Artificial Intelligence Should Not Become a "Black Hole" for Human Agency in Tort Law

Daria Kim*

This article analyses the implications of the tendency to anthropomorphise artificial intelligence (AI) systems for tort law. It shows that the view of AI technology as "autonomous", "unexplainable" and "unpredictable" can misguide the "fit-for-purpose" assessment of the existing liability regimes. The analysis points out that risks and harm associated with AI technology are not inflicted by AI systems as such but are mediated through AI applications, while the main challenge for the allocation of tortious liability lies in the highly distributed causation between human conduct and harm. Overall, it is argued that humans can and should retain agency over mitigating technological risks and internalising harmful effects, even when the sources of those risks and harms are highly distributed.

I. Introduction

AI systems based on machine learning (ML)[1] are often described as "autonomous", "unpredictable" and "unexplainable" (a "black box").[2] The tendency to anthropomorphise AI by endowing AI artefacts – computer software and hardware – with human-like characteristics can be observed in discourse practices of all sorts, from expert literature to mass media outlets to "fit-for-purpose" policy assessments.[3] AI-based systems are portrayed as having "intentions",[4] making "decisions"[5] and being capable of "reasoning".[6] The pervasiveness of anthropomorphic depictions of AI technology, which fail to distinguish between the metaphorical and the literal,[7] has led to the perception of an emerging human-like agency in AI systems.[8]

* MA, LLM, Dr iur, Senior Research Fellow, Max Planck Institute for Innovation and Competition, Munich, Germany.

[1] ML is a field of AI that deploys a class of algorithms that infer ("learn") output based on the input data. Discussions on the disruptive impact of AI on society and law typically imply ML, even though they often allude to "AI" in general. This article primarily addresses issues raised by ML applications.

[2] In computer science, a black box refers to a system in which the inputs and outputs are visible, but the internal workings and details are hidden. While no universally accepted definition of the term "black box" in relation to AI systems exists, this characteristic has become a hallmark of limited interpretability of complex computational models that evolve from the interaction of millions of units within a multi-layer, non-linear model structure. On "black box" as a misleading metaphor of AI, see nn 55–59 and the accompanying text.

[3] In the EU legislative practice, the "fit-for-purpose" assessment of a regulatory framework is carried out as part of the policy intervention design.

[4] *Impact Assessment Accompanying the Proposal Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Commission Staff Working Document, SWD(2021) 84 final, 21 April 2021) Pt 1, 93 (*Impact Assessment*) (assuming that, "even with good *intentions*, AI systems can cause *unintentional* harm" (emphasis added)).

[5] High-Level Expert Group on Artificial Intelligence (HLEG-AI), *A Definition of AI: Main Capabilities and Scientific Disciplines* (18 December 2018) 7 <https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf> (defining AI as "systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, *reasoning* on the knowledge derived from this data and *deciding* the best action(s) to take (according to pre-defined parameters) to achieve the given goal" (emphasis added)).

[6] HLEG-AI, n 5.

[7] See, eg, Isabella Hermann, "Artificial Intelligence in Fiction: Between Narratives and Metaphors" (2023) 38 *AI & Society* 319; Alberto Romele, "Images of Artificial Intelligence: A Blind Spot in AI Ethics" (2022) 35(1) *Philosophy & Technology* 4; Anke Beger, "The Brain Is a Computer and the Mind Is Its Program. Following a Metaphor's Path from Its Birth to Teaching Philosophy Decades Later" in Anke Beger and Thomas H Smith (eds), *How Metaphors Guide, Teach and Popularize Science* (John Benjamins Publishing, 2020) 263; Roberto Musa Giuliano, "Echoes of Myth and Magic in the Language of Artificial Intelligence" (2020) 35(4) *AI & Society* 1009; Marcin Miłkowski, "From Computer Metaphor to Computational Modeling: The Evolution of

The current surge in AI-based products and services has triggered an intense societal debate and necessitated "fit-for-purpose" assessments of the existing legal and regulatory frameworks across policy areas, including liability regimes.[9] Tort law, irrespective of jurisdictional specifics and terminological differences,[10] is in essence concerned with the allocation of liability for harm suffered and the award of compensation to the injured party. Here, the debate has evolved around the questions of how the existing liability rules can and should respond to "AI-induced" harm, whether there are regulatory gaps, and how they should be addressed. Discussions about the challenges that AI poses for tort law often draw on scenarios where harm results from incorrect diagnosis or treatment, failings of self-driving cars and robot surgeons, etc. Such scenarios pose the question of who is liable under existing law or should be held liable if the law were to be reformed.

The widespread perception of AI as "autonomous", "unpredictable" and "unexplainable" creates an impression that human control over risks associated with AI dissipates. Such characteristics may also suggest that the fundamental concepts of tort law – such as fault, causation and reasonable conduct – and the traditional principles of allocating tortious liability may no longer be applicable to harm "caused by AI". Legislative and scholarly proposals for addressing AI challenges for liability have included introducing "a special legal status for robots",[11] adopting the concepts of "corporate personhood" and "corporate liability",[12] subjecting AI-induced harm to strict liability[13] and establishing AI-specific liability insurance.[14]

Computationalism" (2018) 28(3) *Minds and Machines* 515; Harold D Carrier, "Artificial Intelligence and Metaphor Making: Some Philosophic Considerations" (1999) 12(1) *Knowledge, Technology & Policy* 45; Ronald M Biron, "The Computational Metaphor and Computer Criticism" (1993) 5(1) *Journal of Computing in Higher Education* 111.

[8] See, eg, Juliana Schroeder and Nicholas Epley, "Mistaking Minds and Machines: How Speech Affects Dehumanization and Anthropomorphism" (2016) 145(11) *Journal of Experimental Psychology: General* 1427; Francesco Ferrari, Maria Paola Paladino and Jolanda Jetten, "Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness" (2016) 8 *International Journal of Social Robotics* 287.

[9] For instance, the EU law- and policymakers initiated several studies examining the interface between AI and the EU liability frameworks. See Andrea Bertolini, *Artificial Intelligence and Civil Liability* (Study Requested by the JURI Committee of the European Parliament, July 2020); Tatjana Evas, *Civil Liability Regime for Artificial Intelligence. European Added Value Assessment* (Study of the European Parliamentary Research Service, 2020); Andrea Renda et al, "Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe" (Final Report, 2021). See also *Artificial Intelligence for Europe* (Communication from the Commission, COM(2018) 237 final, 25 April 2018) 16 (reporting that the Commission was "assessing whether the safety and national and EU liability frameworks are fit for purpose in light of these new challenges or whether any gaps should be addressed").

[10] While tort law terminology is jurisdiction-specific, this discussion refers to tort liability in the sense of non- or extra-contractual civil liability, the term used in the context of European continental law.

[11] See *European Parliament Resolution with Recommendations to the Commission on Civil Law Rules on Robotics*, 2015/2103(INL) (adopted 16 February 2017) [59(f)] (calling on the Commission to "explore, analyse and consider the implications of […] creating a specific legal status for robots […] and possibly applying electronic personality to cases where robots make autonomous decisions [as a possible legal solution]"). The proposal was criticised (see, eg, *Open Letter of Artificial Intelligence and Robotics Experts to the European Commission* <http://www.robotics-openletter.eu/>) and subsequently withdrawn. See *European Parliament Resolution with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence*, 2020/2014(INL) (adopted 20 October 2020) [7] (noting that "all physical or virtual activities, devices or processes that are driven by AI-systems may technically be the direct or indirect cause of harm or damage, yet are nearly always the result of someone building, deploying or interfering with the systems" and therefore "it is not necessary to give legal personality to AI-systems").

[12] See, eg, Alicia Lai, "Artificial Intelligence, LLC: Corporate Personhood as Tort Reform" [2021] *Michigan State Law Review* 597 (arguing for conferring corporate personhood to AI systems); Carla L Reyes, "Autonomous Corporate Personhood" (2021) 96(4) *Washington Law Review* 1453 (analysing the intersection between the concepts of "AI personhood" and "corporate personhood for autonomous businesses"). But see Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* (CUP, 2020) 3 (noting that "[t]here is not likely to be a single legal change, such as granting AI legal personality similar to a corporation, that will solve matters [brought about by AI technology] in every area of the law").

[13] See, eg, Christiane Wendehorst, "Strict Liability for AI and Other Emerging Technologies" (2020) 11(2) *Journal of European Tort Law* 150; Gerald Spindler, "User Liability and Strict Liability in the Internet of Things and for Robots" in Sebastian Lohsse, Reiner Schulze and Dirk Staudenmayer (eds), *Liability for Artificial Intelligence and the Internet of Things* (Nomos, 2019) 125; Yavar Bathaee, "Artificial Intelligence Opinion Liability" (2020) 35 *Berkeley Technology Law Journal* 113, 162ff. But see Bernhard A Koch, "Liability for Emerging Digital Technologies: An overview" (2020) 11(2) *Journal of European Tort Law* 115, 126 (noting that "will always be residual cases of 'traditional' fault liability involving human conduct, even in a fully automated society").

[14] Herbert Zech, "Liability for AI: Public Policy Considerations" (2021) 22 *ERA Forum* 147.

The justification for a legislative intervention – in the form of either adjusting the existing rules or introducing new ones – crucially depends on the factual understanding of a problem to be resolved. The depiction of AI as "autonomous" and "self-learning" has already been criticised as "an overvaluation of the actual capabilities" of AI systems.[15] Researchers also caution against the tendency to anthropomorphise AI as it distorts the public perception of the technology.[16] Notably, some explain this tendency as a means to create hype[17] around AI and even to increase trust towards AI-based products and services.[18]

On a more rigorous reading, however, the "autonomous decision-making" of ML systems comes down to *automated* "model fitting"[19] based on data corpora and calculation of statistical correlations and probabilities (also known as "predictions").[20] Stripped of anthropomorphic metaphors, ML output is reducible to several objective interacting elements: input data; algorithms incorporating mathematical (optimisation) functions that guide the "training" process; statistics; software engineering; and the impact of the environmental factors in which ML systems operate.[21] How AI technology is framed – as an "autonomous" agent or as a tool used deliberately – is of paramount importance for how AI-specific uncertainties and gaps are defined in tort law.[22]

The policy approach to AI in the European Union (EU) has evolved over time: the initial idea that "autonomous robots" should be granted legal status was abandoned.[23] More recently, it has been acknowledged that the existing liability frameworks at the EU level should be adjusted rather than radically changed. At the time of writing, several proposals of the European Commission for regulating AI are moving through the legislative pipeline:

- the Proposal for a Directive on liability for defective products (adapting the rules on strict liability for defective products to the digital age and AI);[24]
- the Proposal for a Directive on adapting non-contractual civil liability rules to AI (focusing on fault-based liability);[25] and
- the AI Act (representing safety regulation).[26]

[15] *Open Letter*, n 11, [2].

[16] Hermann, n 7; Donna L Hoffman and Thomas Novak, "Object-oriented Anthropomorphism as a Mechanism for Understanding AI" (2020) 48 *Advances in Consumer Research* 918; Alexander K Dewdney, "Misled by Metaphors: Two Tools That Don't Always Work" in Alexander K Dewdney (ed), *The Machine as Metaphor and Tool* (Springer 1993); Diane Proudfoot, "Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game" (2011) 175 *Artificial Intelligence* 950; Arleen Salles, Kathinka Evers and Michele Farisco, "Anthropomorphism in AI" (2020) 11 *AJOB Neuroscience* 88; David Watson, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence" (2019) 29 *Minds and Machines* 417; Adam Waytz, Joy Heafner and Nicholas Epley, "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle" (2014) 52 *Journal of Experimental Social Psychology* 113.

[17] See, eg, Eva Cetinic and James She, "Understanding and Creating Art with AI: Review and Outlook" (2022) 18(2) *ACM Transactions on Multimedia Computing, Communications and Applications* 13 (with further references).

[18] Waytz, Heafner and Epley, n 16.

[19] Josef Bajada, *Artificial Intelligence Demystified* (4 January 2019) Towards Data Science <https://towardsdatascience.com/https-medium-com-josef-bajada-demystifying-artificial-intelligence-6f5f7a8dd1b0> (arguing that a more suitable term for machine learning would be "automated model fitting [which would not sound] cool enough to attract the same level of investment and innovation interest").

[20] Kalev Leetaru, "Our Entire AI Revolution Is Built on a Correlation House of Cards", *Forbes*, 20 April 2019 <https://www.forbes.com/sites/kalevleetaru/2019/04/20/our-entire-ai-revolution-is-built-on-a-correlation-house-of-cards/>.

[21] As also discussed in Part II.B.

[22] Part III explores the implications of such difference in framing for the concept of causation and the incentives for risk prevention.

[23] For references, see n 11.

[24] The European Commission, *Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products* (COM(2022) 495 final, 28 September 2022). The revision of the EU Product Liability Directive 85/374/EEC addressed inter alia the uncertainty as to whether "intangibles" such as data, digital content and software constitute a "product" within the meaning of the Directive.

[25] The European Commission, *Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)* (COM(2022) 496 final, 28 September 2022) (*Proposal for AI Liability Directive*).

[26] The European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (COM(2021) 206 final, 21 April 2021) (*Proposal for EU AI Act*).

Such legislative measures are regarded as complementary instruments to incentivise research in, and development of, reliable and trustworthy AI applications.[27]

Both the draft Directive for (strict) liability for defective products and the draft Directive for non-contractual (fault-based) liability mainly aim to introduce targeted measures to ease the injured party's burden of proof. Such measures – in addition to direct safety regulation under the EU AI Act – are considered necessary in view of the assumed specific characteristics of AI systems, namely, their "autonomy", "unpredictability" and "black box" nature.[28] Substantive law issues – such as what constitutes fault or by which legal standard causality should be determined – remain at the discretion of the national law of the EU Member States. The question of how the characterisation or perception of AI systems as autonomous and unpredictable black boxes may impact the application of traditional tort law concepts by national courts remains open.

This article does not aim to determine whether fault-based or strict liability – "two intellectually tenable and more or less equally respectable answers to the question of what should be the general standard of liability in tort"[29] – should be the preferred choice for addressing harm ensuing from AI applications. Instead, the purpose is two-fold: first, to highlight that the characterisation of AI systems as autonomous black boxes is at a minimum arguable and by no means either denotes agency in a human-like sense[30] or entails the fundamental unfeasibility of human control[31] and, second, to show that, upon closer technical examination, "autonomous" AI systems do not render the concepts of causation and incentives for risk prevention inapplicable or no longer relevant in tort law.

The discussion unfolds as follows. Part II starts with an overview of findings of earlier "fit-for-purpose" analyses of tort law that characterise AI as "autonomous", "unpredictable" and "unexplainable" and, hence, assume a limited human role in preventing risks "caused" by AI (Part II.A). It then introduces a contrasting perspective on AI systems that emphasises that the functional relationship between the input and the output is known, objective and deterministic. It argues that AI systems can play a role in mediating or influencing the occurrence of risks, while the main challenge in allocating liability arises from the complex causation between human conduct and resulting harm, which is often distributed across multiple parties (Part II.B). Part III discusses the implications of conceiving "AI harm" as "AI-*mediated* harm"[32] – as opposed to the conception of harm being "caused by" AI – for the application of the existing concepts and principles of tort law. It focuses on the notion of "causation" and incentives for risk prevention in view of their central importance for the objectives of tort law. The conclusion reiterates that the representation and perception of AI as "autonomous", "unpredictable" and "unexplainable" risks downplaying the ability of humans to take and exercise control. As argued, such view of AI technology can lead to "false negatives" at the policy level (ie the underestimated possibility to prevent or mitigate AI risks, including through tort liability rules) as well as "false negatives" at the individual case level (ie failure to allocate liability based on the causality between the human conduct and AI output).

By "AI risks",[33] the discussion refers to the risks associated with the deployment of ML-based products or services. While this article focuses on the safety risks of AI technology,[34] it is acknowledged that AI

[27] See, eg, *Impact Assessment*, n 4, 7, 33–34.

[28] See, eg, *Proposal for AI Liability Directive*, n 25, 1, 13, 16, 22 (in particular, draft recitals 3, 27, 28).

[29] Stephen R Perry, "The Impossibility of General Strict Liability" (1988) 1(2) *Canadian Journal of Law & Jurisprudence* 147, 147.

[30] Magdalena Żemojtel-Piotrowska, Jarosław Piotrowski and Amanda Clinton, "Agency" in Virgil Zeigler-Hill and Todd K Shackelford (eds), *Encyclopedia of Personality and Individual Differences* (Springer, 2020) 69.

[31] As discussed in Part II.B.

[32] The term "AI-mediated" is used in this paper to highlight the intermediary role of AI system, in contrast to the notion that these systems are autonomous and have agency of their own.

[33] Such risks are understood as situations where "a danger caused directly or indirectly by the development and/or deployment of AI". Renda et al, n 9, 56.

[34] It is worth noting that such characteristics are mainly associated with ML techniques, even though reference is usually made generally to "AI".

applications pose risks to a broader range of legally protected interests.[35] The analysis draws primarily on the EU legal framework for extra-contractual liability. At the same time, it is assumed that, terminological differences aside, basic concerns and substantive legal issues raised by AI technology are relevant across jurisdictions. Most pressingly, there is a need to define a pertinent approach to the allocation of liability for harm caused by using AI applications.

II. From "AI-caused" to "AI-mediated" Understanding of Risks

The maxim that "understanding a problem is half the solution" underlies the methodology of designing policy intervention. In the EU legislative practice, problem analysis refers to the identification of a problem that in the absence of a policy intervention cannot be solved, as well as its driving forces and contributing factors.[36] The answer to the question of how tort law can and should respond to challenges posed by AI technology depends on how such challenges are understood and defined. This section draws on the policy reports carried out in support of two complementary legislative developments at the EU level – the AI Act[37] and the reform of the EU liability regime[38] – and exposes how challenges raised by AI have been defined in the course of law-making.

A. AI "Autonomy" as the Assumed Challenge for Tort Liability

The "fit-for-purpose" assessments attribute the major regulatory challenges posed by AI technology to its specific characteristics such as "autonomy", "unpredictability" and "black box" nature of ML models. Such traits are identified as "problem drivers" – factors raising uncertainty regarding the "fitness" of the existing framework, including liability rules.[39] In the subsequent paragraphs, a few excerpts defining the "problem drivers" are reproduced in the original wording, as the language is key to understanding how the legislative bodies perceive and define the factual and legal challenges to be addressed.

1. "Autonomous Behaviour"

The authors of the Impact Assessment Report have posited that –

> [t]he level of autonomy of AI systems is a continuum, ranging from fully supervised and controlled systems to more independent ones that combine environmental feedback with an analysis of their current situation and can perform tasks *without direct human intervention*, even when operating in a complex environment. While the high-level objectives of AI systems are always defined by humans, [some AI] outputs and the mechanisms to reach these objectives are not concretely specified, enabling *automated decision-making*, within pre-set boundaries, without the involvement of a human operator. […].[40]

Accordingly, AI autonomy is assumed to make it "*difficult to prove the link between a damaging output of the AI system and the action or omission of a potentially liable person*".[41]

2. "Opacity/Lack of Transparency and Explainability"

AI systems have also been viewed as "lack[ing] transparency because the rules followed, which lead from input to output, are *not fully pre-determined* by a human".[42] Hence:

---

[35] Under the European approach, the risk assessment of AI technology is anchored in the protection of fundamental rights.

[36] European Commission, *Better Regulation Guidelines* (SWD(2015) 111 final, 19 May 2015) 90–91.

[37] For the full reference, see n 26.

[38] For the full references, see nn 24–25.

[39] *Impact Assessment*, n 4, 28ff; *Impact Assessment Report Accompanying the Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence* (Commission Staff Working Document, SWD(2022) 319 final, 28 September 2022) 120ff (*Impact Assessment*).

[40] *Impact Assessment*, n 39, 120 (emphasis added).

[41] *Impact Assessment*, n 39, 120 (emphasised in the original).

[42] *Impact Assessment*, n 39, 120 (emphasis added).

> [I]t may not be possible to explain *how the output is exactly derived* from its input in a given context […] *how exactly* [an AI system] functions as a whole, how the algorithm was realised in code and how the programme actually runs in a particular case, including the hardware and input data.[43]

Consequently, it is concluded that "the opacity of an AI system can make it very challenging to establish which input data lead to a specific output creating a damage, and how"[44] and that:

> [i]n the absence of a targeted alleviation of the burden of proof, and targeted rules on the disclosure of information on AI-systems in the context of civil proceedings, *this can make it very hard or even impossible for victims to prove the link between the harmful AI output and a human action or omission*.[45]

3. "Continuous Adaptation and Lack of Predictability"

In the policy assessment, the lack of predictability of AI systems is related to the ability of AI systems to "learn" while in operation, whereby "the rules being followed by the system adapt based on the input it receives while in use".[46] Besides, the limited predictability of AI systems has been attributed to the fact that "data driven AI systems in general have a *probabilistic behaviour* as they are based on data that usually does not represent all possible scenarios".[47] The report also notes that the output of AI systems can display "a high input sensitivity [whereby] even a small change in the inputs can lead to an unpredictable behaviour of the system, and thus a lack of predictability".[48] Hence, the authors of the assessment expect that ML applications "based on continuous adaptation may […] *make it impossible to prove the link between the damaging AI output and a specific cause, in particular, the fault of a person*".[49] ML systems that continue "learning" and "may even change their own behaviour in unforeseen ways" after being released into the environment have been regarded as sources of "new risks that are not adequately addressed by the existing legislation".[50]

Overall, the language used in the above-cited report heavily draws on anthropomorphic metaphors ("behaviour", "decision-making", "learning", etc). In other reports, AI systems are attributed the capacity for "reasoning" and being "rational".[51] Such depictions may create the impression that the causal link between human input and AI output, which constitutes the basis for the ability to control AI systems, may eventually dissipate and that an AI system may gradually acquire "agency of its own". Notably, AI as such is considered – even "suspected"[52] – to cause harm.[53] Before policy measures can be considered to address the challenges of "autonomous, unpredictable and opaque" AI systems, the nature of such challenges needs to be examined more closely.

---

[43] *Impact Assessment*, n 39, 120 (emphasis added).

[44] *Impact Assessment*, n 39, 120.

[45] *Impact Assessment*, n 39, 120–121 (emphasised in the original).

[46] *Impact Assessment*, n 39, 121.

[47] *Impact Assessment*, n 39, 121 (emphasis added).

[48] *Impact Assessment*, n 39, 121.

[49] *Impact Assessment*, n 39, 121 (emphasised in the original).

[50] *Impact Assessment*, n 4, 28.

[51] HLEG-AI, n 5, 3 (describing a "learning rational AI system" as "a rational system that, after taking an action, evaluates the new state of the environment (through perception) to determine how successful its action was, and then adapts its *reasoning* rules and decision making methods" (emphasis added)).

[52] *Impact Assessment*, n 39, 102 (referring to "AI system [being] suspected to have at least contributed to causing harm to the victim").

[53] See, eg, the Commission to the European Parliament, the Council and the European Economic and Social Committee, *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and robotics* (COM(2020) 64 final, 19 February 2020) 7, 8, 15 (*Report on the Safety and Liability*) (referring to "the safety impact caused by the autonomous behaviour throughout the product lifetime", "AI humanoid robots [causing] the immaterial harm […] to users" and the difficulty in "getting compensation […] for damage caused by autonomous AI-applications").

B. No Agency in AI Systems

The tendency to anthropomorphise AI has been met with scepticism and criticism.[54] The "black box" characterisation of AI has also been criticised for being misleading.[55] As one ML researcher explains, if an autonomous car goes off the road, the reason for that is that that car:

> applied a *transparent* and *deterministic* computation using the values of its parameters, given its current input, and this determined its actions [and] if we ask why it had those particular parameters, the answer will be that they are the result of the model that was chosen, the data it was trained on, and the details of the learning algorithm that was used.[56]

In other words, the process of building models is understandable and transparent as all computations are based on mathematical rules and statistical principles and cannot deviate from the algorithmic instructions.[57] While the term "black box" is typically used to refer to computational complexity,[58] such complexity should not be equated with the fundamental non-explainability or incomprehensibility of computations.[59]

Another tendency that is strongly present in the discourse on AI is to use the terms "autonomy" and "automation" interchangeably.[60] This is inaccurate as instances, where a task is carried out without direct human intervention *during* its implementation, should be properly called "automation".[61] When stating that AI systems "can increasingly perform tasks with less, or entirely without, direct human intervention",[62] it should be emphasised that a human might not be directly involved during the task *execution*. However, for that, any computational process needs to be conceived, planned and configured by a human in the first place.[63] In other words, it would be a leap to consider the automated implementation of computation as "autonomous behaviour"[64] of AI systems. Besides, such notion is misleading as "autonomous behaviour" – as a manifestation of agency and personality[65] – involves a broad range of complex cognitive operations and psychological states. In this sense, it is incomparable with automated algorithmic operations.

---

[54] For the references, see n 16.

[55] Miranda Marcus, *Magic Boxes & Machine Learning – Why We Need to Stop Using the Black Box Metaphor* (24 December 2020) Towards Data Science <https://towardsdatascience.com/magic-boxes-machine-learning-why-we-need-to-stop-using-the-black-box-metaphor-c9f345d1bc12>.

[56] Dallas Card, *The "Black Box" Metaphor in Machine Learning* (5 July 2017) Towards Data Science <https://dallascard.medium.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0> (emphasis added).

[57] Meredith Broussard et al, "Artificial Intelligence and Journalism" (2019) 96(3) *Journalism & Mass Communication Quarterly* 673 (defining AI "we have today" as "merely complex and beautiful mathematics" and ML as being based on "computational statistics, albeit on steroids").

[58] *Impact Assessment*, n 39, 121 (explaining that "AI models frequently have more than a billion parameters and process very large amounts of data [and] are usually combined together in complex systems in real world scenarios […] *Due to this multiplicity of elements constituting* such AI systems, they may not be understandable in practice for humans. While some AI systems can be comprehensible from an ex post perspective despite their complexity (eg complex rule-based systems), complexity can contribute to a lack of explainability of outputs in other cases" (emphasis added)).

[59] Daria Kim et al, "Clarifying Assumptions about Artificial Intelligence before Revolutionising Patent Law" (2022) 71 *GRUR International* 295, 311ff.

[60] Such tendency can also be observed in the European Commission's documents: see, eg, *Impact Assessment*, n 39, 120 (referring to "autonomous behaviour" and "the level of autonomy of AI systems" and at the same time to "the automation of a certain task" and "automated decision-making").

[61] Shimon Y Nof, "Automation: What It Means to Us around the World" in Shimon Y Nof (ed), *Springer Handbook of Automation* (Springer, 2009) 14.

[62] *Impact Assessment*, n 39, 120 (emphasised in the original).

[63] See, eg, Greg Michaelson, "Teaching Programming with Computational and Informational Thinking" (2015) 5(1) *Journal of Pedagogic Development* 51, 53–54; Kim et al, n 59, 307.

[64] *Impact Assessment*, n 39, 2, 8, 34, 88.

[65] Mazen El-Baba and Joseph Jamnik, "Personality and Mortality" in Virgil Zeigler-Hill and Todd K Shackelford (eds), *Encyclopedia of Personality and Individual Differences* (Springer, 2020) 2111.

As a corollary to "autonomy", AI systems have also been endowed with the ability to "learn" and make "decisions".[66] In technical terms, "learning" means the process of deriving a model (a numeric function) that captures the functional relationship between inputs and outputs. This relationship is objective in the sense that it pre-exists in data. In the context of ML, a "decision" stands for a "prediction" generated by a trained ML model which, in technical terms, means the calculated numerical probability.[67] The accuracy of predictions is key for the robust functioning of ML-based applications.[68] A number of factors bear on the accuracy of ML models, including the selection of the training datasets and the training process.[69] Consequently, such factors constitute potential sources of technological risks, but such dependency also suggests that ML-induced risks are controllable, at least as regards model accuracy.

One can come across accounts of ML systems as being "unpredictable" and having "implications far beyond the intentions of their developers".[70] This is true to some extent. However, it is important to realise the causes and scope of such uncertainty. Most ML applications are developed for specific tasks, even when based on foundation models. Their functionality and predictive capacity are constrained by their design, input data, and the training process.[71] The same model that was trained to distinguish between malignant and benign cells cannot be readily applicable to generate predictions regarding other biomarkers. Despite being hailed as general-purpose models, large language models do one thing only – "generate statistically likely sequences of words".[72] While the ability of ML models to adapt continuously to new input data has been viewed as the source of unpredictability, it is important to note that such unpredictability primarily arises from environmental factors that affect system deployment, rather than the system's capacity to exercise "free", "subjective" choices.[73] Current ML techniques are deterministic in that they deliver the same output for the same input if all relevant conditions are reproduced.[74] Neither can randomisation be equated with the non-determinism of ML systems.[75]

Attributing the characteristics of "autonomy", "non-explainability" and "unpredictability" to AI can create a false framing, suggesting the presence of "agency" in AI systems. Yet, even though AI systems are often referred to as "agents" they are not endowed with "agency" in the sense of "intentionality and volition".[76] Current ML systems are not "autonomous" in the sense that they avail themselves of

[66] On "self-learning" as a misnomer, see *Open Letter*, n 11.

[67] For a more detailed explanation, see Kim et al, n 59, 307ff.

[68] Francesco Amigoni and Viola Schiaffonati, "The Importance of Prediction in Designing Artificial Intelligence Systems" in Marcello Pelillo and Teresa Scantamburlo (eds), *Machines We Trust: Perspectives on Dependable AI* (MIT Press, 2021) 105.

[69] Accuracy, as a key quality of the performance of ML models, is measured by the proportion of correct predictions relative to the total number of predictions generated by an ML model. The degradation of accuracy of an ML model (also known as "concept drift") can occur due to various reasons, including changes in the distribution of the input data and changes in the environment in which an ML model is used. See also Renda et al, n 9, 62ff.

[70] World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), *COMEST Report on Robotics Ethics* (SHS/YES/COMEST-10/17/2 Rev 42, 14 September 2017) <https://unescoblob.blob.core.windows.net/pdf/UploadCKEditor/REPORT%20OF%20COMEST%20ON%20ROBOTICS%20ETHICS%2014.09.17.pdf>.

[71] This issue pertains to the generalisability of ML model, which is one of the challenges and research areas in ML. Some approaches such as model pre-training aim to solve the problem of limited generalisability. See, eg, Rishi Bommasani et al, "On the Opportunity and Risks of Foundation Models", *Cornell University arXiv* (E-print, 12 July 2022) <https://arxiv.org/pdf/2108.07258.pdf>.

[72] Murray Shanahan, "Talking about Large Language Models", *Cornell University arXiv* (E-print, 7 December 2022) <https://arxiv.org/abs/2212.03551>. For a detailed technical explanation, see Cal Newport, "What Kind of Mind Does ChatGPT Have?", *The New Yorker*, 13 April 2023 <https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have>.

[73] See *Impact Assessment*, n 39, 120–121 (explaining that "built-in processes of the AI system […] depend upon a variety of internal and external factors independent from such conduct, for instance input data collected by sensors, which makes it again more difficult to link a harmful output to a human action or omission").

[74] For a detailed technical explanation, see Kim et al, n 59, 306ff.

[75] ML is based on mathematical optimisation whereby an algorithm guides the calculation of optima within the search space. While randomisation allows for exploration of different computational routes to finding the optima, current randomisation techniques are deterministic in the sense that for exactly the same input they yield the same output.

[76] Żemojtel-Piotrowska, Piotrowski and Clinton, n 30.

the decision-making capacity or the freedom of making choices. The training of a model is carried out within the pre-specified rules, boundaries and assumptions. An ML system cannot violate constraints and instructions: for instance, if an artificial neural network is trained with a particular cost function, such function – or other mathematical rules and statistical principles – cannot be dismissed by a computer.

In light of the foregoing, the notions of AI "autonomy", "non-explainability" and "unpredictability" should not be taken literally, as preordained properties of AI, or as denoting human-like "agency". Instead, the perception of the lack of human control arises from the multiplicity of factors affecting the performance of AI systems and the complexity of their interplay. Thus, the challenge of allocating legal consequences for harmful ML output subsists not in the dissipation of human control, but rather in its dispersion among numerous actors. This insight suggests an important shift in the framing of AI from AI systems presenting the source of risks and causing harm by themselves to AI as playing a *mediating* role between human conduct and the harm occurrence.[77] While it is true that research in the field of AI is advancing, there is currently no consensus on when artificial general intelligence (if it were a proxy for "human-like" intelligence) might be developed.[78] Importantly, computer scientists note that achieving "human-like" capabilities of AI systems is not simply a matter of acquiring more data, and it may require an entirely different approach than the currently dominant data-intensive methods.[79]

In drawing implications of AI characteristics, the European Commission's impact assessment for the AI Act mainly refers to the challenges of the burden of proof in liability cases.[80] At the same time, the question of how substantive concepts and rules should apply in the cases of AI-mediated harm remains open to interpretation. In light of the clarifications provided in this part, the next section focuses on certain substantive aspects of tort law.

III. Implications of the Concept of AI-mediated Harm for Tort Liability

The main teaching of the deterministic view of AI technology is that ML systems are not endowed with the agency to cause harm "by themselves", driven by their "own mind and will", and the human impact on ML outcome does not vanish but becomes highly dispersed among multiple parties, decisions and activities. Accordingly, the core objectives of tort law – defining the standard of care, adequate compensation for harm[81] and incentives for risk prevention[82] – hold in the advent and diffusion of ML systems. However, the attainment of these goals may have nuances in cases of ML-mediated harm. The

[77] Robert C Williamson, *AI as Mediator* (15 December 2021) Machine Learning for Science <https://www.machinelearningforscience.de/en/ai-as-mediator/> (proposing that, just as "good science relies upon being able to trace back through long and complex 'chains of reference', we need to improve our systems for tracing the complexities of decision making when mediated by technological systems […] systems confusingly known as 'artificial intelligence'"). See also Klaus Mainzer, *Artificial Intelligence – When Do Machines Take Over?* (Springer, 2020) 276 (noting that, while "it is the human being who should determine how [AI] is used […], specialization and the growing complexity of technical, social, and ecological contexts lead to a diffusion of responsibility").

[78] Martin Ford, *Architects of Intelligence: The Truth about AI from the People Building It* (Packt Publishing, 2018).

[79] Pat Langley, *The Computational Gauntlet of Human-Like Learning*, The Institute for the Study of Learning and Expertise <http://www.isle.org/~langley/papers/gauntlet.aaai22.pdf>; Michael I Jordan, *Artificial Intelligence – The Revolution Hasn't Happened Yet* (23 June 2019) Harvard Data Science Review <https://hdsr.mitpress.mit.edu/pub/wot7mkc1>.

[80] See, eg, *Impact Assessment*, n 39, 102. The document states: "Identifying the cause of damage and convincing the court of its impact on the turn of events is challenging if it is an AI system that is suspected to have at least contributed to causing harm to the victim. This is due to the very nature of AI systems and their peculiar features such as complexity, opacity, limited predictability, and openness. Identifying harmful conduct will be the more difficult the more independent the behaviour of the AI system is designed, or – figuratively speaking – the more black the box is."

[81] The concept of compensable harm remains relevant no matter how harm was inflicted. See Koch, n 13, 122 (observing that "[e]merging digital technologies may cause the same types of damage as any other source of harm" but "with increased digitisation comes greater dependence on data and in turn higher vulnerability for data corruption or loss"). Given the variety of AI technology applications, the nature of the legally protected interests put at risk by the deployment of ML systems may vary. In addition to harms also associated with traditional technologies (bodily or property), ML-based applications may inflict other types of harm such as mental and emotional. See, eg, Renda et al, n 9, 58 (defining mental consequences of using AI as "individuals' mental health may be impacted by new beliefs that were propagated through fake news or chatbots" and emotional consequences as "the possibility of depression" (with further references)).

[82] *Report on the Safety and Liability*, n 53, 12.

remainder of this section focuses on the implications of distributed AI risks for causation, due care and risk minimisation incentives, the concepts that are central to both objectives of tort law.

A. Causation in Cases of AI-mediated Harm

Causation is a cornerstone concept of tort liability and a relevant factor for both fault-based and strict liability.[83] If one took the view that AI systems may act and inflict harm "autonomously", the notion of causation between human conduct and harm would lose all relevance. However, as previously mentioned, ML outcomes depend on the intricate and multi-faceted decision-making and actions of various individuals throughout the system's design, as well as external factors that impact its operation.[84] Just as a brick falling on a pedestrian passing by a building can be the result of gravitational force, engineering flaws in a building, the pedestrian's decision to take a certain route at a certain time, or even someone's intention to drop the brick, AI-mediated harm would typically result from an interplay of multiple man-made and environmental factors. An ML model may have design errors that can affect its accuracy. Safety risks would materialise if the relevant causal conditions align.[85] The challenge would lie in working out which particular action or omission within the complex AI value chain was decisive for the risk materialisation, how to weigh the relative causative power of multiple factors at play, and whether certain entities were at fault when acting or failing to act.

Multiple or uncertain causation is neither a new phenomenon nor a unique challenge posed by AI techniques for tort law. Jurisdictions have adopted different approaches to dealing with situations in which multiple causal factors might be at play or it is uncertain which conduct is relevant for the allocation of tortious liability.[86] In particular, two substantive approaches have been applied to deal with uncertainty over causation: establishing a threshold for the probability that the defendant has caused the loss (the "all-or-nothing" approach) and proportional liability. Under the first approach, the defendant will be held liable if the 50% threshold probability is exceeded, that is the defendant more likely than not caused the loss.[87] Under the second approach, the defendants will be held liable regardless of uncertainty over causation but the amount of damages will be imposed in proportion to the likelihood of causation.[88]

Accordingly, as a starting point, one could examine to what extent the concepts of "probability" and "likelihood" could be instrumental to deal with uncertain causation in cases of AI-mediated harm. Furthermore, measures of a more procedural nature such as (rebuttable) presumptions and the shift or alleviation of the burden of proof have previously been applied to treat uncertain causation.[89] For instance, the solution proposed by the draft EU AI Liability Directive proposes to relax the burden of proof and introduces a rebuttable presumption of "the causal link between the fault of the defendant and

[83] The concept of causation and the standard of proof vary among jurisdictions. For an overview in the European Union, see Marta Infantino and Eleni Zervogianni, "Summary and Survey of the Results" in Marta Infantino and Eleni Zervogianni (eds), *Causation in European Tort Law* (CUP, 2017) 592 (finding that "the majority of European jurisdictions address causation problems in negligence and strict liability through the same tests"). At the EU level, strict liability applies under the EU Directive for liability for defective products (n 24), according to which the plaintiff needs to prove a causal connection between the damage and the product defect.

[84] Such environmental factors can affect the operation of an ML-based system if its functionality depends on real-time input data that adjusts the model.

[85] For an analysis of hypothetical cases where a model prediction may lead to harm, see *Impact Assessment*, n 39, 233ff.

[86] Infantino and Zervogianni, n 83.

[87] Steven Shavell, "Causation and Tort Liability" in P Newman (ed), *The New Palgrave Dictionary of Economics and the Law* (Palgrave Macmillan, 2002) 211, 212.

[88] Shavell, n 87, 212.

[89] Such as in the case of enterprise liability, whereby the victim does not need to prove what exactly within an enterprise constituted the cause of harm, as long as she can prove that the enterprise activity as such was flawed.

the output produced by the AI system or the failure of the AI system to produce an output",[90] as a means to address uncertain causation due to the specific characteristics of AI.[91]

Given that causation is usually distributed across a *set* of factors and events,[92] the question arises which of the events within the multi-party, multi-input, multi-stage value chain of AI-based products and services should be deemed as not only "causes in fact" but "causes in law". This issue pertains to a longstanding debate about what it means "to cause" harm for the purposes of allocating legal responsibility and how to distinguish between causation in a "basic" (or natural or factual) and a "legal" sense.[93] The debate regarding the legal standard of causation has evolved around two "leading philosophical theories of causation":[94] strong necessity and weak necessity/strong sufficiency. While the "but for" (strong necessity) criterion has been widely applied across jurisdictions, the "weak necessity/strong sufficiency" criterion has been a matter rather of theoretical discussion than of judicial application.[95] An example of the collocation of both approaches can be found under the Principles of European Tort Law (PETL), which define the "conditio sine qua non" of causation as the "but-for" standard[96] but also apply the concept of weak necessity/strong sufficiency to cases of multiple activities to deal with uncertain partial causation[97] and alternative causes.[98] The latter case is also treated under the principle of proportional liability.[99]

Remoteness is another fundamental concept in tort law that may become highly relevant in scenarios of AI-mediated harm. When applying this concept, the legal assessment should determine whether harm mediated by an AI system was too remote or not reasonably foreseeable by a certain party within the AI value chain. What needs to be examined in more detail is what constitutes "reasonable foreseeability" in the context of risks posed by AI applications.[100] Overall, complex distributed technologies such as AI-based applications might present a pertinent opportunity to revisit the debate regarding the relative value of the competing legal standards for causation for resolving complex cases of distributed technological risks.

In addition, it is worth noting that cases that involve multi-factor causation are often difficult from an evidential perspective. Technology-related cases typically necessitate expert assessment.[101] Although these challenges for legal assessment are not new, one would expect them to be exacerbated in situations where harm has occurred through advanced AI applications. While commentary on AI and tort law often laments the unfeasibility of proving causation where harm was "caused" by "autonomous, unpredictable

[90] *Proposal for AI Liability Directive*, n 25, Art 4, references the provisions of the *Proposal for EU AI Act* (n 26) that lay down the safety requirements for high-risk AI systems, including with regard to data governance. The proof of non-compliance with such requirements would trigger the presumption of a "causal link between the fault of the defendant and the output produced by the AI system or the failure of the AI system to produce an output".

[91] *Impact Assessment*, n 39, 120ff.

[92] On the philosophical underpinnings of the debate, Richard W Wright and Ingeborg Puppe, "Causation: Linguistic, Philosophical, Legal and Economic" (2016) 91 *Chicago-Kent Law Review* 461, 464ff.

[93] Wright and Puppe, n 92.

[94] Christian Barry, "Understanding and Evaluating the Contribution Principle" in Andreas Follesdal and Thomas Pogge (eds), *Real World Justice: Grounds, Principles, Human Rights, and Social Institutions* (Springer, 2006) 103, 132.

[95] RT Hon Lord Hoffmann, "Causation" in Richard Goldberg (ed), *Perspectives on Causation* (Hart Publishing, 2011) 3.

[96] *Principles of European Tort Law*, Art 3:101 <http://egtl.org/PETLEnglish.html>.

[97] *Principles of European Tort Law*, n 96, Art 3:105.

[98] *Principles of European Tort Law*, n 96, Art 3:103(1).

[99] In particular, in the case of alternative causes, each activity would be regarded as a cause "to the extent corresponding to the likelihood that it may have caused the victim's damage" (*Principles of European Tort Law*, n 96, Art 3:103(1)).

[100] See, eg, Weston Kowert, "The Foreseeability of Human-Artificial Intelligence Interactions" (2017) 96(1) *Texas Law Review* 181, 203 (emphasising the case-by-case nature of foreseeability analysis in the context of human-AI interactions).

[101] For a discussion in the context of AI, see Sylwia Wojtczak, "Causation in Civil Law and the Problems of Transparency in AI Profile" (2021) 29(4) *European Review of Private Law* 561.

and unexplainable" AI,[102] a closer look at ML technology, its de-anthropomorphised account[103] and the analysis of hypothetical cases of AI-mediated harm[104] show that several focal points are likely to have the strongest impact on the accuracy of an ML model. It is true that the amount and complexity of activities involved in the development and deployment of AI systems might be overwhelming. However, when it comes to safety risks due to the gradual degradation of an ML model's accuracy, it is possible to identify a set of decisive factors.[105] In such cases, it might be feasible to make a case-specific assessment of the likelihood of each factor that could have led to the harmful outcome.[106] For that, relevant information related to model training and device operation must be properly documented, and access to this information ensured.[107]

The point submitted here is that the value chain of AI-based products is highly complex, with multiple parties, inputs and stages, and that this complexity makes instances of AI-mediated harm good candidates to be treated as cases of multiple activities and uncertain, alternative or partial causation. The question of how to deal with such types of causation has been explored in legal theory and the value of that body of work needs to be examined first before tort law is "reinvented" to accommodate AI-mediated harm.

B. The Legal Standard of Due Care

The definition of the due care standard is relevant for both the fair allocation of fault liability and the effective prevention of harm. In the context of AI, it has been argued that both technology users and potential victims suffering harm are unlikely to have knowledge allowing them to take due care and avoid damage. Hence, the incentive effect of fault-based liability may "fail",[108] while strict liability on those who are closest to the product development might be better suited to provide the incentive "to acquire the necessary risk knowledge".[109] While the knowledge of safety risks posed by emerging technologies is inevitably imperfect, especially at the early stages of technology deployment, the accumulation of such knowledge is a matter of conducting research and testing and establishing feedback loops. AI artefacts are products of engineering and the source of uncertainty regarding their functioning lies not in their innate "agency" but in the "unknowns" associated with the input data and the environment in which they operate.[110] Control over technological risks under uncertainty is not a unique challenge for humanity, and the body of knowledge accumulated in this field can inform the design of policy strategies and responses for risk management specific to AI use-cases and types of risks.[111]

With regard to AI-based applications, due care standards should be informed by a robust technical understanding of the conditions under which technological risks of AI systems may materialise.[112] Where the due care requirement is stipulated as a general clause anchored in the conduct of "a reasonable person", it would need to be further interpreted for all activities throughout the AI development cycle, taking into account the relevant factors such as foreseeability, the availability of preventive measures,

[102] As shown in Part II.A.

[103] As discussed in Part II.B.

[104] *Impact Assessment*, n 39, 233ff.

[105] For an explanation, see nn 68–69 and the accompanying text.

[106] For an analysis of prototypical examples, see *Impact Assessment*, n 39, 233ff.

[107] The following provisions have been proposed in the EU to facilitate access to information and alleviate the burden of proof of the injured party: *Proposal for AI Liability Directive*, n 25, Art 3 and *Proposal for* (*revised) Product Liability Directive*, n 24, Art 8.

[108] Zech, n 14, 151.

[109] Zech, n 14, 151.

[110] As discussed in Part II.B.

[111] See, eg, Reva Schwartz, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* (National Institute of Standards and Technology Special Publication 1270, 2022) <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf>; Asaf Tzachor et al, "Responsible Artificial Intelligence in Agriculture Requires Systemic Understanding of Risks and Externalities" (2022) 4 *Nature Machine Intelligence* 104.

[112] As discussed in Part III.A.

the dangerousness of the activity, etc.[113] While many parties are usually involved in the development and design of AI systems, two categories of actors should be distinguished for the purposes of defining the respective due care standards: those who have or are supposed to gain expert knowledge (professional developers and users (operators) of AI systems) and lay persons (ordinary end users). As a general principle, specialists are held to higher standards of care due to their "above average capacities [and] special knowledge and qualifications".[114] Typical examples include engineers and healthcare professionals, who are expected to meet a higher standard of care that reflects their expert level of knowledge and specialised expertise. Accordingly, the due care standard for professional developers or operators of AI should be informed by the state-of-the-art knowledge, best practices and industry codes in the field of AI applications, while the regulator should ensure that there are sufficient incentives in place for such knowledge and information to be created and adequately communicated.[115]

C. Causes within the Victim's Sphere

Given the possibility of contributory negligence, it is important to determine which "activity, occurrence or other circumstance" should be regarded as being within the victim's "own sphere".[116] As discussed earlier, the functionality of ML-based applications depends on predictions generated by an ML model.[117] In this regard, one should differentiate between ML-based applications where the user makes a decision to act based on a prediction of an ML model and those in which predictions are automatically executed while the system is functioning (eg an image recognition system informing automated driving). In the former case, the user decides whether to act on the information generated by an ML model and such decision should be deemed as being within the victim's "own sphere" (for instance, the user of an ML-based medical application is in charge of whether to implement its health recommendations[118]). In the latter case, additional safeguards should be embedded within a product or a service that would allow the detection of abnormal functioning, including due to the decline of model accuracy.

In both cases, the user's understanding and perception of AI technology are key for determining what is "reasonable", which raises the question of whether a "reasonable" user is one with a trusting or rather a sceptical attitude towards AI. Users' attitudes are naturally susceptible to claims made about technology in mass media and commercial speech. Hyped claims about human-like robots or other AI artefacts may trigger unrealistic expectations. Studies[119] show that the anthropomorphisation of AI affects its perception by the lay audience. Notably, it appears to increase trust in autonomous vehicles.[120] Such evidence is worrying as it suggests that lay persons may be prone to miscalculate AI risks. The impact of framing and communication regarding AI on societal perception of AI technology and risk awareness is

[113] For instance, under *Principles of European Tort Law*, n 96, Art 4:102, which defines the required standard of conduct, such factors include "the nature and value of the protected interest involved, the dangerousness of the activity, the expertise to be expected of a person carrying it on, the foreseeability of the damage, the relationship of proximity or special reliance between those involved, as well as the availability and the costs of precautionary or alternative methods".

[114] Pierre Widmer, "Comparative Report on Fault as a Basis of Liability and Criterion of Imputation (Attribution)" in Pierre Widmer (ed), *Unification of Tort Law: Fault* (Kluwer Law International, 2005) 331, 349.

[115] As discussed in Part III.D.

[116] Borrowing from the language of *Principles of European Tort Law*, n 96, Art 3:106.

[117] As discussed in Part II.B.

[118] So-called "AI-assisted clinical reasoning" refers to systems that "can provide statistical reasoning and pattern recognition but interpretation and application to individual patients still require a clinician". See World Health Organization, *Emerging Trends and Technologies: A Horizon Scan for Global Public Health* (Report, 2022) <https://apps.who.int/iris/bitstream/handle/10665/352385/9789240044173-eng.pdf?sequence=1&isAllowed=y>. See also Reed T Sutton et al, "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success" (2020) 3(1) *NPJ Digital Medicine*, doi:10.1038/s41746-020-0221-y (highlighting issues regarding reproducibility, reliability, quality of data, algorithmic biases, and over-reliance).

[119] Ella Glikson and Anita Williams Woolley, "Human Trust in Artificial Intelligence: Review of Empirical Research" (2018) 14(2) *Academy of Management Annals* 627; Jakub Złotowski, Kumar Yogeeswaran and Christoph Bartneck, "Can We Control It? Autonomous Robots Threaten Human Identity, Uniqueness, Safety, and Resources" (2017) 100 *International Journal of Human-Computer Studies* 48.

[120] Waytz, Heafner and Epley, n 16.

a topic of its own that should be studied with the methodologies of cognitive and behavioural disciplines. Insights from that research are highly relevant for tort law, at least in two ways: first, they should inform the definition of the reasonable expectations for AI-based products or services; second, the regulator should ensure that AI innovators have the incentive to generate and *adequately* communicate sufficient information about AI risks.

While the due care standard of tortious liability can stimulate research into and development of safer innovation, tort law is not the only source of such incentives, as discussed in the next section.

D. The Intersection between Tort Liability and AI Safety Regulation

To effectively mitigate AI risks, one needs a robust understanding of the functional linkages between elements within AI systems and operative knowledge of the conditions that may lead to risk materialisation across the AI value chain. For that, there should be sufficient behavioural and economic incentives to generate such knowledge and align the level of care and the level of risk-creating activity with such knowledge. In the case of technological risks, the knowledge life cycle encompasses the stages of generating optimal knowledge (typically through basic and applied research), the adequate communication of information related to the risks of using technology-based products or services, the internalisation of such knowledge by the relevant parties across the value chain including end users, and the feedback loop whereby the experience of deploying technology refines the knowledge of technology developers about the associated risks and preventive measures.

If AI autonomy in the proper sense of agency (ie freedom from man-made rules or constraints) could be assumed, the legislators would need to give up on the idea of risk control and solely focus on damage compensation. Yet a closer look at the factors of AI risks suggests that there is ample room for human control and risk prevention.[121] The regulator's role in this regard is to ensure that intervention by regulatory measures fills in the gap where voluntary incentives for generating optimal knowledge about the factors of safety risk materialisation may fail. Typically, such measures take the form of ex-ante direct safety regulations[122] and ex-post tort liability rules.[123]

Safety regulation and tort law can be viewed as complementary and mutually reinforcing. First, their complementarity means that tort liability rules should ensure that victims are compensated for AI-mediated harm, given that safety regulations – despite the objective to minimise risks – may not guarantee that no damage will occur. Second, even though tort law also aims at incentivising risk-avoiding and caretaking conduct, designing adequate liability rules in the case of novel technologies – in particular, regarding the due care standards – is challenging due to the lack of information about technological risks. In this regard, compliance with mandatory safety requirements is likely to be more effective than the development of technology-specific legal standards of due care by courts. Furthermore, tort liability and safety regulation can be mutually reinforcing – for instance, the Proposal for EU AI Liability Directive treats safety standards established by the Proposal for EU AI Act as due care standards whose violation would trigger the presumption of fault.[124]

Given the common objective to minimise the likelihood of risks materialising and the magnitude of damage, the question is how the interface between tort liability and safety regulation should be optimally structured. In this regard, a distinction needs to be drawn between uncertainty regarding knowledge of *effective* measures of risk prevention (ie, which specific measures should be applied in practice to

[121] See, eg, Peter J Scott and Roman V Yampolskiy, "Classification Schemas for Artificial Intelligence Failures", *Cornell University arXiv* (E-print, 15 July 2019) <https://arxiv.org/ftp/arxiv/papers/1907/1907.07771.pdf>.

[122] As illustrated by the *Proposal for EU AI Act*, n 26, this type of regulation stipulates compliance rules, such as requirements for AI training data and the training process.

[123] For an overview of the relevant regulatory frameworks in the United States and European Union, see Michael Baram, "Liability and Its Influence on Designing for Product and Process Safety" (2007) 45(1) *Safety Science* 11.

[124] For an explanation, see n 90.

prevent the risk) and *efficiency*/optimality of the due care standard (ie, which combination of sufficiently effective measures would minimise the sum of precaution and accident costs[125]).

The question of which measures can effectively prevent or minimise AI risks and who should be in charge of undertaking them has been the subject of detailed technical analysis.[126] The EU legislature has taken the first steps in that direction.[127] Research shows that technological risks can materialise if there are flaws in the design of AI systems and if their performance is not properly tested, controlled, and monitored.[128] In particular, sources of AI risks include wrong background assumptions and other decisions made by AI developers that influence the predictive power of ML models, including unbalanced input data selection, but also third-party activity in bad faith, such as intentional data poisoning, model extraction, malicious cyberattacks, etc.[129] Measures to mitigate these risks can include, in particular, transparency obligations,[130] the use of explainable models,[131] and incorporating risk management systems and matrices for monitoring the system performance within ML-based systems that continue "learning" while in use. As discussed above, the deterministic nature of ML models means that uncertainty regarding their output arises due to objective factors such as correlations existing in the training data and the unpredictability of environmental factors. Such unpredictability and uncertainty can vary significantly (consider, for instance, automated driving on a highway vs. automated lawn trimming in a private garden). Hence, the level of the required precautionary measures should reflect the degree of risk involved.

Law-and-economics analysis finds that both fault and strict liability can induce socially optimal care and reduce the risk of harm, albeit under different conditions and with a different set of incentives for parties' activities.[132] Under the fault-based liability system, producers or suppliers of goods or services may adopt a higher level of care than they would under strict liability and hence impose higher/excessive costs of care on customers.[133] Conversely, a strict liability regime may discourage the level of activity, which in the case of innovation activity might not be socially advantageous, in view of the potential benefits of innovation. The challenge then lies in defining the relevant and optimal standard of care.

The European Union has taken a comprehensive approach encompassing strict and fault-based liability[134] and AI safety regulation[135] to address AI risks and harms. It remains for an economic analysis to determine whether the current framework is optimal from a benefit-cost perspective or whether it could be improved without sacrificing safety. Under the imperfect knowledge about the risks posed by novel technologies, it is understandable that considerations regarding the effectiveness of policy measures may prevail over their efficiency (ie, effectively mitigating such risks through the least onerous measures). Due to uncertainty and lack of knowledge, the legislature might be inclined to enforce more stringent safety standards that might be optimal or even necessary, to be "on the safe side". However, over time, such

[125] Aaron S Edlin, "Due Care" in Peter Newman (ed), *The New Palgrave Dictionary of Economics and the Law* (Palgrave Macmillan, 2002) 653.

[126] See, eg, Matija Franklin et al, "Causal Framework of Artificial Autonomous Agent Responsibility" (Paper presented at the Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022) 276.

[127] *Impact Assessment*, n 39.

[128] Renda et al, n 9, 62.

[129] Model extraction means an attack on an ML model where an attacker attempts to extract parts of different classes (categories or labels) that the model was trained on. Data drift refers to the change in the distribution of data used in a predictive task, which can cause an ML model to lose its predictive power over time. Data poisoning means an attack on the data of an ML model, where the attacker purposefully introduces malicious data to the training dataset in order to impair the performance of the model. For a more detailed explanation and a discussion of other safety risks related to ML models, see, eg, Renda et al, n 9, 62.

[130] *Report on the Safety and Liability*, n 53, 9.

[131] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019) 1 *Nature Machine Intelligence* 206.

[132] Evas, n 9, 34–35 (drawing on the chapter by Louis Kaplow and Steven Shavell, "Economic Analysis of Law" in Alan J Auerbach and Martin Feldstein (eds), *Handbook of Public Economics* (North Holland, 2002) Vol 3, 1661).

[133] Evas, n 9, 35.

[134] For the references, see nn 24–25.

[135] *Proposal for EU AI Act*, n 26.

safety standards should be revisited and adapted to reflect scientific progress and practical experience.[136] This approach would help avoid excessive costs of caretaking while allowing for the realisation of the benefits of innovation.

One final point should be mentioned regarding concerns that compliance regulation can impede innovation processes and incentives. Such arguments are rooted in a longstanding debate over regulation and innovation.[137] Yet it would be absurd to promote innovation for innovation's sake. The choice between innovation and safety is "a false dichotomy"[138] as the ultimate value of innovation should be seen in the improvement of social welfare. It is also worth noting that empirical studies show that the relationship between safety compliance regulation, liability, and innovation is more complex than "the simple view" that "liability chills innovation".[139] The peculiarity of regulating risks associated with innovation activity is evident in the fact that the policy objective should not be focused on reducing the level of activity itself, but rather on directing it towards socially beneficial outcomes. In other words, the goal is to maximise the social benefits while minimising the social risks of new technologies. At the level of policy objectives, the complementarity of innovation and safety goals is acknowledged,[140] and it remains to be seen whether substantive and operative legal rules applicable to AI will uphold this balance.

IV. Conclusion

AI technology should not become the "black hole" into which causation of harmful outcomes by human conduct disappears. This article has raised a concern that the widespread portrayal of AI as "autonomous", "unpredictable" and a "black box" might distort technology perception, mislead the fit-for-purpose analysis of the legal and regulatory framework,[141] and downplay the potential of human control in mitigating technological risks at the stage of system design and deployment. However, contrary to the widespread perception,[142] AI technology is not fundamentally incomprehensible or uncontrollable[143] but requires research to generate knowledge to ensure its safe application.

AI-based products and services have a highly complex value chain distinguished by multiple parties, elements and segments, inputs and activities. Such complexity, in turn, bears on the allocation and apportioning of liability for the suffered harm. This article has argued that risks and harm associated with AI technology are not inflicted *by* AI systems as such but are mediated through AI applications. The proposed view of distributed sources of AI risks holds the promise of enabling a more targeted approach to mitigating AI risks compared to the framing of AI systems as autonomous and incomprehensible agents. On closer examination, legal challenges brought to the fore by AI technology – such as

[136] Consider the evolution in the regulation of genome-editing technology or even more conventional small-molecule drugs. Regulation of genetically modified organisms (GMOs) in the European Union is a pertinent example. While GMOs have been subject to stringent regulation based on the precautionary principle, policymakers have acknowledged that the current approach to risk-assessment should be adapted in view of scientific progress and emerging new genomic techniques. See *Study on the Status of New Genomic Techniques under Union Law and in Light of the Court of Justice Ruling in Case C-528/16* (Commission Staff Working Document, SWD(2021) 92 final, 29 April 2021) 59ff.

[137] See, eg, Richard B Stewart, "Regulation, Innovation, and Administrative Law: A Conceptual Framework" (1981) 69(5) *California Law Review* 1256.

[138] *Product Liability: Hearings Before the Subcommittee on Consumer of the Committee on Commerce, Science, and Transportation, United States Senate, One Hundred Second Congress, First Session, September 12 and 19, 1991* (US Government Printing Office, 1992) Vol 4, 68 (citing the statement by Nicholas A Ashford).

[139] Evas, n 9, 35ff (with further references).

[140] *Report on the Safety and Liability*, n 53, 12 (stating that liability rules "always have to strike a balance between protecting citizens from harm while enabling businesses to innovate").

[141] Watson, n 16, 417 (arguing that "the anthropomorphic rhetoric [with regard to AI] is at best misleading and at worst downright dangerous" and that "[t]he impulse to humanize algorithms is an obstacle to properly conceptualizing the ethical challenges posed by emerging technologies").

[142] As discussed in Part II.A.

[143] As discussed in Part II.B.

uncertainty regarding the adequate due care standard and safety measures[144] or the diminishing role of causation in the legal assessment[145] – are not novel or unique challenges for tort law. Consequently, the appropriateness of the existing concepts and rules in tort law should be carefully considered before re-inventing them.

Discussions about AI technology reveal a recurring theme that often arises with breakthrough technologies. This theme involves the emergence of myths and misconceptions that stem from a lack of communication between experts in technology and law, resulting in misunderstandings about the legal implications of technological developments.[146] The emergence and deployment of ML technology may not call for a radical transformation of liability frameworks but it does necessitate their incremental adjustment. A sounder depiction of AI technology – devoid of anthropomorphic metaphors – would yield a more rigorous analysis of the challenges for liability law and how liability frameworks should respond to them. The point worth reiterating is that technologies are not inherently negative or positive, socially advantageous or disadvantageous, safe or unsafe. Rather, it all depends on the way they are applied and the purposes they serve. Even in the case of highly distributed technological risks, humans can and should retain agency over risk control, the internalisation of AI-mediated harm, and the benevolent use of technology.[147]

[144] For a classical law-and-economics account of uncertainty regarding the legal standard of due care, see Richard Craswell and John E Calfee, "Deterrence and Uncertain Legal Standards" (1986) 2(2) *Journal of Law, Economics, & Organization* 279.

[145] Judith J Thomson, "The Decline of Cause" (1987) 76 *The Georgetown Law Journal* 137, 138–139 (observing a tendency that cases where causality could not be proved nevertheless were decided in favour of the plaintiff and pointing out the "increasing dismissiveness about causality […] in legal theorizing [and] '[t]he Decline of Cause' in moral theorizing").

[146] For the examples in the context of tort law, see Kyle Graham, "Predicting the Future in Tort Law: Applying Forecasting Science to Innovations from Trampolines to Autonomous Vehicles" (2022) 62(3) *Jurimetrics* 303. I thank the reviewer for bringing this paper to my attention.

[147] In the words of ML researcher Robert C Williamson, "rather than worrying about AI 'making decisions' about us, we should pay more attention to who commissioned the chain of technological action using AI rather than the technology itself". Williamson, n 77.