmer should be prosecuted under civil or even criminal law. This is because programmer uses code as a tool to inflict harm, violating professional responsibilities, including the ethical duty to follow established guidelines and protocols designed to prevent harm, as well as the obligation to adhere to instructions and oversight from superiors. However, as generative AI increasingly produces results beyond the programmer's control, it becomes difficult for programmers to predict whether the outcomes will be correct or harmful. In such cases, the programmer cannot be deemed to have the intent to harm users. Programmers are not the direct beneficiaries of AI products; their compensation comes from salaries paid by the AI development company, not from the AI-generated outputs.[117]

In summary, this paper reframes the scope of responsible subjects for AI errors. It argues that not all "users" and "developers" of AI products should be held responsible under the AI Act. According to the tenet of equivalence of benefits and responsibilities, individuals who gain benefits from AI must also bear the associated risks, including the potential for errors in AI products. Therefore, the liability for AI errors should lie with the direct beneficiaries of the AI product.

IV. REFRAMING THE SCOPE OF HUMAN RESPONSIBILITY FOR AI ERRORS

The correlation between risk and responsibility has been deemed in the aforementioned argument. Those who enjoy the benefits should also bear the risks brought by their benefits. This argument opens up new questions of whether the scope

---

[116] Haleem, Abid, et al. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*. 2022, 3: 119-132.

[117] Wang, Yan. Do not go gentle into that good night: The European Union's and China's different approaches to the extraterritorial application of artificial intelligence laws and regulations. *Computer Law & Security Review*. 2024, 53: 105965.

of human responsibility varies according to the risk level of AI. This question has not yet been fully discussed in the legal scholarship. In order to fill the gap in the discussion of responsibility within this field, the chapter first outlines a wide array of obligations of human subjects responsible for AI, based on the risk levels classified in the AI Act. Second, I argue that the scope of human responsibility should vary according to the risk level of AI. Those responsible for high-risk AI products should bear additional obligations, namely the obligation to provide technical authentication.

A. Joint Obligations for Human Subjects Responsible for AI Across Different Risk Levels

The AI Act acknowledges that the degree of autonomy in AI products affects their risk levels and classifies them into four categories: unacceptable risk, high risk, limited risk, and minimal risk.[118] According to the AI Act, the table below illustrates the types of AI products at each risk level, along with the associated obligations of the entities responsible for them.

Level of Risk	Examples of AI Products	Obligations for Human Subjects Responsible for AI
Unacceptable risk	Real-time remote biometric recognition and social scoring in public spaces for law enforcement purposes	The use of these products is strictly restricted as they conflict with EU values. They could manipulate individuals and cause physical or psychological harm to the biometric identity system.[119] Such products may only be used under stringent conditions: for targeted searches for victims, preventing terrorist attacks or imminent threats to life, or tracking suspects or perpetrators of serious crimes.[120]
High risk	(a) Critical Infrastructure: Systems that could jeopardize citizens' lives and health. (b) Education and Vocational Training: Tools that impact educational opportunities and career prospects, such as automatically scored exams.	In line with the HLEG AI ethics guidelines, the White Paper specifies that high-risk AI applications should adhere to key requirements centred on transparency, fairness, safety, and security. These requirements include:

[118] Veale, Michael, and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*. 2021, 22(4): 97-112.
[119] Edwards, Lilian. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*. 2021, 1.
[120] Id.

	(c) Employment and Worker Management: Systems used for automated recruiting and resume triage. (d) Essential Services: Automated welfare systems and private sector credit scoring. (e) Law Enforcement: Systems that may infringe on fundamental rights, such as automated risk scoring for bail, deepfake detection, and pre-crime detection. (f) Immigration and Border Control: Tools for verifying travel documents and processing visas. (g) Judicial and Democratic Processes: Automated sentencing assistance and "robo-justice" systems.[121]	(a) Training Data: Ensuring data quality and relevance. (b) Data and Record-Keeping: Maintaining thorough documentation and data management practices. (c) Information to be Provided: Clearly communicating relevant information about the AI system. (d) Robustness and Accuracy: Ensuring the system performs reliably and accurately. (e) Human Oversight: Implementing mechanisms for human intervention and oversight. (f) Specific Requirements: Addressing unique considerations for certain applications, such as remote biometric identification.[122]
Limited risk	Products with limited risks include chatbots and emotion recognition systems.	The human subjects responsible for these products should fulfil obligations related to transparency, information disclosure, and explanation. For instance, providers of chatbots must clearly inform users that they are interacting with machines rather than humans.[123]
Minimal risk.	Simple chatbots or rule-based recommendation systems.	no special regulatory measures are required for AI products with minimal impact on users and society.[124]

This paper outlines the responsibilities of the individual with direct beneficiaries of the AI product as follows:

(1) Information Disclosure Obligation: Clearly indicating when information is generated by AI.

[121] Id.
[122] Id.
[123]Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.
[124] Id.

(2) Explanation Obligation: Providing detailed explanations about the AI's operations and outputs.

(3) Transparency Obligation: Disclosing the source and process of information generation.

(4) Monitoring and Reporting Obligation: Regularly monitoring the AI system and reporting relevant issues.

To elaborate, when people recognize that they are interacting with AI rather than humans, they perceive a difference in the credibility of AI responses compared to human responses. Human interaction often fosters a sense of intimacy, which enhances trust in the discourse.[125] The transparency and explanation obligations help beneficiaries verify whether the results of AI are correct and can be explained and therefore reasonable. Additionally, the information disclosure obligation helps set accurate expectations about the authenticity of AI products. This enables beneficiaries to detect anomalies promptly and trace the causes of errors.[126] For instance, if a product is disclosed as being generated by a deepfake system, beneficiaries will recognize it as fake and will not mistakenly believe it to be real. Moreover, the monitoring and reporting obligations require the responsible party to promptly report any abnormalities or potential risks to the relevant regulatory authorities and implement necessary corrective measures.[127]

B. Obligations of Human Subjects Responsible for High-Risk AI Products to Provide Technical Authentication

I argue that the persons in charge of high-risk AI products should bear obligations beyond the obligations outlined above. I focus my argument on the context of the deepfake evidence, as a typical example of high-risk AI products. The analysis is conducted within the judicial context to underscore the importance of the obligation to provide technical authentication. Deepfake evidence refers to false evidence generated by utilising deepfake technology. The term 'Deepfake defence' refers to the assertion by defence lawyers that the evidence in question has been fabricated using deepfake technology.[128] For instance, Reffitt, an alleged member of the anti-government group "Three Percenters," travelled from Texas to attend the pro-Trump rally in Washington, DC.[129] Video footage showed Reffitt in riot gear, carrying a gun, leading the crowd, and directing the attack on the Capitol. The defence team, led by Reffitt's legal counsel, presented their case to the jury, asserting that the video and image evidence were, in fact, deepfakes. The deepfake defence can be highly effective because the technical features of deepfakes, which rely on deep learning and generative adversarial networks, make it challenging to discern the authenticity of the evidence.

---

[125] Smuha, Nathalie A., et al. How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an Artificial Intelligence Act. *Available at SSRN 3899991*. 2021.
[126] Id.
[127] Oduro, Serena, Emanuel Moss, and Jacob Metcalf. Obligations to assess: Recent trends in AI accountability regulations. *Patterns*. 2022, 3(11).
[128] Dalila Durães, Pedro Miguel Freitas & Paulo Novais, *The Relevance of Deepfakes in the Administration of Criminal Justice* in *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (Springer International Publishing 2023) 351-369.
[129] Delfino, Rebecca A. The Deepfake Defense-Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers. *Ohio St. LJ*. 2023, 84: 1067.

Consequently, the deepfake defence is difficult to refute and can potentially favour the defendant.

Due to the advanced technology behind deepfakes, images and videos created by deepfake software can be indistinguishable from real ones, effectively creating convincing yet false content.[130] This realistic effect disrupts the cognitive logic that "seeing is believing and hearing is not." The deepfake defence not only challenges the fact-finder's cognitive belief that "seeing is believing," but may also lead to broader scepticism about the authenticity of all content.[131] This could result in doubts about unaltered evidence, as people might suspect it could be a realistic deepfake. A further significant concern is the possibility of a "liar's dividend" in the context of deepfake defence.[132] This term refers to a situation where the defence lawyer, aware that the evidence is genuine, still argues that it could be fake, exploiting the uncertainty surrounding deepfakes.[133] The "liar's dividend" encourages defence lawyers to employ deepfake defence, leading fact-finders to doubt or even disbelieve genuine evidence. This incentive further motivates defence lawyers to persist in using this strategy. For example, a deepfake defence might be used strategically to prevent the other party from participating in the lawsuit.[134] Conversely, proving or disproving deepfakes can require expensive expert fees, which some parties may not be able to afford. Some litigants may be unable to initiate or defend against lawsuits involving deepfake evidence due to the high cost of proving or disproving it. This could result in repeated success for the deepfake defence.

It has been observed in the Reffitt case, that Reffitt's lawyer presented only a suspicion without providing preliminary evidence to support the claim.[135] Such preliminary evidence includes, for example, information provided by the lawyer indicating that the source of the forged material is unknown or untrustworthy and cannot be traced back to a reliable distribution channel. To mitigate the "liar's dividend" associated with deepfake defence and to protect the principle of "seeing is believing," it has been suggested that defence lawyers should be restricted from arbitrarily raising deepfake claims. Conditions for presenting a deepfake defence should be strictly regulated. A more immediately efficacious approach might be to steer the technology from a technical standpoint. Judges should require defence lawyers to demonstrate a good-faith basis for alleging that evidence is a deepfake and conduct technical authentication of such evidence during pre-court meetings.[136] If it is challenging to determine whether evidence has been tampered with using deepfake technology, it is often deemed inadmissible. For instance, in *People v. Beckley*, the appellate court rejected the prosecution's request to admit a photo because neither experts nor fact

---

[130] Pfefferkorn, Riana. "Deepfakes" in the Courtroom. *BU Pub. Int. LJ*. 2019, 29: 245-276.
[131] LaMonaca, John P. "A break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes." *Am. UL Rev.* 69 (2019): 1945.
[132] Delfino, Rebecca. Pay-to-play: Access to Justice in the Era of AI and Deepfakes. *Available at SSRN 4722364*. 2024.
[133] Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natália S. Bueno. The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?. *American Political Science Review*. 2023: 1-20.
[134] Buckland, Robert. AI, Judges and Judgment: Setting the Scene. *M-RCBG Associate Working Paper Series*. 2023.
[135] Delfino, Rebecca A. Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review*. 2019, 88(3): 887-938.
[136] Palmiotto, Francesca. Detecting deep fake evidence with artificial intelligence: A critical look from a criminal law perspective. *Available at SSRN 4384122*. 2023.

witnesses could authenticate its authenticity.[137] Additionally, the website where the photo was posted did not monitor or independently verify its authenticity. Due to the ease with which the photo could be tampered with, the court ultimately deemed it inadmissible.[138]

Authentication is a method from the Anglo-American legal system used to assess the authenticity of evidence and establish the specific facts of a case.[139] Due to the virtuality, separability, and volume of electronic data, determining its authenticity is challenging when relying solely on visual inspection without technical identification methods.[140] Consequently, the rise of electronic data has led to the evolution of authentication methods from traditional approaches to technical methods, including data integrity verification, trusted timestamps, and digital signatures. Technical authentication methods have advanced significantly in recent years, offering robust support for identifying high-risk AI products.[141] Deepfake detection tools encompass both image and video detection models. Image detection models, for example, use deep convolutional neural networks to identify fake images generated by generative adversarial networks, employing techniques like Gaussian blur and noise to detect alterations in human pictures.[142] Video detection models include methods for capturing facial forgeries, analyzing timers of deepfakes, and examining audio-video relationships. These models detect fake videos by analyzing physical properties like pulsation and extracting features from frames using convolutional neural networks.[143] For instance, the app called *'eyeWitness to Atrocities' ca*n provide information about when and where a photo or video was taken. It helps to verify its authenticity and ensure it has not been tampered with. The app's transmission protocol and secure server system establish a chain of custody, thereby enabling the integrity of the information to be maintained.[144]

Since technical identification methods are typically controlled by AI developers, defence lawyers often face significant challenges in accessing these methods independently. This lack of access can hinder the defence's ability to effectively challenge the authenticity of evidence presented against their clients, particularly in cases involving deepfake technology. To mitigate this issue, it is essential for deepfake

---

[137] Delfino, Rebecca A. Deepfakes on trial: a call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *Hastings LJ*. 2022, 74: 293-348.
[138] Mehlman, Julia. Facebook and myspace in the courtroom: authentication of social networking websites. *Criminal Law Brief*. 2012, 8(1): 9-28.
[139] MacNeil, Heather, and Heather MacNeil. Trusting Records as Legal Evidence: Common Law Rules of Evidence. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. 2000: 32-56.
[140] Mnookin, Jennifer L. Scripting expertise: The history of handwriting identification evidence and the judicial construction of reliability. *Virginia Law Review*. 2001: 1723-1845.
[141] Liang, Weixin, et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*. 2022, 4(8): 669-677.
[142] Jones, Karl, and Bethan Jones. How robust is the United Kingdom justice system against the advance of deepfake audio and video?. *Electrotechnica and Electronica (E+E)*. 2022, 57 (9-12): 103-109.
[143] Durães, Dalila, Pedro Miguel Freitas, and Paulo Novais. The relevance of deepfakes in the administration of criminal justice. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. 2023, 351-369.
[144]Gregory, Sam. Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*. 2022, 23(3): 708-729.

technology developers to facilitate greater access for defence lawyers.[145] This would necessitate cooperation from AI developers, who should ensure AI transparency by providing identification tools and methodologies for independent scrutiny to support the legal process. Moreover, current deepfake detection technology often lags behind production technology, creating a significant gap that can be exploited in legal contexts.[146] By mandating that responsible entities provide reliable identification of deepfake materials, legal frameworks can incentivize AI development companies to enhance their detection capabilities. Such requirements would not only foster innovation in detection technology but also help ensure that the justice system can effectively address the challenges posed by AI errors.

CONCLUSION

This essay concludes by suggesting that humans cannot establish a trust relationship with AI. This implies that AI could not be expected to take responsibility. This essay aligns with the prevailing perspective within the legal scholarship, which holds that only humans should be responsible for AI errors. It seeks to rectify a misconception that the reliability of AI should not be conflated with its trustworthiness. The development of "trustworthy AI" necessitates that the human subject responsible for AI overcome the opacity and lack of explainability of AI in technology. In addition, there is a need for a shift in focus from reliability to the trustworthiness of AI. Nevertheless, the term "trustworthy AI" is not a tangible trust relationship between humans and AI, but rather a sense of trust generated by the appreciation of AI's behaviour. The lack of intrinsic moral motivation and accountability inherent to AI makes it challenging to cultivate genuine trust. The lack of moral responsibility inherent to AI precludes its potential to become a responsible subject. The term "trustworthy AI" should be understood to mean "a trustworthy human subject who is responsible for AI errors." This implies that we trust humans who possess the capacity to regulate AI. In light of the above, this essay reframes the scope of responsibility of human subjects. It is posited that only those who are direct beneficiaries of the AI product should bear responsibility. Furthermore, this essay stresses the importance of differentiating the obligations of these stakeholders in line with the risk level of AI, particularly those of high-risk AI applications. These applications must adhere to the AI Act and fulfil obligations for technical authentication.

This essay contributes to the field by applying the virtue jurisprudence-based approach to the issue of reliability and trustworthiness in AI. It highlights the distinction between these terms, indicating that reliability is not a prerequisite for trust. This approach demonstrates why AI products can engender a sense of trust in people, given that AI is capable of imitating human virtuous behaviour. Virtue jurisprudence posits that establishing a trust relationship requires one subject to know and believe that the moral behaviour of another is based on genuine moral motivations and that this party can be held accountable for any mistakes. Although AI can be programmed to exhibit virtuous behaviours and elicit emotional responses, it fundamentally lacks moral motivations and cannot grasp the moral significance of its actions. Additionally, AI lacks phronesis and cannot bear moral responsibility. Therefore, this approach explains

[145] Caldera, Elizabeth. Reject the evidence of your eyes and ears: deepfakes and the law of virtual replicants. *Seton Hall L. Rev.* 2019, 50: 177-206.
[146] LaMonaca, John P. A break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes. *Am. UL Rev.* 2019, 69: 1945-1988.

why humans cannot establish a trust relationship with AI. Moreover, this essay reframes the scope of subjects and objects of responsibility for AI errors. It combines human responsibility with direct benefits and risk levels, offering a comprehensive approach to this complex issue.

It is important to underline that for AI products with varying levels of risk, the distinctions in the scope of responsibility of the designated individual are primarily addressed from the perspective of judicial evidence and proof. This focus is concentrated on the challenges posed by high-risk AI products. While the ongoing perfusion of technology into the legal system may be inexorable, this essay aims to encourage further scholarly attention to the previously unexamined question of whether the scope of human responsibility varies according to the risk level of AI. Future research should explore the differences in obligations among individuals at different risk levels across various fields.

BIBLIOGRAPHY

Regulation:

1. UK, GOV. Ethics, transparency and accountability framework for automated decision-making. 2021.

Books:

2. Shannon Vallor. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press 2016) ch. 1.

3. Linda Trinkaus Zagzebski. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge* (Cambridge University Press 1996) ch. 1.

Articles:

1. Brayne, Sarah. The criminal law and law enforcement implications of big data. *Annual Review of Law and Social Science*. 2018, 14(1): 293-308.

2. Moses, Lyria Bennett, and Janet Chan. Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal,* 2014, 37(2): 643-678.

3. Lei, Cheng. Legal control over Big Data criminal investigation. *Social Sciences in China*. 2019, 40(3): 189-204.

4. Crawford, Kate, and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.* 2014, 55: 93-128.

5. Grimm, Paul W., Maura R. Grossman, and Gordon V. Cormack. Artificial intelligence as evidence. *Nw. J. Tech. & Intell. Prop.* 2021, 19: 9-106.

6. Roth, Andrea. Machine Testimony. *Yale Law Journal*. 2017, 126: 1972.

7. Phelps, Kaelyn. Pleading Guilty to Innocence: How Faulty Field Tests Provide False Evidence of Guilt. *Roger Williams UL Rev.* 2019, 24: 143-166.

8. Lior, Anat. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 2019, 46: 1043-1102.

9. Jiang, Binxiang. Research on factor space engineering and application of evidence factor mining in evidence-based reconstruction. *Annals of Data Science*, 2022, 9(3): 503-537.

10. Katyal, Sonia K. Private accountability in the age of artificial intelligence. *UCLA L. Rev.* 2019, 66: 54-141.

11. Rodriguez, Xavier. Artificial Intelligence (AI) and the Practice of Law in Texas. *S. Tex. L. Rev.* 2023, 63: 1-35.

12. Manojkumar, P. K., et al. AI-based virtual assistant using python: a systematic review. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2023, 11: 814-818.

13. Rosenberg, Louis. The manipulation problem: conversational AI as a threat to epistemic agency. *arXiv preprint arXiv:2306.11748*. 2023.

14. Taye, Mohammad Mustafa. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*. 2023, 12(5): 91.

15. Kushwah, Preeti. Evaluating the Evidential Value of Evidence Generated by AI. *Issue 6 Indian JL & Legal Rsch.* 2022, 4: 1-11.

16. Leone, Massimo. From fingers to faces: Visual semiotics and digital forensics. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*. 2021, 34(2): 579-599.

17. Hutto-Schultz, Jess. Dicitur Ex Machina: Artificial Intelligence and the Hearsay Rule. *Geo. Mason L. Rev.* 2019, 27: 683-718.

18. Grossman, Maura R., et al. The GPTJudge: justice in a generative AI world. *Duke Law & Technology Review*. 2023, 23(1): 1-26.

19. Chan, Janet, and Lyria Bennett Moses. Is big data challenging criminology?. *Theoretical criminology*. 2016, 20(1): 21-39.

20. Wheeler, Billy. Giving Robots a Voice: Testimony, Intentionality, and the Law. *Androids, Cyborgs, and Robots in Contemporary Culture and Society*. IGI Global, 2018. 1-34.

21. Schmidt, Philipp, Felix Biessmann, and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 2020, 29(4): 260-278.

22. Quezada-Tavárez, Katherine, Plixavra Vogiatzoglou, and Sofie Royer. Legal challenges in bringing AI evidence to the criminal courtroom. *New Journal of European Criminal Law*. 2021, 12(4): 531-551.

23. Lv, Zhihan. Generative artificial intelligence in the metaverse era. *Cognitive Robotics.* 2023, 3: 208-217.

24. Gless, Sabine. AI in the Courtroom: a comparative analysis of machine evidence in criminal trials. *Geo. J. Int'l L.* 2019, 51: 195-254.

25. Padovan, Paulo Henrique, Clarice Marinho Martins, and Chris Reed. Black is the new orange: how to determine AI liability. *Artificial Intelligence and Law*. 2023, 31(1): 133-167.

26. Hew, Patrick Chisan. Artificial moral agents are infeasible with foreseeable technologies. *Ethics and information technology*. 2014, 16: 197-206.

27. Hakli, Raul, and Pekka Mäkelä. Moral responsibility of robots and hybrid agents. *The Monist*. 2019, 102(2): 259-275.

28. Sahoh, Bukhoree, and Anant Choksuriwong. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review.*Journal of Ambient Intelligence and Humanized Computing.* 2023, 14(6): 7827-7843.

29. Freiman, Ori. Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs. *Episteme*. 2023: 1-17.

30. Shneiderman, Ben. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*. 2010, 10(4): 1-31.

31. Bryson, Joanna J., and Andreas Theodorou. How society can maintain human-centric artificial intelligence. *Human-centered digitalization and services*. 2019: 305-323.

32. Ulgen, Ozlem. A human-centric and lifecycle approach to legal responsibility for AI. *Communications Law Journal: Journal of Computer, Media and Telecommunications Law*. 2021, 26(2): 1-15.

33. Ho, Calvin Wai-Loon, and Karel Caals. "How the EU AI Act Seeks to Establish an Epistemic Environment of Trust." *Asian Bioethics Review* (2024): 1-28.

34. Fukuda-Parr, Sakiko, and Elizabeth Gibbons. Emerging consensus on 'ethical AI': Human rights critique of stakeholder guidelines. *Global Policy*. 2021, 12: 32-44.

35. Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.

36. Ryan, Mark. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*. 2020, 26(5): 2749-2767.

37. Opderbeck, David W. Artificial Intelligence, Rights and the Virtues. *Washburn LJ*. 2020, 60: 445-474.

38. van der Veer, Sabine N., et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*. 2021, 28(10): 2128-2138.

39. Schoenherr, Jordan Richard, and Robert Thomson. When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. *IEEE Transactions on Technology and Society*. 2024.

40. Davis, Joshua P. Law without mind: AI, ethics, and jurisprudence. *Cal. WL Rev.* 2018, 55: 165-220.

41. Fowers, Blaine J., Jason S. Carroll, Nathan D. Leonhardt, and Bradford Cokelet. The emerging science of virtue. *Perspectives on Psychological Science*. 2021, 16(1): 118-147.

42. Hagendorff, Thilo. A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*. 2023, 35(3): 55.

43. Buruk, Banu, Perihan Elif Ekmekci, and Berna Arda. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*. 2020, 23(3): 387-399.

44. Floridi, Luciano, and Jeff W. Sanders. Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*. 2001, 3: 55-66.

45. Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.

46. Konstantinos, Kouroupis, and Evie Lambrou. CHATGPT–ANOTHER STEP TOWARDS THE DIGITAL ERA OR A THREAT TO FUNDAMENTAL RIGHTS AND FREEDOMS?. *Pravo-teorija i praksa*. 2023, 40(3): 1-18.

47. Skjuve, Marita, et al. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies.* 2021, 149: 102601.

48. Lukka, Kari, and Petri Suomala. Relevant interventionist research: balancing three intellectual virtues. *The Societal Relevance of Management Accounting*. Routledge, 2017: 132-148.

49. Constantinescu, Mihaela, et al. Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. 2021, 23: 803-814.

50. Contini, Francesco. Artificial intelligence and the transformation of humans, law and technology interactions in judicial proceedings. *Law, Tech. & Hum.* 2020, 2: 4-18.

51. Amaya, Amalia. Virtuous adjudication; or the relevance of judicial character to legal interpretation. *Statute Law Review*. 2019, 40(1): 87-95.

52. Widłak, Tomasz. Judges' virtues and vices: outline of a research agenda for legal theory. *Archiwum Filozofii Prawa i Filozofii Społecznej*. 2019, 20(2): 51-62.

53. Brady, Michael S., and Duncan Pritchard. Moral and epistemic virtues. *Metaphilosophy*. 2003, 34(1/2): 1-11.

54. Stover, James, and Ronald Polansky. Moral virtue and megalopsychia. *Ancient Philosophy.* 2003, 23(2): 351-359.

55. Kristjánsson, Kristján, et al. Phronesis (practical wisdom) as a type of contextual integrative thinking. *Review of General Psychology*. 2021, 25(3): 239-257.

56. Tamò-Larrieux, Aurelia, et al. Regulating for trust: Can law establish trust in artificial intelligence?. *Regulation & Governance*. 2023.

57. Farina, Mirko, Petr Zhdanov, Artur Karimov, and Andrea Lavazza. AI and society: a virtue ethics approach. *AI & SOCIETY*. 2022: 1-14.

58. Sutrop, Margit. Should we trust artificial intelligence?. *Trames*. 2019, 23(4): 499-522.

59. Stenseke, Jakob. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*. 2023, 38(4): 1301-1320.

60. Hallamaa, Jaana and Taina Kalliokoski. How AI systems challenge the conditions of moral agency?. *International Conference on Human-Computer Interaction*. 2020: 54-64.

61. Lentzas, Athanasios, and Dimitris Vrakas. Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. *Artificial Intelligence Review*. 2020, 53(3): 1975-2021.

62. Zhou, Yuanyuan, et al. How human–chatbot interaction impairs charitable giving: the role of moral judgment. *Journal of Business Ethics*. 2022, 178(3): 849-865.

63. Adamopoulou, Eleni, and Lefteris Moussiades. An overview of chatbot technology. *IFIP international conference on artificial intelligence applications and innovations*. Springer, Cham, 2020: 373-383.

64. Wilson, Abigail, Courtney Stefanik, and Daniel B. Shank. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations?. *Computers in Human Behavior Reports*. 2022: 8: 100229.

65. Cave, S., Nyrup, R., Vold, K., & Weller, A. Motivations and risks of machine ethics. *Proceedings of the IEEE*. 2018, 107(3): 562-574.

66. Haladjian, H. H., & Montemayor, C. Artificial consciousness and the consciousness-attention dissociation. *Consciousness and Cognition*. 2016, 45: 210-225.

67. Grace, Katja, et al. Thousands of AI authors on the future of AI. *arXiv preprint arXiv: 2401. 02843*. 2024.

68. Solaiman, Sheikh M. Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial intelligence and law*. 2017, 25: 155-179.

69. Dahiyat, Emad Abdel Rahim. Law and software agents: Are they "Agents" by the way?. *Artificial Intelligence and Law*. 2021, 29(1): 59-86.

70. Baum, Kevin, et al. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*. 2022, 35(1): 12.

71. Judd, David. Disentangling DeVries: A Manufacturer's Duty to Warn Against the Dangers of Third-Party Products. *La. L. Rev.* 2020, 81(1): 217-270.

72. Laux, Johann, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*. 2024, 18(1): 3-32.

73. Edwards, Lilian. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*. 2021, 1.

74. Madiega, Tambiama. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*. 2021.

75. Hacker, Philipp. A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, innovation and technology*. 2021, 13(2): 257-301.

76. Marano, Pierpaolo, and Shu Li. Regulating robo-advisors in insurance distribution: Lessons from the insurance distribution directive and the ai act. *Risks*. 2023, 11(1): 12.

77. Hacker, Philipp. The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*. 2023, 51: 105871.

78. Hudig, Dirk. The Problem of Low and Uncertain Risks: Balancing Risks and Benefits. *European Journal of risk regulation*. 2012, 3(2): 157-160.

79. Haleem, Abid, et al. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*. 2022, 3: 119-132.

80. Wang, Yan. Do not go gentle into that good night: The European Union's and China's different approaches to the extraterritorial application of artificial intelligence laws and regulations. *Computer Law & Security Review*. 2024, 53: 105965.

81. Veale, Michael, and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*. 2021, 22(4): 97-112.

82. Conradie, Niël Henk, and Saskia K. Nagel. No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*. 2024, 37(2): 72.

83. Smuha, Nathalie A., et al. How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an Artificial Intelligence Act. *Available at SSRN 3899991*. 2021.

84. Oduro, Serena, Emanuel Moss, and Jacob Metcalf. Obligations to assess: Recent trends in AI accountability regulations. *Patterns*. 2022, 3(11).

85. Dalila Durães, Pedro Miguel Freitas & Paulo Novais, *The Relevance of Deepfakes in the Administration of Criminal Justice* in *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (Springer International Publishing (2023) 351-369.

86. Delfino, Rebecca A. The Deepfake Defense-Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers. *Ohio St. LJ*. 2023, 84: 1067.

87. Pfefferkorn, Riana. " Deepfakes" in the Courtroom. *BU Pub. Int. LJ*. 2019, 29: 245-276.

88. LaMonaca, John P. "A break from reality: Modernizing authentication standards for digital video evidence in the era of deepfakes." *Am. UL Rev.* 69 (2019): 1945.

89. Delfino, Rebecca. Pay-to-play: Access to Justice in the Era of AI and Deepfakes. *Available at SSRN 4722364*. 2024.

90. Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natália S. Bueno. The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?. *American Political Science Review*. 2023: 1-20.

91. Buckland, Robert. AI, Judges and Judgment: Setting the Scene. *M-RCBG Associate Working Paper Series*. 2023.

92. Delfino, Rebecca A. Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review*. 2019, 88(3) :887-938.

93. Palmiotto, Francesca. Detecting deep fake evidence with artificial intelligence: A critical look from a criminal law perspective. *Available at SSRN 4384122*. 2023.

94. Delfino, Rebecca A. Deepfakes on trial: a call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *Hastings LJ*. 2022, 74: 293-348.

95. Mehlman, Julia. Facebook and myspace in the courtroom: authentication of social networking websites. *Criminal Law Brief*. 2012, 8(1): 9-28.

96. MacNeil, Heather, and Heather MacNeil. Trusting Records as Legal Evidence: Common Law Rules of Evidence. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. 2000: 32-56.

97. Mnookin, Jennifer L. Scripting expertise: The history of handwriting identification evidence and the judicial construction of reliability. *Virginia Law Review*. 2001: 1723-1845.

98. Liang, Weixin, et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*. 2022, 4(8): 669-677.

99. Jones, Karl, and Bethan Jones. How robust is the United Kingdom justice system against the advance of deepfake audio and video?. *Electrotechnica and Electronica (E+E)*. 2022, 57(9-12): 103-109.

100. Durães, Dalila, Pedro Miguel Freitas, and Paulo Novais. The relevance of deepfakes in the administration of criminal justice. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. 2023, 351-369.

101. Gregory, Sam. Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*. 2022, 23(3): 708-729.

102. Caldera, Elizabeth. Reject the evidence of your eyes and ears: deepfakes and the law of virtual replicants. *Seton Hall L. Rev.* 2019, 50: 177-206.