1.

AI Liability in Europe: Addressing Harms to Fundamental Rights and the Problem of the Human in the Loop

Beatriz Botero Arcila *

2. INTRODUCTION

Who should compensate you if you get hit by a Tesla in "autopilot" mode: the safety driver or the car manufacturer?[1] What about if you find out you were unfairly discriminated against by an AI decision-making tool that was being supervised by an HR professional? Should the developer compensate you, the company that procured the software, or the HR professional that was "supervising" the system's output?[2] Does anything change if the harm occurs when a doctor uses, or decides not to use, an AI system to assist in a procedure?[3]

These situations all involve the liability for harms that are caused by or with an AI system. In the examples above, there is also an individual participating in the process that led to the harm. Who should be held liable, the AI, the human, both? Why? As it turns out, these questions do not have easy answers. This Article explores these questions in the context of the European Union's ongoing efforts to regulate artificial intelligence (AI) and, in particular, adjust civil liability to the complexities of AI systems.

---

[1] Andrew J. Hawkins, *The world's first robot car death was the result of human error – and it can happen again*, The Verge, (Nov. 20, 2019) https://www.theverge.com/2019/11/20/20973971/uber-self-driving-car-crash-investigation-human-error-results .

[2] Meilssa Heikkilä, *Dutch scandal serves as a warning for Europe over risks of using algorithms,* Politico.eu (March 29, 2022) https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/?tpcc=nleyeonai.

[3] *See also*, James Vincent, *OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host*, The Verge, (June 9, 2023) https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit .

When a harm occurs, liability law, an ex-post mode of regulation, requires a sufficient justification for shifting the loss from the person who suffered the damage (the victim) to the person or legal entity who allegedly caused the damage (the injurer).[4] While liability laws vary across countries, in the EU they typically require that the victim show that (1) they suffered a harm or loss, (2) that the injurer was at fault, and (3) that there is a link of causality between the fault and the loss.[5]

Establishing liability when an AI system is involved requires establishing *who* is responsible for the action that led to the harm and establishing that that action was in breach of a duty of care (fault). This isn't, a priori, impossible in AI related harms: Humans are everywhere in automated decision-making systems, not only right at the end as the examples above perhaps suggest: They design them, select the training data and inputs, ask the questions that the system answers, implement the conclusions, and often conduct ex-post evaluations.[6] Similarly, liability law already deals with many other complex industries, such as the energy industry or the pharmaceutical industry. What is particular about AI systems, however, is that AI systems are opaque in the sense that recipients of outputs rarely have a concrete sense of how that output was arrived at from inputs;[7] complex in the sense that AI systems involve many actors and parts and their behavior arises in non-linear ways from those of their parts;[8] and autonomous, in the

---

[4] Christiane Wendehorst, *Liability for Artificial Intelligence: The Need to Address Both Safety Risks and Fundamental Rights Risks*, The Cambridge Handbook of Responsible Artificial Intelligence (Silja Voeneky et al., eds. 2022), at 192.

[5] *See* e.g. European Group on Tort Law, Principles of European Tort Law, available at: http://egtl.org/docs/PETL.pdf [*hereinafter* PETLs]; Miriam Buiten, Alexandre de Streel and Martin Peitz, *The law and economics of AI liability*, 48 Computer Law & Security Review (2023), at 4. These are the regular requirements in the EU, where most countries come from a civil law tradition. In Common Law countries, such as the US or Canada, these three requirements are usually presented as four: that (1) an actual harm or loss occurred, (2) that the injurer owed a legal duty to the plaintiff to act reasonably and avoid causing harm, (3) that the injurer breached their duty of care, (4) and that the breach of duty directly caused the victim's injuries or damages. *See* Samuel Beswick Tort Law: Cases and Commentaries (2022), available at: https://commons.allard.ubc.ca/fac_pubs/705/ .

[6] *See* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop,* 76 Vand. L. Rev. 429 (2023), at 443 ; Meg Leta Jones, *The Ironies of Automation: Tying Policy Knots with Fair Automation Practice Principles,* 18 Vand. J. of Ent. & Tech. L. 77 (2015), at 84.

[7] *See* Jenna Burrel, *How the machine 'think': Understanding opacity in machine learning algorithms*, 3 Big Data & Society 1 (2016).

[8] *See* Jennifer Cobbe & Jatinder Singh, *Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges*, 42 COMPUTER L. & SECURITY REV.

sense that those outputs can be of a kind that they were not explicitly programmed to produce. These characteristics complicate tracing back the damage to one behavior specifically, make liability claims difficult or over-costly to prove for victims of AI harms and creates legal uncertainty for AI producers and deployers.[9] For similar reasons, some AI harms are harder to disincentivize through litigation.[10]

In human-AI hybrid systems, an additional challenge is identifying what should have been the responsibility of the person who was working with or supervising the algorithm when the harm occurred. So-called "humans-in-the-loop" are very often inserted into AI systems to respond to concerns about bias or safety risk in AI systems.[11] They focus on placing a human at the end of the system to oversee its output or decision-making processes of the AI. This is based on the truth that humans are better at complex and contextual analysis, while AI systems are better at repetitive tasks.[12] Human machine complementarity is thus very promising when neither the state-of-the-art algorithm nor the human dominate performance in all instances.[13]

Researchers have shown, however, that the relationship between humans and machines in hybrid systems is more complex than that and challenges this basic assumption: adding a human into a machine system rarely results in the best of both worlds. Humans often defer to machines, deviate from

---

105573, 39 (2021); Ian Brown, *Expert explainer: Allocating accountability in AI supply chains*, Ada Lovelace Institute, June 29, 2023 https://www.adalovelaceinstitute.org/resource/ai-supply-chains/ ; Arvid Narayanan, Understanding Social Media Recommendation Algorithms, Knight First Amendment Institute at Columbia University (2023) available at https://s3.amazonaws.com/kfai-documents/documents/4a9279c458/Narayanan---Understanding-Social-Media-Recommendation-Algorithms_1-7.pdf , a 11.

[9]Commission Report on safety and liability implications of AI, the Internet of Things and Robotics, 16, https://commission.europa.eu/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en (last visited Aug 26, 2023) [*hereinafter* European Commision, *Report on the safety and liability implications of AI*].

[10] *See* Margot E. Kaminski, *Regulating the Risks of AI*, 103 Boston U. L. Rev. (forthcoming 2023), at 18.

[11] *See* European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and amending Certain Union Legislative Acts, Sec (2021) 167 final, Art. 14 [*hereinafter* AI Act] *see also* Ben Green, *The flaws of policies requiring huma oversight of government algorithms*, 45 Comp. L. & Sec. Rev. (2022) (surveying 41 policies in the US that prescribe human oversight of government algorithms).

[12] Crootof et. al, *supra* note 6, at 438.

[13] Ruijang Gao,et al. *Human-AI Collaboration with Bandit Feedback,* Accepted at IJCAI 2021, a 1 https://arxiv.org/pdf/2105.10614.pdf .

algorithmic advice in biased ways, or may not be able to take effective control of a situation when control is transferred to them.[14] Despite this, scholars in the US have shown that courts tend to locate liability for automated systems in the human in the loop, and not in the company that deployed or designed the algorithm.[15] In Europe, an expert group convened by the European Commission pointed out that liability laws in EU Member States tend to assume that the human who is on the front end of a system - such as the driver of an automated vehicle - is in full control of the system, and is likely to be held responsible for an accident involving a hybrid system.[16]

This situation is problematic from a liability perspective because the general objectives of liability laws are to provide corrective justice for victims of illegal harm and to create incentives to avoid harm for those who are well positioned to take measures. If the human in the loop is being held liable but is ultimately less capable of taking measures to avoid harm this objective is not being met. If they also have less control over the system, there is also a corrective justice question about them being held liable. How should liability law account for the particularities of AI systems, and the complexity of human-AI interactions?

In the fall of 2022, the European Union launched two directive proposals that seek to update its civil liability rules to meet the challenges posed by AI and answer some of these questions: An AI Liability Directive (AILD) and a revision of the Product Liability Directive (PLD). The proposed AILD and PLD complement the EU's risk regulation, the Artificial Intelligence Act (AI Act), and seek to facilitate the challenges victims of AI harms may face when seeking redress, given the characteristics of AI systems.

Briefly, the PLD Proposal updates the EU's current product liability framework and extends it to software, including AI systems. It is restricted to damages suffered by natural persons to life, health, property, and loss of data. In these realms, it harmonizes Member States' strict liability rules for manufacturers of AI systems.[17] The proposed AILD applies to all harms

---

[14] Crootof et. al, *supra* note 6, at 469 *see infra* Part I, C.

[15] Ryan Calo, *Robots in American Law* (U. Wash. Sch. L. Legal Stud. Rsch. Paper No. 04, 2016), at 24, available at https://duckduckgo.com/?q=Ryan+Calo%2C+Robots+in+American+Law+36+(U.+Wash.+Sch.+L.%2C+Legal+Stud.+Rsch.+Paper+No.+04%2C+2016)%2C+https%3A%2F%2Fpapers.ssrn.com%2Fsol3%2Fpapers.cfm%3Fabstract_id%3D2737598&t=brave&ia=web .

[16] *See* Expert Group on Liability and New Technologies, New Technologies Formation, Liability for Artificial Intelligence and other emerging digital technologies (2019).

[17] *See infra* Part III, B.

caused by or with an AI system, including affectations to fundamental rights, and harmonizes procedural aspects of fault-based liability under Member State law. In essence, under the AILD, claimants of damages caused by high-risk AI systems will be empowered to request information from providers that may be relevant evidence for a claim. The AILD also establishes a rebuttable presumption of the causal link between the damage and the fault of the defendant. These rules not only cover claims against manufacturers but also against professional and non-professional users. Interestingly, one of the main objectives of the AILD is to enable the effective private enforcement of fundamental rights when AI risks to fundamental rights materialize.[18]

This Article focuses, in particular on how these proposals address the information asymmetries that victims of AI harms face when seeking to prove the elements of liability, the complexities of establishing liability in human-AI hybrid systems and, given the particularities of the EU's regulatory framework, its fitness to guarantee the protection of fundamental rights. Importantly, this Article does not consider the role of insurance - which is a main issue and factor in contemporary liability law and policy - and seldomly focuses on the difficulties and particularities of Member States' liability regimes and the EU-specific difficulties of harmonizing them.[19]

To begin exploring these issues, this Article proceeds as follows, Part I serves as the thematic background section and introduces the general difficulties of regulating AI, human-AI hybrid systems, and ensuring AI accountability in general. It begins by making a non-exhaustive survey of the risks posed by AI, and a discussion of the characteristics of AI systems that complicate AI accountability. It then presents three real-world situations where harm occurs by or with the participation of a human-AI hybrid system and discusses them in light of the socio-technical research on human-AI systems and interactions.

Part II discusses the specific challenges that the characteristics of AI systems, and human-AI hybrid systems pose for liability law. Specifically, it

---

[18] *See infra* Part III C.

[19] As explained by Professor Christiane Wendehorst, "regulating AI liability is more complicated than regulating the issues addressed by the AI Act and other forms of digital regulation, precisely because Member States already have very sophisticated and longstanding liability rules, so it is not obvious from the outset that action by the EU institutions is required." *See* Christiane Wendehorst, *AI Liability in Europe: anticipating the EU AI Liability Directive*, Ada Lovelace Institute (2022), at 6, https://www.adalovelaceinstitute.org/wp-content/uploads/2022/09/Ada-Lovelace-Institute-Expert-Explainer-AI-liability-in-Europe.pdf .

focuses on the different challenges victims of an AI harm face when establishing the main elements of fault based liability: (1) that a "damage" occurred (whereas damages to life or property are easy to prove, an illegal affectation to a fundamental right may be harder to prove); (2) that the AI provider or deployer acted with fault, especially when AI systems are complex, opaque or autonomous, (3) and causality, similarly, because AI's technical and organizational opacity, complicates identifying how a bug or the process behind an AI system led to a particular incident - except, of course, when there is a human right at the end that is easier to blame. This Part also outlines the policy and distributive implications of choosing a liability regime - such as shifting the burden or proof or choosing strict liability - both for EU development and deployment in the EU and for the humans involved in AI systems.

Part III presents and analyzes the European Union's AI Liability rules, the proposed AILD and PLD. It situates them within the general regulatory framework introduced by the AI Act and analyzes them by discussing how they would apply to two hypothetical case studies: one pertaining to a safety harm, and another one a harm to a fundamental right. In doing so, it surfaces some of the still remaining limitations of these proposals. Information asymmetries will persist in cases that do not pertain to high-risk systems; the regime is contradictory in how it treats hybrid systems, and the idoneity of the AI liability regime to protect fundamental rights raises several questions, specifically because liability law typically requires that harm occurs.

Part IV concludes by providing suggestions for how EU policymakers could address some of these challenges in the AI liability regime. These include (a) extending courts' power to request sufficient evidence to systems that are opaque and complex, regarding whether they are high risk or not; (b) expanding the application of the AILD to all sorts of human-AI hybrid systems, regardless of the particular role the human at issue is supposed to play; (c) extend some of the procedural innovations of the AI liability regime to other procedures and regulations focused in the protection of fundamental rights.

3.

I. AI, HUMANS, AND AI ACCIDENTS

What is particular about artificial intelligence systems and their interaction with human operators? How does the legal system react when AI harm occurs?

This Part outlines the now well-known specific risks posed by AI and presents three case studies that outline the liability questions that arise when AI harm occurs and there is a human in the loop. It then finishes with an overview of the research on human-AI interactions, which suggests that the human operators and supervisors do not necessarily have as much control over AI systems as human-in-the-loop requirements and current liability rules tend to suggest.

A. AI Harms and Risks

There is a vast literature on the benefits and risks of AI systems.[20] It is well recognized that AI systems can enhance efficiency and productivity, and enable more accurate data analysis, aiding in better decision-making in a variety of fields.[21] At the same time, it is also well documented that AI systems pose several risks and can cause a variety of harms.[22] AI systems like automated vehicles or appliances can pose safety risks, to life, bodily integrity, or property; AI-powered decision-making software poses risks to fundamental rights, privacy, human dignity, and equality; and AI systems also pose epistemic risks, for example, they may change how we conceptualize the world via statistical knowledge and by increasingly relying on profiling or sorting algorithms.[23]

---

[20] AI is used in this piece to refer to software systems that analyze their environment and take actions – with some degree of autonomy – to achieve specific goals, displaying some form of so-called intelligent behavior. *See* European Commission, Communicatio From the Commission, Artificial Intelligence for Europe, {SWD(2018) 137 final}; The European regulations seem to have adopted the OECD definition after long debates on the matter, which defines an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate output such as predictions, recommendations, or decisions influencing physical or virtual environments. Lucas Bertuzzi, *EU lawmakers set to settle on OECD definition for AI*, Euractiv (Mar. 7 2023), https://www.euractiv.com/section/artificial-intelligence/news/eu-lawmakers-set-to-settle-on-oecd-definition-for-artificial-intelligence/.

[21] *See* European Commission, White Paper on Artificial intelligence. A European approach to excellence and trust, Brussels, Feb. 19, 2020, COM(2020) 65 final, (2020) [*hereinafter* White Paper on AI] (eventually add more footnotes)

[22] *But see* Margot E. Kaminski, *Regulating the Risks of AI*, 103 Boston U. L. Rev. (forthcoming 2023), at 18. (arguing that labeling harm as risk is not only descriptive but also a normative move. As she explains, "some harms are aggregate in nature, do raise complex causal issues, and might best be dealt with ahead of time. But labeling harm as risk also constructs the problem in particular ways, invoking a specific set of legal practices and policy conflicts.") at. 9.

[23] *See e.g.* Juan Ortiz Freuler, Dataification, *Identity, and the Reorganization of the Category*

What is particular about AI from a liability perspective, however, is that when harms occur AI systems' characteristics makes them hard to scrutinize. Take the example of Machine learning (ML) algorithms which are used in many of the AI-powered tools consumers are often in contact with, such as assisted driving, healthcare, and home appliances like Amazon's Alexa.[24] They are also used to make classifications, predictions and to decide what can be the best action in a particular situation.[25] ML algorithms make classification decisions (for example whether a given object is a person or not, or whether someone should get alone), using analytics algorithms that work with high-dimension data to determine what features are relevant to that decision. The number of features can run into the tens of thousands which, even if it is replicating work done by humans, involves a qualitatively different decision-making logic from that of humans.[26] Trained machine learning algorithms define decision-making rules to handle new inputs which do not need to be understood by a human operator.[27]

The high dimensionality and illegibility of ML accentuate their "riskiness" from an accountability perspective. Harms caused by AI systems may be hard to detect, it is hard to find or understand their cause, and access or explanations to how they reach certain answers may not be available; they can reach conclusions that they were not programmed to reach; and so many parts and actors can intervene in their development and deployment that is is rarely straightforward to ascertain the cause or responsible party behind harm that arises. This is to say, AI systems are opaque, autonomous and complex.

Individual, 65 Temple L. R. 4 (2023); Brent D. Mittelstadt et al. *The ethics of algorithms: Mapping the debate,* 3 Big Data & Society 2 (2016).

[24] Bernard Marr, *Machine learning in Practice: How Does Amazon's Alexa Really Work?,* Bernard Marr & Co. (n.a.) available at: https://bernardmarr.com/machine-learning-in-practice-how-does-amazons-alexa-really-work/; *How Machine Learning is Used in Autonomous Vehicles,* Rinf.Tech (n.a.) available at: https://www.rinf.tech/how-machine-learning-is-used-in-autonomous-vehicles/#:~:text=An%20autonomous%20vehicle%20can%20use,the%20world%20around%20a%20car.

[25] ML is broadly defined as "any methodology and set of techniques that employ data to come up with novel patterns and knowledge and generate models that can be used for effective predictions about the data." *See* Mittelstadt et al. *supra* note 23 (citing Martin Van Otterlo, A machine learning view on profiling. In: Hildebrandt M, de Vries K (eds) *Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology*, Abingdon: Routledge (2013) at 41–64).

[26] *See* Mittelstadt et al. *supra* note 23, at 3.

[27] *See* Mittelstadt et al. *supra* note 23 (citing Mathias).

What follows explains some of these properties, often interrelated, in some more detail, as they represent key challenges for AI accountability.
1.

1. Opacity

Algorithms are opaque in the sense that recipients of the output of an algorithm rarely have a concrete sense of how the output was arrived at from the inputs – or what those inputs were.[28] This is complicated from a liability perspective because to succeed in a liability claim, plaintiffs must show that the output caused them harm, and that the output was produced because of the faulty behavior of someone (a deployer or developer for example).

Algorithms are opaque in different ways. Many forms of opacity are a function of proprietary protection of corporate or state secrecy, or because of generalized technical illiteracy.[29] Intentional secrecy and technical illiteracy hinder tort lawsuits because they obstruct effective inspection of AI systems – either because developers assert confidentiality, because there is inadequate documentation, or because of the difficulty in scrutinizing their forms of "reasoning."[30] Opaqueness by virtue of corporate secrecy or technical illiteracy can be addressed by making code available for scrutiny through regulatory means or by widespread educational efforts to make key actors (i.e. journalists) or the public at large more knowledgeable about these mechanisms.[31] Part III,discusses how the EU liability framework does some of this.

Algorithms can be opaque in a more fundamental way, however, by virtue of a mismatch between the forms of human reasoning and semantic interpretation and the "mathematical optimization in the high-dimensionality characteristic of machine learning,"[32] This means that when a computer learns and builds its own representation (correlations and probabilistic reasoning) to classify objects or predict the likelihood of an event it does so

---

[28] Burrel, *supra* note 7, at 1.

[29] *Id.* at 3.

[30] In addition, access to the algorithm and the data could be impossible without the cooperation of the potentially liable party. In practice, victims may thus not be able to make a liability claim. In addition, it would be unclear, how to demonstrate the fault of an AI acting autonomously, or what would be considered the fault of a person relying on the use of AI. *See* Henry Fraser, Rhyle Simcock, Aaron J. Snoswell, *AI Opacity and Explainability in Tort Litigation*, FAccT' 22, June 2022,, at 1; European Commision, *Report on the safety and liability implications of AI supra* note 9, at 16.

[31] *Id.* at 4.

[32] Burrel, *supra* note 7, at 2.

without regard for human semantic comprehension and reasoning modes.[33] The difficulty in understanding how machines make choices is aggravated as computational resources expand and the number of features grows way beyond what can be grasped by a human.[34] The scale, complexity and non-linear characteristics of many complex AI systems may make them inherently inscrutable, even to their developers.[35]

Organizational and technical opacity often reinforce each other. In 2019, for example, the Dutch Data Protection Authority found that a system that had been used by the Dutch Tax Authority since 2013 to allocate some subsidies tended to systematically identify high-risk claims by parents with double citizenship – an irrelevant feature. This was discriminatory, and the use of such data was unlawful, but it was hard to identify – it took 6 years! - and, later, hard to understand.[36]

2. Complexity

AI systems are complex in two ways. They are complex in the sense that they are socio-technical systems where a variety of actors and elements participate throughout the system's life cycle. They are also complex in the sense that their behavior arises in a nonlinear, often unpredictable way from that of its parts.[37]

The first one, often referred to as the problem of many hands, refers to the fact that several tasks need to be completed in the development and deployment of an AI system, from problem definition to data collection, labeling, cleaning, model training and fine-tuning, and testing and deployment. While some developers do all these inhouse, these activities are

---

[33] *Id.* at 2, 10

[34] *Id.* at 9.

[35] *See* Burrel, *supra* note 7 at 8 ; Stefan Buijsman & Herman Veluwenkamp, *Spotting When Algorithms Are Wrong*, Minds & Machines (2022), https://doi.org/10.1007/s11023-022-09591-0 (last visited Sep 1, 2023). at 16 ;
Fraser et al. *supra* note 30, at 1.

[36] EU Law Enforcement, *The Dutch benefits scandal: a cautionary tale for algorithmic enforcement*, EU Law Enforcement, (Apr. 30, 2021), https://eulawenforcement.com/?p=7941; Dutch DPA, *Methods used by Dutch Tax Administration unlawful and discriminatory,* July 17, 2020 https://autoriteitpersoonsgegevens.nl/en/current/methods-used-by-dutch-tax-administration-unlawful-and-discriminatory .

[37] *See* DONELLA H. MEADOWS, THINKING IN SYSTEMS: A PRIMER (Chelsea Green Publishing, 2008).

often carried out by different actors in an AI supply chain.[38] Additionally, AI systems are not static and can be subject to frequent or continuous change. AI systems change after having been placed in the market through online updates, but also through data feeds, cloud-based services and, as described above, self-learning. The plurality of actors makes it increasingly difficult to find out who might be liable for the damage caused, and the most salient or immediate causal antecedent of blame may not converge with the locus of decision-making.[39] An example of this situation is OpenAI's descriptions of its services associated with ChatGPT. It foresees, for example, that it will "pre-train" and "fine-tune" these models, but that these will then be "customizable by each user up to limits defined by society."[40] It will be hard to determine which actor is making the key decisions, and to what extent providers, customers, and even end-users will be considered liable for harm in this scenario.

This is aggravated by the second way in which AI systems can also be complex. System complexity also refers to the idea that their outputs are "emergent," which means that outputs arise in nonlinear, often unpredictable ways from those of their parts.[41] Complex systems are subject to feedback loops that are often unpredictable. Research on social networks shows, for example, that social media is a complex emergent property where user

---

[38]Ian Brown, *Expert explainer: Allocating accountability in AI supply chains*, Ada Lovelace Institute, June 29, 2023 https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

[39] Expert Group on Liability and New Technologies *supra* note 16 at 33 *see also* Helen Nisenbaum already observing in 1996, "[w]here a mishap is the work of 'many hands,' it may not be obvious who is to blame because frequently its most salient and immediate causal antecedents do not converge with its locus of decision-making. The conditions for blame, therefore, are not satisfied in a way normally satisfied when a single individual is held blameworthy for a harm." *see* Helen Nissenbaum, Accountability in a computerized society. *Science and engineering Ethics,* 2(1), 29. But see Noorman arguing that "in order to "qualify" as the problem of many hands, the component decisions should be benign, or at least far less harmful if examined in isolation; only when the individual decisions are collectively combined do we see the most harmful result. In this understanding, the individual decision-makers should not have the same moral culpability as they would if they made all the decisions by themselves." M. Noorman, *Computing and moral responsibility.* The Stanford Encyclopedia of Philosophy.

[40] OpenAI, How Should AI Systems Behave, Who Should Decide?, OPENAI, https://perma.cc/6A3TU4UR; James Vincent, *OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host*, The Verge, (June 9, 2023) https://www.theverge.com/2023/6/9/23755057/openai-chatgpt-false-information-defamation-lawsuit .

[41]Arvid Narayanan, *Understanding Social Media Recommendation Algorithms*, Knight First Amendment Institute at Columbia University (2023); Available at https://s3.amazonaws.com/kfai-documents/documents/4a9279c458/Narayanan---Understanding-Social-Media-Recommendation-Algorithms_1-7.pdf , at 11.

behavior interacts with content and social media algorithms, and where the reach and virality of a particular post are basically unpredictable.[42] The unpredictability of outcomes and of the impact that certain actions can have in complex systems also challenges the traditional conditions for assigning blame.[43]

3. Autonomy and generativity

Autonomy refers to AI systems' ability to perform a task without every step of the task being pre-defined and with little or no human control or supervision. It has been criticized, as a term, for its association with human will, yet it is simply a result of AI systems' self-learning capacity stemming from their mathematical optimization in high-dimensionality processes, the same leading to opacity.[44] The self-learning features of AI autonomy may, for example, lead an AI system to alter the initial algorithms due to their self-learning capacity from the environment.[45] This same feature makes it difficult for outsiders to predict with full accuracy how a system will behave in the future.[46] It is also a priori unclear how much should an AI producer be expected to foresee some of these changes.[47]

Generativity has come to the general attention with newer AI systems, such as multimodal models capable of processing image and text inputs and producing text outputs that replicate human linguistic patterns. However, it is also well known that despite its ability to generate convincing answers, some of these models, like ChatGPT, are not fully reliable. They often replicate biases from training data, but they also produce made-up facts and make reasoning errors.[48] In New Zealand, an app put in place by a supermarket to generate meal plans was reported offering recipes for chlorine gas, when customers entered a wider range of household shopping list items.[49] A notice on the meal planner warns that recipes are not reviewed by

---

[42] Arvid Narayanan, Understanding Social Media Recommendation Algorithms, Knight First Amendment Institute at Columbia University (2023) available at https://s3.amazonaws.com/kfai-documents/documents/4a9279c458/Narayanan---Understanding-Social-Media-Recommendation-Algorithms_1-7.pdf , at 16.

[43] Helen Nissenbaum, Accountability in a computerized society. *Science and engineering Ethics,* 2(1), 29.

[44] European Commision, *Report on the safety and liability implications of AI supra* note 9.

[45] *Id.*.

[46] Expert Group on Liability and New Technologies *supra* note 16, at 35.

[47] European Commision, *Report on the safety and liability implications of AI supra* note 9.

[48] OpenAI, GPT 4 Technical report (Mar. 27, 2023) https://arxiv.org/pdf/2303.08774.pdf

[49] Tess McClure, *Supermarket AI Meal Planner App Suggests Recipe That Would Create*

humans, and that users must use their own judgment before relying on making any recipe.[50] OpenAI warns that "[g]reat care should be taken when using language model outputs, particularly in high-stakes contexts, with the exact protocol (such as human review, grounding with additional context, or avoiding high-stakes uses altogether)."[51]

If harm occurs, however, there remains uncertainty regarding how or if liability should be assigned to the developers or deployers of generative AI systems or what would be the responsibility of the person relying on the use of the AI system.

b. B. Three AI Accidents

So, what happens when AI systems cause harm? How does the legal system frame and react to these disputes? This section provides three case studies of real-world examples.

1. AI and Safety: The case of Tesla's Autopilot[52]

Since 2019, Tesla's Autopilot feature has been involved in at least 736 crashes only in the United States, including 17 fatalities.[53] Here we look at

---

Chlorine Gas, The Guardian, Aug. 10, 2023, https://www.theguardian.com/world/2023/aug/10/pak-n-save-savey-meal-bot-ai-app-malfunction-recipes (last visited Aug 25, 2023).

[50] *Id.*

[51] OpenAI, *supra* note 48.

[52] In France, an accident involving a Model 3 Tesla claimed the life of one person and injured a further 21 in Paris in December 2021. The driver, a 57-years-old off-duty Taxi driver was driving his family to dinner in his Tesla when he claims he lost control of the car, which then accelerated to a speed of 180 kilometers per hour down an avenue before crashing into a series of metal poles and a white van. Despite Tesla claiming that they could not detect any sign of technical fault on the autopilot's part (Tesla has previously alleged that they are able to tell exactly what happens onboard their vehicles as a result of the data they collect off of their onboard cameras and sensors), investigators for the STAJ, the French accident judicial service, "suspect that the Tesla suffered a technical failure" according to an article by the Telegraph. The driver is now filing a criminal complaint against Tesla on the grounds that their vehicles put the lives of its users and others in danger. Henry Samuel, *'Accelerator Went Ballistic': French Driver Files Complaint against Tesla after Fatal Crash,* The Telegraph, Mar. 21, 2022, https://www.telegraph.co.uk/world-news/2022/03/21/accelerator-went-ballistic-french-driver-files-complaint-against/ (last visited Aug 25, 2023).

[53] Sebastian Blanco, *Report: Tesla Autopilot Involved in 736 Crashes since 2019,* Car and Driver, June. 13 2023, https://www.caranddriver.com/news/a44185487/report-tesla-

one recent ruling in which the driver was also "the victim" of the accident:

In 2019, Justine Hsu, a resident of Los Angeles was involved in an accident while using Tesla's Autopilot feature in her Model S: The vehicle swerved into a curb, causing the airbag to deploy violently, which resulted in injuries to her face, including a fractured jaw, knocked-out teeth, and nerve damage. In 2020, Hsu sued Tesla, alleging defects in the design of the airbag, and alleging that the Autopilot feature had failed to operate safely. Tesla denied liability for the accident and argued in its court filing that Hsu used Autopilot on city streets, despite a user manual warning against doing so.[54]

Tesla calls its driver-assistant systems Autopilot or Full Self-Driving, and at least until October 2022, a video on the company's website says "The person in the driver's seat is only there for legal reasons. He is not doing anything. The car is driving itself."[55] However, the company has also explicitly warned drivers that drivers should keep their hands on the wheel and be "prepared to take over at any moment."[56] Ms Hsu's attorney remarked, however, that she only received a warning to put her hands on the wheel less than a second before the strike.[57]

The jury of the case reached a verdict stating that Tesla's Autopilot feature did not fail to perform safely (it also found that the airbag did not fail to perform safely, but that is not our issue here).[58] One of the jurors explained to the press that one of the key factors supporting their decision was that Tesla had clearly warned that the partially automated driving software was not a

---

autopilot-crashes-since-2019/ (last visited Aug 26, 2023).

[54] Abhirup Roy, Dan Levine & Hyunjoo Jin, *Tesla Wins Bellwether Trial over Autopilot Car Crash*, Reuters, Apr. 22, 2023, https://www.reuters.com/legal/us-jury-set-decide-test-case-tesla-autopilot-crash-2023-04-21/ (last visited Aug 25, 2023).

[55] Mike Spector, Dan Levine & Mike Spector, *Exclusive: Tesla Faces U.S. Criminal Probe over Self-Driving Claims,* Reuters, Oct. 27, 2022, https://www.reuters.com/legal/exclusive-tesla-faces-us-criminal-probe-over-self-driving-claims-sources-2022-10-26/ (last visited Aug 25, 2023).

[56] *Id.*

[57] Abhirup Roy, Dan Llevine, and Hyunjoo Jin, *Exclusive: Tesla's Autopilot never claimed to be self-pilot, juror says,* Reuters, April 22, 2023 https://www.reuters.com/business/autos-transportation/teslas-autopilot-never-claimed-be-self-pilot-juror-2023-04-21/#:~:text=After%20the%20verdict%20on%20Friday%2C%20juror%20Mitchell%20Vasseur%2C%2063%2C,confessed%20to%20be%20self%20pilot. (last visited Aug 24. 2023)

[58] Abhirup Roy, Dan Levine & Hyunjoo Jin, *Tesla Wins Bellwether Trial over Autopilot Car Crash*, Reuters, Apr. 22, 2023, https://www.reuters.com/legal/us-jury-set-decide-test-case-tesla-autopilot-crash-2023-04-21/ (last visited Aug 25, 2023).

self-piloted system and that the direct fault for the accident was the driver's distraction.[59] The jury awarded Hsu zero damages.[60]

It is worth noting that in 2022, the US National Highway Traffic Safety Administration launched an investigation on Autopilot because over a period of four years a dozen Tesla cars crashed into parked first respondent vehicles. The investigation is focused on the Autopilot feature and is specifically looking into whether Autopilot feature ultimately undermines "the effectiveness of driver's supervision."[61] At the time of writing, however, there is no apparent response from Tesla to the investigation, and it seems Tesla will be subject to civil penalties for not responding to NHTSA's requests.[62]

2. AI and Safety: The case of the Boeing 737 Max aircraft

In late 2018 and early 2019 two Boeing 737 MAX aircraft crashed in Indonesia and Ethiopia minutes after takeoff, killing almost 250 people in total. The 737 MAX is the fourth generation of the insignia Boeing 737. It has more efficient engines and some aerodynamic changes and, at the time of the accident, included a new software system to support navigation. The 737 MAX was certified by the United States Federal Aviation Administration in March 2017 and released into the market in May 2017,[63] but the entire fleet of 737 MAX Jets was grounded after these two accidents. An improved version was cleared for flying again at the end of 2020.[64]

After the accidents, Boeing initially blamed the pilots and argued that the pilots could have prevented the accidents if they had followed the security protocols for the 737.[65] This is important because airplane crash liability law

---

[59] *Id.*

[60] *Id*

[61] Lauren Aratani, *Tesla Investigation Deepens after More than a Dozen US 'Autopilot' Crashes,* The Guardian, Jun. 9, 2022, https://www.theguardian.com/technology/2022/jun/09/tesla-autopilot-crashes-investigation-nhtsa (last visited Aug 25, 2023).

[62] Ars Contributors, *Fed's Deadline Comes and Goes without Tesla's Reply to Autopilot Questions*, Ars Technica (2023), https://arstechnica.com/cars/2023/07/tesla-misses-deadline-to-inform-nhtsa-about-autopilot-problems/ (last visited Aug 25, 2023).

[63] *See e.g.* Joseph Herkert, Jason Borenstein & Keith Miller, *The Boeing 737 MAX: Lessons for Engineering Ethics*, Science and Engineering Ethics (2020) 26:2957–2974.

[64] Dominic Gates, *Boeing 737 MAX can return to the skies, FAA says*, The Seattle Times, Nov. 18, 2020, https://www.seattletimes.com/business/boeing-aerospace/boeing-737-max-can-return-to-the-skies-says-faa/ (last visited Aug 26, 2023).

[65] Douglas MacMillan, *'Our Daughter Died in Vain': What Boeing Learns from Plane*

typically holds the carrier or airline for airplane accidents because they are flight operators because they are in the best position to prevent aircraft accidents, not manufacturers. This consideration would not hold, however, if the disaster was caused by a technical malfunction of the aircraft, as this would be Boeing's responsibility.[66] (The pilot's error will rarely give rise to liability on the part of the pilot, unless they are found to have erred in stirring the planes. If the training was inadequate, then the party responsible for providing the training can be found liable).[67]

A subsequent investigation by the Indonesian authority, published in October 2019, placed some of the blame on the pilots and maintenance crews, the supplier of a component, but concluded Boeing and the Federal Aviation Administration were primarily responsible.[68] Data from the flight revealed that the Maneuvering Characteristics Augmentation System (MCAS) had forced the nose of the aircraft down 26 times in 10 minutes. [69] MCAS was an automated system unique to the Boeing 737 MAX, which had been added to the new 737 to fix for a potential nose-up stall that could appear under certain flight conditions which resulted from a change of the size and place of the engines in the new model.[70] Prior to the accident, MCAS had been acting on faulty sensor data that indicated an impending stall, which caused the plane to nose down and forced pilots to try to compensate.[71]

The Indonesian report noted that the design and certification of MCAS had not adequately considered the likelihood of loss of control of the aircraft. Boeing had also not included any information about MCAS in pilot training or manuals (and had also withheld it from the Federal Aviation

---

Crashes, Washington Post, Oct. 29, 2019, https://www.washingtonpost.com/business/2019/10/28/our-daughter-died-vain-what-boeing-learns-plane-crashes/ (last visited Aug 26, 2023).

[66] Airplane Crash Liability In International Law - Aviation - Worldwide, https://www.mondaq.com/aviation/903784/airplane-crash-liability-in-international-law (last visited Sep 3, 2023). *citing* Warsaw Convention of 1929, Montreal Convention of 1999,

[67] *Id.*

[68] Herkert et al. *supra* note 63 at 2959.

[69] Scott Neuman, *Indonesia Report: Pilots, Ground Crew Share Blame With Boeing For Lion Air Crash*, NPR, Oct. 25, 2019, https://www.npr.org/2019/10/25/773291951/pilots-ground-crew-share-blame-for-lion-air-737-max-crash-indonesian-report-says (last visited Aug 26, 2023).

[70] Some have argued that this in itself an ethical issue, as the decision to use software to "mask" the repositioning of the engines that disrupted the aerodynamics of the airframe can be questionable. *See* e.g. Herkert, et al. *supra* note 63 at 2960.

[71] Neuman, *supra* note 69.

Administration).[72] Both failures contributed to the pilot's inability to really understand what was happening.[73]

At the same time, similar faults had occurred on the previous flight of that same plane and the crew and maintenance staff had failed to report them, which would have resulted in grounding the plane according to existing protocols.[74] A crucial censor measuring the plane's angle of attack was out of calibration before the accident, and standard operating procedures required the crew to return and land the plane. But this didn't happen, nor was the issue reported. The report also noted that the crew had failed to coordinate their responses to multiple failure alerts and the first officer on the plane was unfamiliar with security procedures, had shown in training that he had problems handling the aircraft, and had failed to follow a checklist procedure that could have stopped MCAS from operating.[75] The fatal accident occurred because the pilots struggled to counteract repeated nose downs activated by MCAS due to the faulty sensor data.[76]

A few weeks after the crash in Indonesia, Boeing issued a new Flight Crew Operations Manual Bulletin, containing procedures for responding to flight control problems due to possible faulty sensor data.[77] According to the reports, this may have been the first time the airline pilots learned about the existence of MCAS.[78] In March 2019, however, the second crash occurred in Ethiopia 6 minutes after takeoff. After the crash, concerns were raised that the pilot of the plane had not received adequate training, and experienced observers commented that the accidents would have been prevented by more experienced pilots using customary stall prevention procedures, even with the lack of information on MCAS.[79] Subsequent reports by Ethiopian Airlines

---

[72] *Id.*

[73] Dominic Gates and Lewis Kamb, *Indonesia's devastating final report blames Boeing 737 MAX design, certification in Lion Air Crash*, The Seattle Time, Oct. 24, 2019, https://www.seattletimes.com/business/boeing-aerospace/indonesias-investigation-of-lion-air-737-max-crash-faults-boeing-design-and-faa-certification-as-well-as-airlines-maintenance-and-pilot-errors/ (last visited Aug 26, 2023)

[74] *Id.*

[75] *Id.*

[76] Neuman, *supra* note 69.

[77] Herkert et al. *supra* note 63 at 2958.

[78] Herkert et al. *supra* note 63 at 2959.

[79] Selam Gebrekidan, *Ethiopian Airlines Had a Max 8 Simulator, but Pilot on Doomed Flight Didn't Receive Training On It,* The New York Times, Mar. 20, 2019, https://www.nytimes.com/2019/03/20/world/africa/ethiopian-airlines-boeing.html (last visited Aug 26, 2023); Herkert et al. *supra* note 63, at 2961 citing William Langewiesche Langewiesche, *What really brought down the Boeing 737 MAX?,* The New York Times

Accident Investigation, and other experts, however, indicated that the pilots had followed the new checklist, which placed blame for the accident on design flaws in the plane.[80] It seems too that the pilots did not know how to disable the faulty software.[81]

In January 2023 there were at least 68 civil liability cases pending against Boeing. These, however, are expected to be settled as in 2021, Boeing accepted sole liability for the accident in Ethiopia, explicitly agreed that the pilots were not at fault and exonerated the MAX suppliers that had built the sensor and the one that produced, to Boeing's specification, the MCAS software.[82] The agreement includes a stipulation that was signed by most families. In exchange, Boeing got an explicit exclusion of any claim for punitive damages and put an end to legal discovery processes that could reveal further evidence of wrongdoing by the company.[83]

3. AI and Fundamental Rights: The Dutch Scandal

A significant amount of the literature on AI harms and risks is focused not on safety and security harms, however, but on harms to fundamental rights and, specifically, harms to the right to equality. Take, as an example of such harm, a 2019 scandal in the Netherlands over an algorithmic decision-making system used by the tax authorities to identify fraudsters in the national childcare benefit scheme. The system, which had been deployed since 2013, falsely accused tens of thousands of parents and caregivers, which resulted in dire financial consequences for many: victims had to wrongly repay large sums of money, lost access to benefits and had their properties seized. Additionally, several victims ended up affected by mental health issues, divorces, thousands of children taken into foster care and some even committed suicide.[84]

(September 18, 2019) https://www.nytimes.com/2019/09/18/magazine/boeing-737-max-crashes.html .

[80] Herkert et al. *supra* note 63 at 2959, 2961.

[81] Airplane Crash Liability In International Law - Aviation - Worldwide, https://www.mondaq.com/aviation/903784/airplane-crash-liability-in-international-law (last visited Sep 3, 2023).

[82] Dominic Gates, *Boeing accepts sole responsibility for 737 MAX crashes, wins agreement that avoids punitive damages,* The Seattle Times, Nov 10, 2021, https://www.seattletimes.com/business/boeing-aerospace/boeing-accepts-liability-for-737-max-accidents-wins-agreement-that-avoids-punitive-damages/ (last visited Aug 26, 2023)

[83] *Id.*

[84] Heikkilä, *supra* note 2.

In this case, it was hard to establish that illegal harm was occurring. Only around 2019 did it start becoming apparent that the system was classifying many people with dual nationality as potential fraudsters (often people of vulnerable communities, such as migrants).[85] A report by the Dutch Data Protection Authority explained that the self-learning algorithm was supposed to learn which claims had the highest risk of being false. The system, however, tended to brand as fraudulent applications with minor errors, such as missing signatures, and systematically identify high-risk claims by parents with double citizenship, most of which belonged to ethnic minorities.[86] An audit also revealed that the authority focused on non-western people, especially Turkish and Moroccan nationals.[87] People accused of fraud were forced to pay back thousands of euros given by the government for childcare, with no means of redress.[88]

Algorithmic opacity was an important factor that aggravated the situation: Applications with the highest risk scores were sent for manual review by a civil servant, but reviewers were not given information as to why the system had given the application a high-risk score for inaccuracy.[89] Victims also had little way to understand what was happening. According to a report by Amnesty International "(…) parents and caregivers were requested by civil servants to provide additional evidence to prove their entitlement to benefits. However, parents and caregivers who tried to find out what information was considered incorrect or false, or what evidence was deemed missing, were often met with silence; the tax authorities consistently refused to clarify their decisions."[90] At the same time, the Dutch authorities

---

[85] *Id. see also* Dutch DPA, *Methods used by Dutch Tax Administration unlawful and discriminatory,* July 17, 2020 https://autoriteitpersoonsgegevens.nl/en/current/methods-used-by-dutch-tax-administration-unlawful-and-discriminatory .

[86] Anna Holligan, *Dutch Rutte government resigns over child welfare fraud scandal,* BBC, Jan. 15, 2021, https://www.bbc.com/news/world-europe-55674146; https://www.tweedekamer.nl/nieuws/kamernieuws/eindverslag-onderzoek-kinderopvangtoeslag-overhandigd (last visited Aug 26, 2023)

[87] Heikkilä, *supra* note 2.

[88] *Id.*

[89] Amnesty International, Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal, Amnesty International (2021), https://www.amnesty.org/en/documents/eur35/4686/2021/en/ (last visited Aug 26, 2023); Dutch Data Protection Authority, *Belastingdienst/Toeslagen : De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag*, 15,https://www.autoriteitpersoonsgegevens.nl/uploads/imported/onderzoek_belastingdienst_kinderopvangtoeslag.pdf

[90] Amnesty International, *supra* note 89.

were secretive for a long time about the use of the AI system, and its working, and did not even make information available to the public about their use of algorithmic decision-making systems.[91]

A 2020 parliamentary report found several issues in the process that led to the scandal. These ranged from the Cabinet and Parliament enacting legislation "that was nail-biting and therefore lacked the ability to do justice to individual situations," the mass implementation of the program, with little consideration for particular situations, to an inadequate provision of information to the government regarding the extended and the hard-line approach to childcare scandal, and administrative courts that for years maintained the "nail-biting implementation of child care regulations."[92] Administrative courts apparently never noticed that the cases concerning child care benefits largely concerned people with dual nationality.[93] After the report, about 15'000 people were compensated with E 30'000 in compensation.[94]

Later in 2021, the Dutch Data Protection Authority also fined the Dutch tax administration with 2.75 million Euros for the unlawful, discriminatory and improper retention and use of nationality data.[95]

c. C. AI-Human interactions

In the above-presented incidents of AI harms and accidents, there is always a human directly interacting with an AI system: the Tesla driver, the airplane pilot, and civil servants. In some instances, human negligence and bias may have contributed to the harm. And yet, as the examples above also show, the characteristics of the AI system, their governance and institutional setting, the lack of training and access to relevant information and by the

---

[91] Amnesty International, *supra* note 89.

[92] Tweede Kamer der Staten Generaal, Eindverslag onderzoek kinderopvangtoeslag overhandigd (Dec. 17, 2020) https://www.tweedekamer.nl/nieuws/kamernieuws/eindverslag-onderzoek-kinderopvangtoeslag-overhandigd.

[93] *Highest Dutch court apologises to childcare benefit scandal victims,* Dutch News (nov. 19, 2021),

[94] *Id.*

[95] Björn ten Seldam & Alex Brenninkmeijer, *The Dutch benefits scandal: a cautionary tale for algorithmic enforcement*, April 30, 2021 https://eulawenforcement.com/?p=7941; Dutch DPA, *Methods used by Dutch Tax Administration unlawful and discriminatory,* July 17, 2020 https://autoriteitpersoonsgegevens.nl/en/current/methods-used-by-dutch-tax-administration-unlawful-and-discriminatory .

people interacting with the algorithm, and the interaction interface contributed to making these individuals not effective at preventing harm.[96] How should regulation, and specifically liability law, account for the complexity of AI-Human interactions?

This subsection starts addressing this question by preceding the role humans are supposed to have when they are supervising and interacting with AI systems and recent findings on the effectiveness of these interactions and hybrid systems.

1. What is a human-in-the-loop and its key assumptions

AI regulations, best practices and ethical documents often recommend inserting a "human in the loop." In early 2020 Fjield et al. found in an analysis of 36 AI ethics documents that about 70% of these mentioned human control of technology as a guiding principle that requires that important decisions remain subject to human review.[97] In 2021, Ben Green surveyed 41 policy documents from around the world that provide some form of mandate or guidance regarding human oversight of algorithms in the public sector, including the AI Act.[98]

Rebecca Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, Humans in the Loop, (2022), https://papers.ssrn.com/abstract=4066781 (last visited Aug 29, 2023). and coauthors have explained that the ambitions of such policies are varied: humans are expected to play a corrective role as they improve the performance of a system by correcting for errors, situations and bias; a resilience role, as they may be expected to act as a failure mode or stop the system from working in case of emergency; a justificatory role to increase a system's legitimacy and guarantee the principle, especially in government, that decisions should respond to the circumstances of individual people; relatedly a dignitary role as they may be expected to protect the dignity of people affected by the decision; insert friction, and play an accountability role as a means to ensure that someone is legally liable and/or morally responsible for the system's decisions – as in cases when Tesla's

[96] To be fair, many of today's daily interactions are machine-human interactions that assist us in all sorts of decision-making, from how we use Google Maps to, most recently, ChatGPT. Many of these interactions are however harmless.

[97] Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, (2020), https://papers.ssrn.com/abstract=3518482 (last visited Aug 27, 2023).

[98] Green *supra* note 11.

hand off control to human drivers before a crash.[99] In some instances, the mere fact that a human is involved is raised as a normative and ethical value.[100]

One of the key assumptions that underlie these mandates is that hybrid systems can bring the best of humans and machines to decision-making systems: Humans are flexible decision-makers which allows us to exercise discretion, generalize and jump across context, and we can justify our decisions, even if our decision-making processes are also opaque.[101] Algorithms are fast, capable of making decisions based on far more information and factors than humans, consistent, and at scale.[102] They are, however, bad at ethics or following norms, do not justify their decisions, and are especially dependent on their training data and the data fed into models, which makes them prone to reproduce the biases in them. They are thus bad at edge cases.[103] Hybrid systems thus promise to bring the best of both worlds by allocating tasks to either an individual or a machine, based on lists of what each is supposed to be better at. [104] This is known as the Men-Are-Better-At/Men-Are-Better-At method (MABA-MABA), which was developed in 1951 when the National Research Council attempted to characterize human-machines interactions before developing a national air traffic control system.[105]

---

[99] Crootof, et al. *supra* note 12, at 474-487. (citing at 482 Bettina Berendt & Sören Preibusch, *Towards Accountable Discrimination- Aware Data Mining: The Importance of Keeping the Human in the Loop – and Under the Looking Glass,* 5 Big Data 135 (2017). *See also* Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, (2022), https://papers.ssrn.com/abstract=3921216 (last visited Aug 27, 2023).

[100] Even if important, I do not engage here in the scenarios in which humans are kept in control or in the loop as a means to avoid the risk of replacement.

[101] Crootof et. al, *supra* note 6, at 462, citing Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, (2019), https://papers.ssrn.com/abstract=3329669 (last visited Aug 27, 2023)

[102] *Id.*, at 464 *see also* recent research that are defining new types of interactions between humans and machine learning algorithms at the learning process. Eduardo Mosqueira-Rey et al., *Human-in-the-Loop Machine Learning: A State of the Art*, 56 Artif Intell Rev 3005 (2023), https://doi.org/10.1007/s10462-022-10246-w (last visited Aug 27, 2023).

[103] *Id.* at 465.

[104] This is known as MABA-MABA: 'Men Are Better At Machines Are Better At' approaches that propose that designers divide the tasks between humans and machines by considering what people or machines are supposedly better at. *See* S. W. A. Dekker & D. D. Woods, *MABA-MABA or Abracadabra? Progress on Human–Automation Co-Ordination,* 4 Cognition Tech Work 240 (2002), https://doi.org/10.1007/s101110200022 (last visited Aug 27, 2023).

[105] Dekker & Woods *supra* note 104.

In this vein, the EU's Artificial Intelligence Act proposals, for example, impose an obligation for developers and deployers of high-risk systems to design and develop them so that they can be effectively overseen by a natural person while in use. In the Act, human oversight must be aimed at preventing or minimizing the safety and fundamental rights risks raised by the AI system.[106]

2. The challenges of effective human-machine hybrid systems

The optimism about the promises of using AI for decision-making and human-machine interaction is tempered by evidence, especially in the context of algorithmic decision-making, that humans provided with machine recommendations do not perform as well as machine-predictions alone, but also may lead to patterns that increase disparities in decision outcomes across socioeconomic and racial groups.[107] Indeed, hybrid systems do not necessarily bring the best of both worlds: hybrid systems have dynamics of their own and it is difficult to design effective hybrid systems that require collaboration between humans and automated technologies.[108]

Research on human-machine interactions has shown that this occurs for a variety of reasons:

First, the assumption that people and computers have fixed strengths and weaknesses that can be easily capitalized on or used to compensate for the other party's weaknesses is not accurate.[109] Certainly, machines are better at several things: it is difficult for humans to sustain focused attention for more than 20-30 minutes, which is what is needed to, for example, fly a plane. This is true in many other domains reduced to motor memory.[110] In hybrid systems, however, the level of functions to be considered for function

---

[106] Art. 14; *See infra* Part III A

[107] Maria De-Arteaga, Riccardo Fogliato & Alexandra Chouldechova, *A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores*, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems 1 (2020), http://arxiv.org/abs/2002.08035 (last visited Aug 27, 2023).

[108] Green *supra* note 11 at 2 citing Lisanne Bainbridge, *Ironies of automation,* Automatica (1983).

[109] Dekker & Woods *supra* note 104.

[110] Mary Missy Cummings, *Man versus Machine or Man + Machine?*, 29 IEEE Intell. Syst. 62 (2014), https://ieeexplore.ieee.org/document/6949509/ (last visited Aug 27, 2023).

allocation may not be non-problematic or self-evident. Dekker and Woods argued in 2002 that capitalizing on some strengths and weaknesses creates new human strengths and weaknesses: For example, when automation can perform complex and repetitive tasks for an extended period, it increases the difficulty for humans to remain attentive and vigilant *to* the system. In other words, as automation takes over tasks that used to require constant human monitoring, people may become less focused or attentive to the overall process. This can lead to a potential problem known as "vigilance decrement."[111] Thus, because human-machine interactions create new dynamics it is a priori not obvious how to capitalize on different strengths. At the same time, much of the MABA-MABA literature do not explain "the cognitive work that might be involved in deciding how and when to intervene or how to switch from level to level [and] (…) leaves unspecified how humans should decide when and whether to intervene or when to back off."[112]

Second, in instances of human interaction with the algorithmic recommendation, there are two competing tendencies that have been observed: automation bias and algorithmic aversion.[113] Algorithmic aversion is the tendency to ignore tool recommendations after seeing that they can be erroneous. It originates from a lack of agency and lack of transparency, and studies have shown that users will prefer to sacrifice accuracy for control over the algorithm's output.[114] Indeeds, reports have shown that humans can override machine predictions even when these are highly reliable. Studies have also shown that trust in systems is related to how accurate the system is perceived to be.[115]

---

[111]Dekker & Woods *supra* note 104.

[112] *Id.*

[113]Maria De-Arteaga, Riccardo Fogliato & Alexandra Chouldechova, *A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores*, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems 1 (2020), http://arxiv.org/abs/2002.08035 (last visited Aug 27, 2023).

[114] *Id.*- Recent human-machine interaction approaches incorporate Levels of Automation scale (LOAs) to allocate roles between automation and humans in complex systems. LOAs range from fully manual to fully automated systems. An example by Ge Wang illustrates this with an AI tool translating legal documents, offering a slider to adjust the level of jargon, enabling user customization and adaptability. - Ge Watt, *Humans in the Loop: The design of Interactive AI Systems,* Stanford University Human-Centered Artificial intelligence, Oct. 20, 2019

[115]Kun Yu et al., *Trust and Reliance Based on System Accuracy: 24th International Conference on User Modeling, Adaptation, and Personalization*, UMAP 2016, UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization 223 (2016), http://www.scopus.com/inward/record.url?scp=84984881638&partnerID=8YFLogxK (last visited Aug 29, 2023).

Automation bias, on the other hand, refers to individuals' tendency to defer to automated systems.[116] This can lead to omission errors, instances where the human does not detect problematic cases, or fails to act even if they do; a famous example are pilots who tend to rely blindly on automated cues and don't remain vigilant.[117] Commission errors are instances where humans take action based on an erroneous recommendation, failing to incorporate additional external information.[118] Studies have shown that time pressure, facing complex tasks, and the user's self-confidence on their decisions tend to contribute to automation bias.[119]

Drawing from empirical research in algorithmic decision-making in the government context, Ben Green has argued that an underlying problem with algorithmic oversight algorithmic oversight measures have an underlying "mismatch of skills and responsibilities: (...) Human oversight therefore means asking people to perform quality control for systems that perform at a higher prediction quality than people do, and often in inscrutable ways."[120] Referring to the specific instance of algorithmic decision-making systems adopted in government Green has also argued that human oversight policies are unable to provide reliable oversight because by not taking those difficulties into account, they are often designed to simply be "a rubber stamp," they, for example, often lack clear measures of success. Thus humans have no real incentives or means to change an AI system's decision.[121]

3. Towards effective humans in the loop and what this means for liability

A vast field of research has sought to identify ways to overcome some of the challenges of effective human-in-the-loop systems (much of it developed in the past 30 years for aviation and surface transportation settings). Recently an important part of that conversation has also focused on the context of complex and high-risk AI systems.

---

[116] Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, (2022), https://papers.ssrn.com/abstract=3921216 (last visited Aug 29, 2023). (citing Parasuraman & Manzey, 2010; Kitka et al. 1999)

[117] De-Arteaga et al., *supra* note 113 at 2 citing K. L. Mosier et al., *Automation Bias: Decision Making and Performance in High-Tech Cockpits*, 8 Int J Aviat Psychol 47 (1997).

[118] De-Arteaga et al., *supra* note 113

[119] *Id.*

[120] Green *supra* note 11 (citing Parasuraman & Manzey, 2010; Kitka et al. 1999).

[121] *Id.*

Dekker and Woods point out, for example, that allocating a particular function to machines also creates new functions for humans – and a transformation of the workplace in their examples -, which must be accounted for in training and interaction. These include typing, searching for the right display, or making decisions about prioritizing health care with the help of a software.[122] Thus, for example, pilots invest considerable time learning procedures to understand how to interact with the system. For instance, when they encounter a warning light, they understand the need to refer to a manual to identify the appropriate procedure, as it is not feasible to memorize a vast number of procedures. In some instances, what procedure to apply also requires interpretation if, for example, a single sound alert corresponds to different failure scenarios, making it less evident which specific procedure to follow.[123]

They also emphasize the need to depart from the quantitative and substitutional practice of function allocation and move towards supporting better coordination between people and automation.[124] Recent scholarship points in similar directions.[125] This requires, for example, making the operations of automated systems observable to humans, and making it easy and efficient for human operators to direct the system, especially in novel episodes.[126] Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, Humans in the Loop, (2022), https://papers.ssrn.com/abstract=4066781 (last visited Aug 29, 2023). et al. similarly point out the importance of designing human-in-the-loop systems that "promote effective interaction between the human and the system," and with careful attention to the different levels of expertise, experience and training of the individuals interacting with these systems.[127]

Recent work has explored approaches to characterizing human-machine complementarity. Maria de Arteaga and co-authors identify, for example, that supervisors of an algorithmic system in child-welfare in the US were able to

---

[122] Dekker & Woods *supra* note 104.

[123] Cummings, M.L., *Adaptation of Licensing Examinations to the Certification of Autonomous Systems,* Safe, Autonomous and Intelligent Vehicles, Eds: Li,X.,Murray, R., Tomlin, C.J., Yu, H., Unmanned System Technologies series, Springer, in press.,

[124] Dekker & Woods *supra* note 104.

[125] *See* Dale Richards et al., *Designing for Human-Machine Teams: A Methodological Enquiry,* 2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS), Orlando, FL, USA, 2022, pp. 1-4, doi: 10.1109/ICHMS56717.2022.9980612. (Check)

[126] Dekker & Woods *supra* note 104.

[127] Crootof et. al, *supra* note 6, at 498.

correct for a glitch in the system because they had access to the underlying administrative data. This provided an alternative view of the case than what was being shown into the risk score calculation.[128] Other researchers have examined the role of providing explanations of how algorithmic decisions are being made.[129] Other studies have shown that inserting social accountability is a way to decrease automation bias.[130]

From a governance perspective, Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, Humans in the Loop, (2022), https://papers.ssrn.com/abstract=4066781 (last visited Aug 29, 2023). et al. have drawn from the experience of successful regulation of human-machine systems in safety-critical systems, to emphasize that hybrid human-AI systems require detailed rules for system designers and operators. This is the case of railroads, nuclear reactors, and medical devices.[131] Depending on the context, humans and algorithms bring different things why incorporating one, or the other, to a decision-making process should be clarified.[132] Similarly, and also drawing from the experience of critical instances Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, Humans in the Loop, (2022), https://papers.ssrn.com/abstract=4066781 (last visited Aug 29, 2023). et al. point out that regulatory agencies should require that product designers create technological systems around the people operating the system, that the devices are designed and labeled sufficiently for effective use, and address training and organizational policies.[133]

4. II. LIABILITY FOR AI: MAPPING THE DEBATE

The previous Part outlined how AI's characteristics and the characteristics of AI-hybrid systems complicate identifying who is to blame when a mishap occurs. This has led different countries and actors, but

---

[128] De-Arteaga et al., *supra* note 113.

[129] *Id.* at 4

[130] L. J. Skitka et al., *Automation Bias and Errors: Are Crews Better than Individuals?*, 10 Int J Aviat Psychol 85 (2000).

[131] Crootof et. al, *supra* note 6, at 494-496.

[132] Crootof et. al, *supra* note 6, 490-491.

[133] Crootof et. al, *supra* note 6, at 466; Green also emphasizes the importance of strengthening institutional oversight of algorithms and requiring justifications as to why it is appropriate to incorporate an algorithm into decision-making and to provide evidence that the algorithmic system can be effectively overseen, for example after lab tests with lay people. Green *supra* note 11, at 14.

particularly the EU, to advance novel regulatory frameworks for AI and to start considering whether and how to update their civil liability framework. A particular challenge of this effort is that as AI accountability rules and policies introduce humans-in-the-loop, the blame for accidents are harm may be assigned to them - because their role is to supervise systems - but the research also shows that human-machine interactions and possibility for actual supervision is more complex than that.

How, however, does present liability law in the EU deal with the specific challenges posed by AI systems, and human-AI hybrid systems?

An Expert Group on Liability and New Technologies convened by the European Commission concluded in 2019 that existing liability rules offer baseline protection. The experts, however, also identified that the characteristics of AI systems and their applications may make it more difficult to offer victims a claim for compensation in all cases where it seems justified and that the allocation of liability can be unfair or inefficient.[134]

This Part lays out these challenges by situating the policy issues identified in the previous Part within the two main liability regimes in the EU: fault-based liability, and strict liability (which includes product liability).

Before we dive into this, however, a methodological note on this section. Liability law in the European Union is largely not harmonized, except for product liability law,[135] some aspects of liability for infringing data protection law,[136] and liability for infringing competition law.[137] There are similarities, however, in the general regimes, and efforts to understand EU-liability law in a systematic way. What follows draws mostly from the work of the Expert Group and the European Tort Law Group and their developed Principles of European Tort Law.

A. Fault-Based Liability and the Challenges posed by AI

---

[134] Expert Group on Liability and New Technologies *supra* note 16, at 1.

[135] Council Directive 85/375/EC on liability for defective products, OJ L 210, 7.8.1985.

[136] Regulation (EU) 2016/ 679 General Data Protection Regulation, OJ L 119, 4.5.2016 , Art. 2 [*hereinafter* GDPR]

[137] Directive 2014/104/EU of the European Parliament and of the Council of 26 November 2014 on certain rules governing actions for damages under national law for infringements of the competition law provisions of the Member States and of the European Union Text with EEA relevance, 349 OJ L (2014), http://data.europa.eu/eli/dir/2014/104/oj/eng (last visited Aug 29, 2023).

Liability law is concerned with the responsibility and redress for harms to legally protected interests.[138] Legally protected interests are affectations to fundamental rights (such as discrimination, or manipulation) and physical harms (such as damages to property, personal injury and death). [139] To do so, liability law is generally concerned with identifying who is the person or entity to whom that damage is legally attributable to compensate for that damage.[140] What constitutes "legally attributed" depends on the type of liability: In domestic regimes, the general rule for liability attribution is fault-based, which requires that the injurer's objectionable and avoidable conduct - fault - caused the damage.[141] The objective is both to provide corrective justice and provide the right incentives to avoid harm, which economic analysts of liability law call "cheapest cost avoidance."[142]

This subsection first explains, very briefly, the basics of fault-based liability and then develops the challenges AI characteristics pose to it.

1. Fault-Based Liability 101

Fault-based liability requires proving (1) the existence of damage, (2) the fault of the injurer, which is usually intent or negligence, and (3) a causal link between the damage and the faulty conduct.[143]

A damage requires material or immaterial harm to a legally protected interest and their protection varies on its nature and value.[144] Typically, life, bodily or mental integrity, human dignity, and liberty enjoy the highest protection.[145] Property and property rights enjoy extensive protection,[146] whereas the protection of pure interests in contractual relationships is more limited in scope.[147]

What constitutes fault in each case depends on the standard of conduct

---

[138] PETS *supra* note 6, Art. 2:102 (noting, also that pure economic interests or contractual relationships may be more limited in scope).

[139] Wendehorst *supra* note 4 at 189 (Some, however, may not. For example risks are related to how AI affects how we conceptualize the world.)

[140] PETS *supra* note 6, Art. 1:101.

[141] PETS *supra* note 6, Art. 1:101(2) a).

[142] *See* Buiten et al. supra note 4..

[143] PETS *supra* note 6, Art. 101:101

[144] PETS *supra* note 6, Art. 2:101 (a)

[145]PETS *supra* note 6, Art 2: 101(b)

[146] PETS *supra* note 6,Art 2: 101(c)

[147] PETS *supra* note 6, Art 2: 101(d)

expected in different situations. The standard of conduct varies depending on the nature of the interest involved, the dangers of the activity, the foreseeability of the damage, and the expertise of the person carrying it, amongst others and with several differences under Member State law.[148] In the case of negligence - the failure to take reasonable care to prevent harm - the different jurisdictions usually require that the injurer's negligence was objectionable, for example, because it violated the law.[149] Liability can also be triggered simply by not complying with particular laws, such as under the General Data Protection Regulation (GDPR), where Article 82 attaches liability to any infringement of the requirements set out by the regulation.

Lastly, the causal link is the requirement that the faulty action must be a conduct or omission that, had not taken place, the damage would not have occurred (*Condition sine qua non*).[150] If there are multiple causes, where each alone would have caused the damage, each conduct is regarded as a cause of the damage to the extent corresponding to the likelihood that it may have caused the damage.[151] If it remains uncertain which of the multiple causes was the one that caused the damage, each conduct is regarded as a cause "to the extent to the likelihood that it may have caused the victim's damage."[152] If it is impossible to estimate the likelihood, it is assumed that all causes contributed in equal shares.[153] If the victim caused the damage to a certain extent, the victim may bear its loss to the extent corresponding to their action (if the victim caused the damage to a full extent, no one else will be held liable).[154]

2. Challenges posed by AI systems

This subsection explains how AI's characteristics complicate proving the existence of harm, establishing fault, attributing causation and, in complex cases, dividing responsibility among different parties.[155]

---

[148] PETS *supra* note 6, Art. 4:102.

[149] *See* Wendehorst, *supra* note 4, at 192.

[150] PETS *supra* note 6, Art 3: 101

[151] PETS *supra* note 6, Art 3: 102

[152] PETS *supra* note 6, Art. 3(103)

[153] PETS *supra* note 6, Art. 3: 105

[154] PETS *supra* note 6, Art 3: 106.

[155] See Buiten et al. *supra* note 4, identifying three gaps in the existing liability regime with respect to AI: establishing fault, proving causality and dividing responsibility among different parties, subsuming fault and attribution as the same one.

a. Damage

First, is the fact that it may not be obvious that harm to a protected interest is occurring or has occurred. This is specially the case to some harms to fundamental rights, and some economic harms. In the Dutch scandal, for example, the algorithmic decision-making system used by the tax authorities falsely accused tens of thousands of parents and caregivers.[156] Yet, only in 2019, did it become apparent that the system was biased, while the system had been in place since 2013, even if victims maybe had a sense that something wrong was going on. At first sight, there is nothing extraordinary about some people getting subsidies readjusted, or different sorts of applications denied – from admissions to colleges to jobs and loans.

This can be a function of operational and organizational opacity. In the Dutch scandal,it was difficult – if not impossible – for injured parties but also the system's operators to understand what was happening:[157] Reviewers were not given information as to why the system had given the application a high-risk score for inaccuracy.[158] Victims also reported difficulty in finding out what information was considered incorrect or false, or what evidence was deemed missing. [159]

b. Fault

Second, once the harm to a protected interest is established – or identified – victims must prove that someone's faulty behavior caused it. This is also difficult on at least three interrelated accounts: First, the AI system's opacity complicates identifying what went wrong. Second, AI's complexity makes identifying who was at fault hard. Third, a lack of behavioral standards makes it difficult to establish what is the standard of care that different parties must follow.[160]

Take, as an example, the case of the Tesla car that collided against the street curb and where the security driver was injured. In such an instance, the harm could be attributed to various factors: poor car design, incorrect or misinterpreted data, flawed software updates, or user negligence.[161] In similar

---

[156] *See supra* Part I.B.

[157] Expert Group on Liability and New Technologies *supra* note 16, at 26.

[158] Amnesty International, *supra* note 89, at 15.

[159] Amnesty International, *supra* note 89, at 12

[160] *See* Buiten et al. *supra* note 4, at 7

[161] Expert Group on Liability and New Technologies *supra* note 16.

instances, the car may have been fed data from other automated vehicles around it and from road operators. A partner technology company may have developed the self-driving software. A third party may have contributed to the GPS mapping system, and so on.[162]

Under a traditional fault-based liability the victim would have to prove that someone breached a duty of care. This is not only complicated given the system's complexity, but it would also imply showing what the appropriate standard of care is.[163] This would require showing, for example, how others in the industry or field would have acted in similar circumstances and proving that the defendant's actions fell short of this expected standard, something that is hard to do given, in general, the opacity of the AI industry.[164]

In the case of human-AI hybrid systems, the lack of clarity of how a particular system is supposed to improve human decision-making, and vice versa, creates additional difficulties in establishing to what extent the human in the loop contributes to harm.[165] Recall that in the Tesla accident, the jury concluded that the victim's distraction had been the cause of the accident, but the research shows it is not uncommon for humans to rely on AI systems and clarity on roles and design interfaces play an important role.[166] The Expert Group has highlighted that assigning liability when a victim's actions contribute to their own harm is not a new challenge. However, the difficulties stemming from opacity, complexity and the absence of established standards also apply to cases where the victim's actions play a role. This includes the challenge of determining what should be considered the expected level of care from the human or victim in such situations.[167]

c. Causality

---

[162] *See* Water Street Partners, *Autonomous Vehicle Partnerships: How Tech Companies and Automakers are Collaborating to Innovate the Future*, Medium Blogpost, Nov. 8 2016, https://medium.com/@water.street/autonomous-vehicle-partnerships-how-tech-companies-and-automakers-are-collaborating-to-innovate-cf44bc9e85a .

[163] *See* Expert Group on Liability and New Technologies *supra* note 16, at 20; *see* Part III A *infra* presenting the AI Act.

[164] Expert Group on Liability and New Technologies *supra* note 16, at 26; Buiten et al. supra note 4. (arguing that, in the case of autonomous AI systems, this is aggravated by the fact that some outputs can't be anticipated. This challenge may be mitigated however, upon interacting with other risk-mitigating regulations where AI systems are specifically trained to avoid certain harmful outputs); *see also* ChatGPT technical document where GPT is trained to moderate content, with certain success.

[165] *See* supra Part I.B.

[166] *See supra* Part I.C.

[167] Expert Group on Liability and New Technologies *supra* note 16, 31.

Third, the victim would have to prove the cause-and-effect relationship between the defendant's actions or omissions and the resulting harm. Given AI's technical and organizational opacity, doing things such as identifying how a bug in intricate software code, or the process behind an AI system's decision-making leads to a specific outcome, or gathering relevant evidence is more difficult, time-consuming, and expensive.[168] As explained before, AI system's complexity and autonomy may also make it difficult, if not impossible, to trace back outcomes to a specific actor's decision.[169]

Professor Christiane Wendehorst has argued that the difficulty in proving causality may lead to situations where there is an accountability gap because the victim won't be able to prove whose actions caused an accident.[170] In other instances, all parties may be found jointly liable. However, according to Buiten et al., this could also be undesirable for companies that didn't cause the harm, especially those that acquire AI systems from third-party providers and have limited technical capacity to prove harm for redress.[171] Buiten et al. note that this accountability gap can lead to situations where owners or controllers of AI systems provide suboptimal levels of precaution. [172]

B. Strict Liability, Product Liability and AI

The advent of AI systems isn't, of course, the first-time liability regimes must be adapted to new technologies or important information asymmetries between victims and tortfeasors.[173] Courts and regulators have historically altered the burden of proof of some of the above elements in some cases where it may be difficult for the victim to prove each of the conditions of a

---

[168] Expert Group on Liability and New Technologies *supra* note 16, at 26.

[169] Buiten et al. *supra* note 4., at 8.

[170] *See e.g.*, Bénédict Winiger et al. (eds), Digest of European Tort Law I: Essential Cases on Natural Causation (2007),

[171] The Expert Group also adds that "The problem of who really caused the harm in question will therefore often not be solved in the first round of litigation initiated by the victim, but on a recourse level, if ever. More modern approaches provide for proportional liability at least in some cases, reducing the victim's claim against each potential tortfeasor to a quota corresponding to the likelihood that each of them in fact caused the harm in question." Expert Group on Liability and New Technologies *supra* note 16 , at 24.

[172] Buiten et al. *supra* note 4., at 9.

[173] Expert Group on Liability and New Technologies *supra* note 16 , at 27; Miquel Martín-Casals, *Technological Change and the Development of Liability for Fault: A General Introduction,* The Development of Liability in Relation to Technological Change (Miquel Martín-casals, et al. eds. 2010).

fault–liability regime. This is done considering the gravity or dangerousness of the activity, the seriousness of the possible damage, the likelihood that such damage may occur and how easy – or hard – it may be for the victim to access evidence.[174] In medical malpractice cases, for example, the burden of producing evidence tends to be on the party who is or should be in control of the evidence.[175] The effect of doing so is facilitating access to compensation and reduces the information asymmetry between the victim and the injurer.[176]

Particularly from the 19th century onwards, legislators often responded to risks brought about by new technologies - like trains and motor vehicles - by introducing strict liability, a liability regime that does not require the injurer's conduct to have been faulty but merely that their conduct caused harm. This subsection briefly explains the main requirements of strict liability, how it is enshrined in the EU's product liability regime, and the challenges and opportunities of introducing it to AI systems and, specifically, human-AI hybrid systems. It finishes by laying out the discussion in the EU of changing the burden of proofs or extending a strict liability regime to AI systems.

1. Strict liability

Strict liability only requires proving the link of causation between harm and an activity or conduct without which the damage would have not occurred, and which effectively caused the harm.[177] It does not require proving fault. Such conduct can be, for example, driving. If, when driving, someone accidentally damages someone's property, under a strict liability regime that person will be held liable, even if there was no mal conduct at issue. The main defence against strict liability is *force majeure*, which requires that the damage was caused by an "unforeseeable and irresistible event," or conduct by a third party.[178]

The rationale behind strict liability is that the person who carries out and benefits from activities that are considered "abnormally dangerous," which means they create a foreseeable and significant risk of damage even when due care is exercised, or the damage is likely and/or serious,[179] should carry

[174] PETS *supra* note 6, Art. 4:201
[175] Expert Group on Liability and New Technologies *supra* note 16, at 23.
[176] Buiten et al. *supra* note 4., at 4.
[177] PETS *supra* note 6, Art. 3:101.
[178] PETS *supra* note 6, Art. 7:102(a).
[179] PETS *supra* note 6, Art. 5:101.

the loss inflicted on third parties when an accident occurs.[180] It is also often imposed in instances where it would be extremely difficult for the victim to prove a defect or a fault. [181]

A special form of strict liability is vicarious liability, in which a person pursuing a lasting enterprise for economic or professional purposes is held strictly liable for the harm caused by their auxiliaries provided that the auxiliary violated the required standard of conduct.[182] There are also examples in national law where something close to strict liability is extended to the owners of things that cause harm, but this is rather exceptional and often narrowed down by case law.[183] Strict liability for the operation of computers, software, or similar technologies remains largely unfamiliar in Europe.[184]

2. Product liability

EU law establishes the principle of strict liability for producers of material products - such as home appliances or smartphones - as an important part of the European consumer protection law. Product liability was harmonized at the Union level by the Product Liability Directive (PLD) in 1985 and is based on the principle that "the producer" of a product – understood as tangible movables[185] - is liable for damages to life, health and property caused by a defect in a product they have put into the market as part of their business regardless of whether the defect is their fault.[186] It does not extend to harm to fundamental rights.[187]

---

[180] Buiten et al. *supra* note 4. at 5.

[181] It must be noted that strict liabilities are often accompanied by liability caps or other limitations to counterbalance the increased liability risks faced by technology beneficiaries. The chilling effect of tort law is arguably even stronger when liability remains entirely unresolved and unpredictable. In contrast, the introduction of specific statutory solutions at least partially defines and delimits the risks involved, thereby contributing to their insurability. These caps are frequently justified as facilitating insurability, as strict liability statutes often require adequate insurance coverage for such liability risks. Expert Group on Liability and New Technologies *supra* note 16 at 28.

[182]PETS *supra* note 6, Art. 4:202.

[183] Responsabilité du fait des choses, Article 1242 French Civil Code; Wendehorst, *supra* note 4, at 193.

[184] Expert Group on Liability and New Technologies *supra* note 16, at 28 citing See §§ 89e, 91b paragraph 8 of the Austrian Gerichtsorganisationsgesetz (Court Organisation Act).

[185] Product Liability Directive, Recital 3, Art. 1 - , and has been interpreted by the ECJ as also applying to products used while providing any service. (find case)

[186] Product Liability Directive, Art. 4 and 7..

[187] Wendehorst, *supra* note 4 at 197.

The PLD defines a defect as a deviation from the standards that are reasonably expected from that enterprise, product, or service.[188] The PLD requires victims to prove that the product was defective at the time when it was placed in the market and left the producer's sphere of control, that a damage was suffered, and that the defect caused the harm. Since it can be difficult for consumers to do this for any technically complex product, national courts have developed ways to facilitate the burden of proof by, for example, including disclosure obligations for the producer.[189] Developers can defend themselves by proving that the state of the art in science and technology could not have detected the defect when the product was put in circulation.[190]

European authorities and the Expert Group for AI Liability have identified that the PLD regime is not fit to meet the risks of emergent technologies like AI: the PLD was designed with traditional markets and business models in mind; material objects, placed on the market by a one-time action (selling), after which the producer does not have any control over the product.[191] The complexity of AI systems challenges the "notion of a "product" and a "defect."

First, is the definition of a product. The 1985 PLD covers tangible products, which means that the hardware components of an AI system, and software integrated into tangible AI systems are covered by the PLD, but it is unclear whether software alone is covered.[192] Similarly, where the producer does not control the software or other technical features are later installed or learned by the are most likely not covered by the Directive. This leaves out applications and services that are downloaded from the Internet, but which

---

[188] PETs, *supra* note 4, Art. 4:202(2)

[189] *See* Daily Wuyts, *The product liability directive–more than two decades of defective products in Europe* 5 Journal of European Tort Law (2014).

[190] Wendehorst, *supra* note 4, at 193.

[191] Expert Group on Liability and New Technologies *supra* note 16 , at 28; (Fifth) Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the Application of the Council Directive on the approximation of the laws, regulations, and administrative provisions of the Member States concerning liability for defective products (85/374/EEC), COM(2018) 246 final, 8 f.

[192] *See* Buiten et al. supra note 4. at 15, (citing Gerhard Wagner, 'Robot Liability' in Sebastian Lohsse, Reiner Schulze, and Dirk Staudenmayer (eds), Liability for Artificial Intelligence and the Internet of Things (Nomos 2019), at 604. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018SC0157)

may not be so different from counterparts that may have been previously provided on CDs. [193]

The second challenge comes from the notion of a defect. Under the PLD, defectiveness is determined based on the safety expectations of the average consumer, but so long as the defects could have been known at the time the product was placed on the market. Deviations in the decision-making of AI autonomous and generative systems, however, raise questions about whether all harms caused by AI should be treated as a defect or whether it should be affected that a well-functioning AI system could cause harm.[194]

Third, the Expert Group found that meeting the standard of the burden of proof – proving that a product is defective - is challenging for injured persons in complex cases.[195]

Lastly, as the PLD focuses on the moment when a product was put into circulation as the moment that defines the producer's liability, this cuts off claims over subsequent additions – by the producer or someone else – over updates or upgrades or a system. It also does not account for software updates which are in fact often meant to make products safer, or the fact that AI systems are supposed to continue learning once they are placed in the market. Consequently, the PLD does not provide duties to monitor products after circulation. The fact that neither the producer nor the user may have full control over an AI system's operation – because it requires data provided by third parties or is collected from the environment and depends on self-learning processes – dilutes the role of the producer into a multitude of actors that contribute to the design and functioning of an AI system.[196]

3. Should AI systems be covered by strict liability?

The primary advantage of strict liability for harms caused by an AI system is evident: it would relieve them from the burden of proving any wrongdoing by the defendant or establishing a causal connection between such wrongdoing and their loss and they can focus solely on whether the technology's risk materialized and caused harm. It would also provide a

---

[193] Expert Group *supra* note 19.

[194] Expert Group *supra* note 19, at 28.

[195] *See* European Commission, New PLD Explanatory memorandum at 2 https://single-market-economy.ec.europa.eu/system/files/2022-09/COM_2022_495_1_EN_ACT_part1_v6.pdf [*hereinafter* Explanatory Memorandum New PLD] .

[196] Expert Group on Liability and New Technologies *supra* note 16 , at 30.

functional equivalent to users, giving them the same protection when the driver is a human or an AV.[197] Christiane Wendehorst has thus recommended, for example, that a harmonized regime of vicarious liability be adopted so that "a principal that employs AI for a sophisticated task faces the same liability under existing Member State law as a principal that employs a human auxiliary."[198] This would address the difficulty victims have in proving fault. Legislators and courts would not need to have information on the optimal level of precaution in designing and deploying AI-based systems.[199]

It is less useful, however, in cases when the AI systems are complex and there is a human in the loop:

Take, for example, the case of automobiles offered by the Expert Group: In the case of conventional vehicles, the owner is typically held responsible for any damages resulting from the vehicle's use. In such situations, strict liability is commonly enforced, irrespective of whether the owner acted intentionally or negligently. This is because the owner is the one who generally benefits from operating the vehicle, and exercises the highest level of control over the associated risks by determining when, where, and how to use, maintain, and repair it. Consequently, the owner is also the most cost-effective party to avoid and manage risks and obtain insurance coverage. This changes with autonomous vehicles (AVs). AVs are privately owned, the individual owner decides when to use the AV and puts the destination into the system, but many other decisions are taken by algorithms provided by the producer of the AV or a third party acting on the producer's behalf. In some cases, the producers may also be in charge of operating the vehicle's navigation system. The driver, however, is often required to be vigilant when the vehicle is driving and be ready to take control. Who should, in such a scenario, be held strictly liable in the case of an accident?[200]

Buiten et al. note that in a scenario where producers are under a strict liability regime, they will have incentives to take optimal precautions.[201] However, a strict liability regime fails to create incentives for victims and deployers to take care when they can, and when their actions can also lead to an

---

[197] Expert Group on Liability and New Technologies *supra* note 16 , at 28.

[198] Wendehorst, *supra* note 4, at 208.

[199] *Id.*

[200] *See* Expert Group on Liability and New Technologies *supra* note 16, at 37.

[201] *See* Buiten et al. *supra* note 4., at 10.

accident.[202] Such a regime may also be unfair from the perspective of the participant whose actions didn't contribute to the harm.

The Expert Group also noted that strict liability may have important impacts on technological advancement. Some individuals or entities may become more hesitant to actively promote technological research if the risk of liability is perceived as a deterrent.[203] The flipside of this is that activities that are beneficial to society but also risky may be reduced below the optimal level because costs will be internalized while positive externalities will not flow back directly to developers.[204] It could lead to those causing harm to be less interested in developing or deploying AI applications that inherently carry risks, even when sufficient precautions are in place, reducing the beneficial use of AI applications below their optimal level. This could be the case in instances where AI's exceptional performance reduces harm to society compared to not using AI at all.[205] For instance, autonomous car features can enhance safety, AI diagnostic tools can outperform humans in disease detection, and algorithms generate various digital services that consumers find enjoyable. While the use of AI reduces harm when compared to other options, there are also opportunity costs associated with not utilizing AI.[206]

a. C. Summary to this section

The discussion presented above presented the scholarly and policy conversation that has identified that current liability law in the EU makes it hard for victims of AI harms to claim compensation in all cases where it seems justified and that the allocation of liability can be unfair. It also showed that the decision to adopt a particular liability regime for AI harms is a policy question where the answer necessarily balances and chooses amongst different welfare and accountability objectives.

This is the case for five main reasons:

---

[202] *Id.*

[203] Expert Group on Liability and New Technologies *supra* note 16, at 28.

[204] Expert Group on Liability and New Technologies *supra* note 16 at 10.

[205] This may be the case of some instances in medicine and some autopilots, like airplanes.

[206] Buiten et al. *supra* note 4. at 10 discussing from a law & economics perspective how "the chosen liability regime should therefore be seen in the context of public policy towards innovation."

First, from the victims of an AI harm perspective, important information asymmetries exist at two levels: (1) establishing that harm occurred - especially if this is a harm to a fundamental right, and less so when it is a harm to property or physical integrity - and (2) the fault of the injurer. These asymmetries occur because, a priori, establishing that a decision was made wrongly, or that the developer of an AI system breached a standard of conduct, requires understanding how the system works, which may be hard from a victim's perspective. This is in part a function of AI's opacity and complexity.

Second, the deployers, producers and participants of an AI system perspective, face similar challenges to AI-harm victims in establishing to what extent particular damage may be attributable to the other actors they interact with.

Third, there are important difficulties in establishing who should be held accountable for a defect that arises *after* an AI system is placed in the market. This is especially the case of systems that learn from third parties but are particularly difficult, from their environment. It is also aggravated by an absence of clarity about what should be the reasonable expectations of standards of care of different actors in the AI lifecycle (should providers monitor systems? Must users install all system updates? Etc.)

Fourth, and important for our purposes, the complex interaction between human operators and AI systems is not often well addressed by liability laws. Balancing liability in light of the victim's conduct contributing to their harm is not a new problem in liability law, the challenges to establish fault also apply to the contributing victim. As mentioned earlier, the lack of clarity regarding standards of care and reasonable expectations from the human in the loop complicates the assignment of fault and responsibility. It's important to note that while the Expert Group distinguishes the standard of care – referring to the model of careful and prudent conduct required from the perpetrator of the damage – from standards of quality and safety established by law or standard-setting bodies, certain legal and technical standards may still play a significant role in determining what is reasonable to expect from the various parties involved.

Fifth and last, adopting a particular liability regime for AI harms is a policy question where the answer necessarily balances and chooses amongst different welfare and accountability objectives. Importantly these include guaranteed compensation for victims of AI harm, creating appropriate incentives to prevent harm by all actors in the AI system, and, given AI's political and economical potential, doing so in a way that does not disincentivize the development of AI systems excessively. Thus, even if, from a victim's perspective, many of the above issues could be addressed by imposing a strict liability regime for AI developers and deployers, some commentators have noted that this could disincentive the development and adoption of AI systems, which would also come at a significant social cost. Additionally, it could be unfair from the perspective of deployers who have less control over the system.

The next Part presents how the EU addressed these issues in its current proposal for AI liability in Europe.

III. THE EUROPEAN PROPOSAL FOR AI LIABILITY

In the fall of 2022 the European Commission proposed two Directives to address many of the questions explored in the previous Parts on AI liability: An AI Liability Directive (AILD) and a revision of the Product Liability Directive (PLD). The proposed AILD and PLD are elements of the European AI strategy, which was first announced in 2017.[207] The strategy seeks to establish a general EU-wide coordinated approach "to make the most of the opportunities offered by AI and to address the new challenges that it brings."[208] Elements of the approach include stepping up investment in AI research and deployment and efforts to bring AI to small businesses and users. It also includes social policies oriented at addressing changes in the labor market and, importantly, ensuring an appropriate ethical and legal framework that would support "an environment of trust and accountability around the development and use of AI."[209]

---

[207] European Commission, Communication from the Commission, Artificial Intelligence for Europe, COM(2018) 237 final.
[208] *Id.*
[209] *Id.*

Three interrelated legal initiatives seek to create the ecosystem of trust sought by the commission: The AI Act, which seeks to address fundamental rights and safety risks; a civil liability framework, which is composed of the directives at issue here, the revision of the PLD and a new Artificial Intelligence Liability Directive; and a revision of sectoral safety legislation, such as Machinery Regulation and the General Product Safety Regulation.[210] (This piece does not discuss directly the relevant sectoral safety regulations).[211]

Specifically, the proposed legal framework for AI Liability in Europe seeks to ensure that victims of AI harms enjoy the same level of protection as people harmed by other technologies. In so doing, the framework will also instill confidence in AI from the consumer part, while also providing legal certainty for providers and deployers.[212] It does so by creating a strict liability regime for safety harms by AI-powered products to natural persons and enabling judges to require the disclosure of evidence from AI producers and developers during liability procedures.[213]

This Part presents and discusses the two proposed AI liability directives as they relate to the framework first developed by the AI Act. It shows that, even though the directives make progress in important directions, because they strongly rely on the AI Act they may be falling short of meeting its objectives in three ways:

First, the AI Act mandates that only a limited range of AI systems' producers and deployers must furnish technical documentation about the

---

[210] European Commission, "A European approach to artificial intelligence", Shaping Europe's Digital Future, https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[211] The proposed Machinery Regulation and the proposed General Product Safety Regulation which revise the existing Machinery Directive and General Product Safety Directive, aim, in their respective fields, to address the risks of digitalization in the area of product safety, but not liability. *See* European Commission, The general Product Safety Directive, https://commission.europa.eu/business-economy-euro/product-safety-and-requirements/product-safety/consumer-product-safety_en; European Commission, *Machinery,* https://single-market-economy.ec.europa.eu/sectors/mechanical-engineering/machinery_en.

[212] *See* European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)) at B; European Commission, Liability Rules for Artificial Intelligence, https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en .

[213] *See* infra

system's operation, maintain records on its functionality, and design the system to facilitate user comprehension and human oversight.[214] Given that the AI liability rules rely on documentation disclosure to address information asymmetry particular to AI, this measure predominantly benefits the victims of systems subject to AI Act obligations, and creates uncertainty about the effectiveness of the proposed framework to facilitate redress in cases where less-regulated systems are involved.

Second, the liability regime treats human-hybrid systems in a somewhat contradictory manner. On the one hand, many of the liability questions related to AI-human hybrid systems will end up being about whether a particular system was designed to actually enable human supervision, and if it was, may still result in outcomes that situate responsibility on the human at issue. Though focusing the attention on whether a system is "fit for purpose" is an important improvement from the status quo, proving this may again be hard in the case of systems that are not subject to higher standards by the AI Act. On the other hand, the AILD also *excludes* from its application systems where AI are only advising humans but not effectively deciding. As we discussed in the previous part, this seems to be a too simplistic assumption about human-AI interactions. This rule, however, may create incentives for AI designers to design systems so that humans have apparent control, even if more collaborative or even automated systems may be safer and better.

Thirdly, the AI liability regime includes an important vulnerability for the effective safeguarding of fundamental rights. The exclusion of systems where AI recommends outputs, this may leave out several AI systems used to make decisions that are consequential to fundamental rights. Even when the AILD applies, EU Law and Member State liability laws typically require actual harm for compensation. Though in some cases immaterial harms are granted for the affectation of fundamental rights, the requirements for this vary from member-state to member state, and, oftentimes, immaterial harms may also not have occurred. While AI harm victims can still seek redress through fundamental rights law, the AI liability regime is the one that will offer a more accessible burden of proof to counteract the challenges posed by the opacity, complexity, and generative nature of AI systems. This raises questions about the appropriateness of this chosen approach.

To show and discuss these elements in some more detail, this Part proceeds as follows. First, it presents the general framing of AI regulation in the EU, the AI Act. This section can easily be skipped by readers who are

[214] *See infra* Part I 1.

familiar with the general orientation of the provision. The second and third subsections present PLD and the AILD briefly. The fourth section simulates how their application would affect three AI accidents, similar to the examples presented in Part I. Finally, the fifth section concludes with an analyzis.

A. The AI Act

The cornerstone of the European AI regulation is the AI Act.[215] The Act will be an umbrella and union-wide regulation that proposes a risk-based approach to AI regulation, which seeks to ensure that products and services integrate safety and security by design.[216] According to Article 1, its purpose is "to promote the uptake of human-centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and the rule of law, and the environment (…)."[217] Similarly, it establishes as one of the guiding principles of the Act "promote a coherent human-centric European approach to ethical and trustworthy Artificial Intelligence" and establishes that "AI systems shall be developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans."[218]

This section provides a concise overview of the main safety requirements introduced by the AI Act. It then delves into the human supervision requirement and, then highlights the significance of standardization and conformity assessments in the subsequent implementation of the act but, also, in defining the expectations from the various parties involved in AI systems.

1. Levels of risk and key safety requirements

The AI Act applies to providers and deployers of AI systems in the EU.[219] It defines providers as the natural or legal person that develops an AI system with a view of placing it in the market, and deployers as the natural

---

[215] European Commision, *Report on the safety and liability implications of AI supra* note 9, at 4.
[216] *Id.*
[217] AI Act – IMCO-LIBE Draft Compromise Amendments, (14 June 2023), Amendment 140, Article 1.
[218] AI Act – IMCO-LIBE Draft Compromise Amendments, *supra* note 217, Amendment to Article 4.
[219] AI Act, *supra* note 12, Art. 2.

or legal person that uses the AI system.[220] It then divides AI systems into four different levels of risk based on their intended use and regulates them differently: there is a limited set of systems that pose an unacceptable risk and are therefore banned.[221]

Most of the Act is concerned with the safety requirements for high-risk systems, which are a limited set of systems identified in Annex III. These include systems that are either intended to be used as a safety component in motor vehicle security, those used in the management and operation of critical infrastructures, like road traffic or the supply of utilities, biometric identification systems, and AI systems intended to be used in educational and employment settings to determine, respectively, access to institutions or recruitment.[222] Generative AI systems are a separate category, which will have to comply with similar requirements to high-risk systems. Lastly, limited risk systems, are all other AI systems, which must simply comply with minimal transparency requirements that will allow users to use and interact with them in an informed manner.[223] What follows focuses only on the obligations for providers and users of high-risk systems and Generative AI systems.

Chapter 2 of the Act establishes the requirements for high-risk systems. Providers of high-risk AI systems will have to comply with seven key requirements:

(1) *Risk management:* Providers must establish, implement, and maintain a risk management system that runs throughout the entire lifecycle of the AI system. The system will have to identify and analyze the known and foreseeable risks associated with the AI system, estimate, and evaluate the risks, and adopt suitable risk management measures.[224]

(2) *Data governance:* The Act also establishes certain requirements on

---

[220] AI Act, *supra* note 12, Art. 3 (2), (5) Note that individuals who are subject to AI systems have no role to play in the AI act, more on that later. This is according to the latest version of the Act. In former versions, the Act has referred to deployers, as users.

[221] European Parliament *EU AI Act: first regulation on artificial intelligence*, News-European Parliament, June 15, 2923, https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[222] AI Act, *supra* note 12, Chapter II; *see also* Lilian Edwards, Expert Explainer: The AI Act proposal, Ada Lovelace Institute, April 8, 2022.

[223] European Parliament, EU AI Act: first regulation on artificial intelligence, *News, European Parliament*, Aug. 6, 2023, https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[224] AI Act, *supra* note 12, Art. 9.

the data involved in the training of models and how it is governed. Producers will also have to evaluate the availability, quantity and suitability of the data, examine it for possible biases, and identify any gaps or shortcomings and how those gaps will be addressed.[225] Training, validation and testing data "shall be relevant, representative, free of errors and complete" and have appropriate statistical properties.[226]

(3) *Produce technical documentation:* Draw up technical documentation of all high-risk systems before the system is placed on the market to demonstrate compliance with all the above compliance. This documentation must be kept up to date during the system's life cycle.[227]

(4) *Record keeping*: Design AI systems to automatically record events while operating so that the AI system's functioning is traceable throughout its lifecycle.[228]

(5) *Transparency:* Design AI systems so that their operation is "sufficiently transparent to enable users to interpret the system's output and use it appropriately."[229] AI systems must also be accompanied by instructions for use, which shall include the characteristics, capabilities and limitations of the system; and the human oversight measures as well as the technical measures put in place to facilitate the interpretation of the outputs of AI systems.[230]

(6) *Human oversight*: The Act requires that high-risk AI systems are so designed and developed that they can be effectively overseen by a natural person while in use. Human oversight shall aim at preventing or minimizing the risks raised by the AI system.[231]

(7) *Accuracy, robustness and cybersecurity*: lastly, AI systems shall be designed and developed so that, depending on their intended purpose they achieve an appropriate level of accuracy, robustness of cybersecurity.[232]

Distributors, importers, and deployers of high-risk AI systems are subject to the obligations of the Providers when they (a) put their name or trademark on a high-risk AI system already placed in the market, (b) make a substantial modification to a high-risk AI system that has already been placed

---

[225] AI Act, *supra* note 12, Art. 10.2.
[226] AI Act, *supra* note 12, Art 10.3.
[227] AI Act, *supra* note 12, Art.11.
[228] AI Act, *supra* note 12, Art. 12.
[229] AI Act, *supra* note 12, Art. 13.1.
[230] AI Act, *supra* note 12, Art. 13.2.
[231] AI Act, *supra* note 12, Art. 14.
[232] AI Act, *supra* note 12, Art. 15.

in the market, or (c) they modify the intended purpose of the system.[233] Where these circumstances occur, the provider of the system will no longer be considered the provider of that specific AI system.[234] Deployers of high-risk systems will also have to conduct an assessment of the impact on fundamental rights of high-risk systems in the context of use.[235]

Lastly, the EU parliament added a few amendments for providers of foundation models in Article 28b in the spring of 2023, as generative AI took the world by surprise. In sum, foundational models shall, prior to making these models available, (a) analyze and mitigate for reasonably foreseeable risks to health, safety, fundamental rights, the environment, and democracy (b) process and incorporate only datasets that are subject only to appropriate data governance measures, (c) design and develop the model to achieve appropriate levels of performance, predictability, interpretability safety throughout its lifecycle, (d) design the model to reduce its energy use, (e) draw technical documentation with instructions to use, (f) establish a quality management system to ensure compliance with these requirements and (g) register in an EU database for foundation models.[236]

2. The human-in-the-loop requirement

One of the key objectives of the EU's regulatory framework is to promote the development of AI systems that serve people, respect human dignity and personal autonomy, and that function "in a way that can be appropriately controlled and overseen by humans."[237] Consequently, and as described above, the AI Act establishes in Article 14 that high-risk systems must be so designed and developed that they can be effectively overseen by a natural person while in use.

The latest version at the time of writing seems to have tried to

---

[233] AI Act, *supra* note 12, Art. 28.

[234] AI Act, *supra* note 12, Art. 28.2

[235] The assessment shall in particular pay attention to the categories of people likely to be affected by the system, reasonably foreseeable impacts on fundamental rights, and a detailed plan on how those harms will be mitigated, amongst others. Amendment Article 29 a, Fundamental rights impact assessment for high-risk AI systems, AI Act – IMCO-LIBE Draft Compromise Amendments, *supra* note 217.

[236] AI Act – IMCO-LIBE Draft Compromise Amendments, *supra* note 217, Amendment Article 28 b Obligations of the provider of a foundation model.

[237] AI Act – IMCO-LIBE Draft Compromise Amendments, *supra* note 217, Amendment Article 4.

accommodate some of the research, and critiques to human-in-the-loop requirements presented in Part I C. Thus, the current version of the Article emphasizes, for example, the design of "appropriate human-machine interface tools" so that high-risk AI systems can be "effectively overseen by natural persons." Similarly, it also requires that individuals in charge of the oversight must have sufficient AI literacy, and are appropriately enabled to understand the system, be aware of the possibility of relying and over-relying on the system, interpret its output, "be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system; [and] be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure."[238]

---

[238] *See* AI Act – IMCO-LIBE Draft Compromise Amendments, *supra* note 217, Article 14 *Human Oversight*:

1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they are effectively overseen by natural persons, as proportionate to the risks associated with those systems. Natural persons in charge of ensuring human oversight shall have sufficient level of AI literacy (…) and the necessary support and authority to exercise that function, during the period in which the AI system is in use and to allow for thorough investigation after an incident.
2. Human oversight shall aim at preventing or minimising the risks to health, safety, fundamental rights or environment that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter and where decisions based solely on automated processing by AI systems produce legal or otherwise significant effects on the persons or groups of persons on which the system is to be used.
3. Human oversight shall take into account the specific risks, the level of automation, and the context of the AI system and shall be ensured through either one or all of the following types of measures:

a. identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service;
b. identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the user.

2. For the purpose of implementing paragraphs 1 to 3, the high-risk AI system shall be provided to the user in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate to the circumstances:

a. fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;
b. remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
c. be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;

3. Conformity

Importantly, the Act establishes that providers of high-risk systems and foundation models can demonstrate conformity with these requirements through self-assessment based on internal control, either by their own plan or by following a relevant harmonized technical standard. If high-risk AI systems are in conformity with harmonized standards, they will be presumed to be in conformity with those requirements and in compliance with the requirements of the Act and EU law protecting fundamental rights and values.[239] These standards are yet to be developed but, as several commentators to the act have noticed, they will be critical to achieving the substantive goals of the Act.[240]

B. Liability for "material damages caused to natural persons by AI-Powered products:" The Revised PLD

The revision of the Product Liability Directive, adopted in 1985, seeks to adapt the EU's product liability regime to new technologies. Recall that, in essence, previous studies of the existing PLD had found that it needed updating because (1) it was legally unclear how to apply the PLD's definition of products to software, and AI-powered movable objects, and (2) it is difficult for injured people in complex cases, like those involving smart or AI-enabled products, to prove that the product was defective and caused the damage they suffered.[241]

---

d. be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the a high-risk AI system;
e. be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure.

(...)

[239] AI Act, *supra* note 12, Article 40 establishes that "[h]igh-risk AI systems and foundation models which are in conformity with harmonised standards or parts thereof the references of which have been published in the Official Journal of the European Union (…) shall be presumed to be in conformity with the requirements set out in Chapter 2 of this Title or Article 28b, to the extent those standards cover those requirements.

[240] *See* Edwards *supra* note 222; similarly, Michael Veale and Frederik Zuiderveen Borgesius arguing that "standardization is arguably where the real rule-making in the Draft AI Act will occur." Michael Veale and Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act - Analysing the good, the bad, and the unclear elements of the proposed approach*, 22 Computer Law Review International 4, 2021, at 8, 9.

[241] *See supra* Part II. B. 2; an additional element identified by the Commission in the Explanatory Memorandum is that "the rules excessively limited the possibility of making compensation claims. Property damage worth less than Eur 500 is not recoverable under the

The revision of the PLD thus aims to ensure that liability rules reflect the nature and risks of the new digitally, powered products, or ease the burden of proof in complex cases and ease restrictions on making claims "while ensuring a fair balance between the legitimate interests of manufacturers, injured persons and consumers in general."[242] The Directive, as its predecessor, does so by establishing the principle of strict liability of the relevant economic operators "as the sole means of adequately solving the problem of a fair apportionment of the risks inherent in modern technological production."[243] To do so the Directive is structured in 4 chapters, which are described in what follows with a focus on how they apply to AI systems:

Chapter I on general provisions lays out the subject matter, the scope of the directive, and key definitions. According to Article 1, the directive (1) lays out liability rules for economic operators, (2) for damages caused to natural persons, (3) by defective products.[244] Article 4 defines economic operators as the manufacturers of a product or a component, the provider of a service, and the importer or de distributors.[245] Natural or legal persons that modify a product substantially after it has already been placed in the market will also be considered economic operators.[246] Damages are defined as material losses, which can be death or personal injury or harm to, or destruction of, property (but excludes the product itself);[247] and a product as tangible or intangible movable and include software in the definition.[248] The Directive, thus, expands the application of the PLD regime to software and AI systems and AI-enabled goods (such as many smart-home devices).[249] It also expands its application to digital services that are integrated or integrated with a product "in a way that would prevent the product from performing one of its functions,"[250] such as navigation software in an autonomous vehicle.[251]

---

existing PLD). *See* Explanatory Memorandum New PLD *supra* note 195, at 1.

[242] *Id.* at 2

[243] European Commission, Proposal for a Directive of the European Parliament and the Council on liability for defective products, COM/2022/495 final, Recital 2 [*hereinafter* Proposal New PLD]

[244] Proposal New PLD *supra* note 243, Art. 1.

[245] Proposal New PLD *supra* note 243 Art. 4(16).

[246] Proposal New PLD *supra* note 243 Art. 7(4).

[247] Proposal New PLD *supra* note 243 Art. 4(6).

[248] Proposal New PLD *supra* note 243 Art. 4(1).

[249] *See* Proposal New PLD *supra* note 243, Explanatory Memorandum at 3, *see also* recital 12.

[250] Proposal New PLD *supra* note 243, Art. 4(4).

[251] Proposal New PLD *supra* note 243, Recital 15.

It, however, limits its application to material losses (bodily injury - including psychological health - and property), and to losses suffered by natural persons. Affectations to fundamental rights are thus left out, as also harms that may be suffered by legal entities, or other actors in the AI supply chain.

Chapter II lays out the key rights and obligations of the product liability regime. According to Article 5, every natural person who suffers damage from a defective product is entitled to compensation.

Article 6 defines defectiveness as the circumstances when a product "does not provide the safety which the public at large is entitled to expect," taking into account the presentation of the product including instructions for installation and maintenance,[252] and the expectations of the end-users for whom the product is intended,[253] reasonable use and misuse of the product,[254] the safety requirements of the product.[255], the moment in time when the product was placed in the market and the moment in time when the product leaves the control of the manufacturer.[256] The distinction between the moment in time at which a product is placed in the market, and the moment at which it leaves the manufacturer's control seeks to reflect that many products, such as AI systems, remain within the manufacturer's control even after being placed in the market.[257] Similarly, the recitals of the proposed directive explain that "the defectiveness of a product should be determined by reference not to its fitness for use but to the lack of the safety that the public at large is entitled to expect. . . . [this] should be assessed by taking into account, inter alia, the intended purpose, the objective characteristics and the properties of the product in question as well as the specific requirements of the group of users for whom the product is intended."[258] Thus, some products, like medical devices, may be held to higher standards of high safety expectations.[259]

Importantly, the Proposed PLD establishes that it is on the claimant to prove the elements of product-strict liability: defectiveness of the product, the damage suffered and the causal link between the effectiveness and the

---

[252] Proposal New PLD *supra* note 243, Art 6(a).
[253]Proposal New PLD *supra* note 243, Art 6(h).
[254] Proposal New PLD *supra* note 243, Art 6(b).
[255]Proposal New PLD *supra* note 243, Art. 6(f).
[256] Proposal New PLD *supra* note 243, Art 6(e).
[257] Proposal New PLD *supra* note 243. Recital 23.
[258] Proposal New PLD *supra* note 243, Recital 22.
[259] *Id.*

damage.[260] Additionally, the directive establishes a rebuttable presumption of fact to further alleviate the claimant's burden of proof in certain conditions.[261] The defectiveness will be presumed if the claimant establishes that the product does not comply with mandatory safety requirements when the damage was caused by an obvious malfunction of the product during normal use and circumstances.[262] Article 8 establishes that national courts must be empowered to order the defendant to disclose relevant evidence that is at its disposal, upon request of an injured person claiming compensation for damage caused by a defective product, and when the claimant has presented facts and evidence sufficient to support the plausibility of the claim for compensation.[263]

Defectiveness will be also presumed when the defendant fails to comply with the obligation to disclose relevant evidence,[264] and when it is established that the product is defective and the damage caused is of a kind typically consistent with the defect in question.[265]

Chapter III establishes some general provisions on liability. Regarding the possible defenses, manufacturers and distributors will not be liable if they are able to prove that it is probable that the defect that caused the damage did not exist when the product was placed on the market or put into service;[266] that the defectiveness is due to compliance of the product with mandatory regulations;[267] that the product is up to the state of the scientific and technical knowledge at the time it was placed in the market.[268] If an economic operator proves that they did not place the product on the market they will also not be liable.[269] However, economic operators will not be exempted from liability when the defect is due to software updates or upgrades, or a lack thereof.[270]

Regarding the interaction with third parties who may be involved, Article

---

[260] Proposal New PLD *supra* note 243, Art. 9.
[261] Proposal New PLD *supra* note 243, Recital 33.
[262] Proposal New PLD *supra* note 243, Art. 9
[263] Proposal New PLD *supra* note 243, Art. 8
[264] Proposal New PLD *supra* note 243, Art. 9
[265] Proposal New PLD *supra* note 243, Art. 9
[266] Proposal New PLD *supra* note 243, Art 10(c)
[267] Proposal New PLD *supra* note 243, Art. 10(d)
[268] Proposal New PLD *supra* note 243, Art. 10(e)
[269] Proposal New PLD *supra* note 243, Art. 10(a), (b) *see also* Art. 11. … where two or more economic operators are liable for the same damage pursuant to this Directive, they can be held liable jointly and severally.
[270] Proposal New PLD *supra* note 243, Art. 10.2

12 establishes that economic operators can not reduce their liability when a third party's actions or omissions contributed to the harm.[271] However, when the damage was caused by the defectiveness of the product and the faulty action of a third party or the victim, their liability may be reduced.[272] As explained in Recital 36, this is established in the interest of a fair apportionment of risk.[273]

Chapter IV deals with some procedural issues regarding the implementation of the directive in member states.

C. All other AI Harms (with a focus on fundamental rights): The AILD

Unlike the PLD, which seeks to adapt an already existing but specific regime to AI and other new technologies, the AILD seeks to adapt, in general, national liability rules to the challenges posed by claims for damages caused by AI-enabled products and services. As explained in Part II, current civil liability rules typically establish that victims of harm need to prove the existence of harm, wrongful action or omission by another person, and the causal link between the harm and the action.[274]

The AILD is a relatively short directive - it has only 9 articles, four of which are concerned with its implementation.[275] Its objective, however, is to address the ways in which the characteristics of AI systems - including complexity, opacity and autonomy - "may make it difficult or prohibitively expensive for victims to identify the liable person and prove the requirements for a successful liability claim."[276] Thus, the AILD seeks to lay out a set of uniform rules at the EU level that ensures that victims of damage caused by AI have an equivalent level of protection under civil liability rules as victims of equivalent harms caused without AI systems.[277]

---

[271] Proposal New PLD *supra* note 243 Art. 12.1

[272] Proposal New PLD *supra* note 243 Art 12.2 This echoes the principle of the contributory conduct tor activity of the victim *see supra xx*; PETS *supra* note 6, Article 8:101

[273] Proposal New PLD *supra* note 243, Recital 36.

[274] *See supra xx*

[275] Proposal New PLD *supra* note 243, Arts. 6, 7,8, 9.

[276] European Commission, Explanatory Memorandum, 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, at 1 [*hereinafter* Proposal AILD].

[277] European Commission, Explanatory Memorandum, AILD at 10, *see also* Proposed AILD art. 1

Article 1 establishes its subject matter and scope: to lay out common rules on "the disclosure of evidence on high-risk artificial intelligence (AI) systems to enable a claimant to substantiate a non-contractual fault-based civil law claim for damages."[278] Second, it lays down common rules on "the burden of proof in the case of non-contractual fault-based civil law claims (...) for damages caused by an AI system."[279] In doing so, the AILD expects to ease the burden of proof through the use of disclosure and rebuttable presumptions, relying on the information that will be documented pursuant to the AI Act.[280] It explicitly does not adopt more far-reaching changes in standard fault-based liability (such as reversal of the burden of proof, or an irrebuttable presumption) because of how costly this could be for developers or deployers.[281]

Importantly for the purposes of this piece, Recital 15, establishes that the Directive should only cover claims for damages "when the damage is caused by an output or the failure to produce an output by an AI system through the fault of a person," for example the provider or the deployer.[282] According to the Recital

> "[t]here is no need to cover liability claims when the damage is caused by a human assessment followed by a human act or omission, while the AI system only provides information or advice which was taken into account by the relevant human actor. In the latter case, it is possible to trace back the damage to a human act or omission, as the AI system output is not interposed between the human act or omission and the damage, and thereby establishing causality is not more difficult than in situations where an AI system is not involved."[283]

Thus, the alleviations introduced by the directive will not apply if a human agent intercedes between the AI output and the damage. As other commentators have noted, this may leave significant amounts of the AILD proposal inapplicable, as the AI Act will require that high-risk systems shall be designed and developed so that they *can* be effectively overseen by natural persons (as proposed in the text by the Commission) or so that "they be

---

[278] Proposal AILD, *supra* note 276, Art. 1(a)
[279] Proposal AILD, *supra* note 276, Art. 1(b)
[280] European Commission, Explanatory Memorandum, AILD
[281] European Commission, Explanatory Memorandum, AILD
[282] Proposal AILD, *supra* note 276, Recital 15.
[283] Proposal AILD, *supra* note 276, Recital 15.

effectively overseen by natural persons" (as proposed by the Parliament).[284]

At the same time, it is important to highlight that one of the key objectives of the AILD proposal is to enable the effective private enforcement of fundamental rights and preserve the right to remedy where AI-specific risks have materialized.[285] It is thus a central complement of the AI Act, and other regulations already in place in Europe such as the GDPR, the DSA, and EU non-discriminatory and equal treatment regulations.[286]

Article 2 covers key definitions - in general referring back to the AI Act,[287] and Articles 3 and 4 introduce the key measures of the Directive:

Article 3 establishes that national courts will be empowered to demand the disclosure of relevant evidence from high risk systems which is "a provider, a person subject to the obligations of a provider (...) or a user to disclose relevant evidence at its disposal about a specific high-risk AI system that is suspected of having caused damage."[288] To make such a request claimants must meet two conditions: First, they must present facts and evidence sufficient to support the plausibility of a claim for damages.[289] Second, they must have previously asked the provider - or the person acting as such - for evidence about that specific high-risk system, and the request must have been refused.[290] Where a defendant fails to comply with an order to disclose or preserve evidence at its disposal, a rebuttable presumption of not compliance

---

[284] Proposal AI Act, June 14, European Parliament adopted negotiating position. *See* Philip Hacker, The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future, Working Paper, at 19. Available at https://arxiv.org/pdf/2211.13960.pdf .

[285] Proposal AILD, *supra* note 276, Explanatory Memorandum, at 10.

[286] Proposal AILD, *supra* note 276, Explanatory Memorandum; *see also* Proposal AILD, *supra* note 276 Recital 2 "While such requirements intended to reduce risks to safety and fundamental rights are meant to prevent, monitor and address risks and thus address societal concerns, they do not provide individual relief to those that have suffered damage caused by AI."

[287] *See* Proposal AILD, *supra* note 276, Recital 15.

[288] Proposal AILD, *supra* note 276 Art. 3.1 "Member States shall ensure that national courts are empowered, either upon the request of a potential claimant who has previously asked a provider, a person subject to the obligations of a provider pursuant to [Article 24 or Article 28(1) of the AI Act] or a user to disclose relevant evidence at its disposal about a specific high-risk AI system that is suspected of having caused damage, but was refused, or a claimant, to order the disclosure of such evidence from those persons."

[289] *Id.*

[290] Proposal AILD, *supra* note 276, Art. 3.1 paragraph 2.

with a duty of care will be established.[291]

Article 4 establishes a rebuttable presumption of a causal link in the case of the fault of the defendant. This is meant to address the challenges claimants face to establish a causal link between non-compliance with an output produced by an AI system or a failure of an AI system.[292] This is subject to three requirements: (a) the claimant must have established the fault of the defendant (or the court has presumed it pursuant to Article 3), and the existence of harm,[293] (b) it is "reasonably likely, based on the circumstances of the case," that the fault or the failure of the AI system led to the output,[294] and (c) the claimant demonstrated that the output gave rise to a damage.[295]

In the case of high-risk systems, the "fault of the defendant consisting of the non-compliance with a duty of care" will be presumed where the complainant has demonstrated that the provider, or the person subject to the provider's obligations, failed to comply with any of the safety requirements contained in the AI Act for high-risk systems.[296] It will also be presumed on the basis of non-compliance with a court order for disclosure of evidence.

If the claimant is the provider of the system, these requirements are that the AI was not trained using training, validation and testing data sets that meet the quality criteria established in the AI Act;[297] the AI system was not designed and developed to meet the transparency requirements laid down in the AI Ac;[298] the AI system was not design and developed in a way that

---

[291] Proposal AILD, *supra* note 276, Art. 3.5 Importantly, Article 2 defines the duty of care as "a required standard of conduct, set by national or Union law, in order to avoid damage to legal interests recognised at national or Union law level, including life, physical integrity, property and the protection of fundamental rights." Proposed AILD, Art. 2(9)

[292] Proposal AILD, *supra* note 276,, Explanatory Memorandum, at 10.

[293] Proposal AILD, *supra* note 276, Art. 4.1(a).

[294] Proposal AILD, *supra* note 276,, Art. 4.1 (b).

[295] Proposal AILD, *supra* note 276, Art.4.1(c).

[296] Proposal AILD, *supra* note 276, Art. 4.1 Article 4.5 and 4.6 also establishes that "in the case of a claim for damages concerning an AI system that is not a high-risk AI system, the presumption laid down in paragraph 1 shall only apply where the national court considers it excessively difficult for the claimant to prove the causal link mentioned in paragraph 1.(...) In the case of a claim for damages against a defendant who used the AI system in the course of a personal, non-professional activity, the presumption laid down in paragraph 1 shall apply only where the defendant materially interfered with the conditions of the operation of the AI system or if the defendant was required and able to determine the conditions of operation of the AI system and failed to do so."

[297] Proposal AILD, *supra* note 276, Art. 4.2(a).

[298] Proposal AILD, *supra* note 276, Art. 4.2(b).

"allows for an effective oversight by a natural person," as laid out in the Act;[299] the system was not designed and developed to achieve an appropriate level of accuracy, robustness and cybersecurity in the light of its intended purposes, pursuant to the AI Act;[300] and lastly "the necessary corrective actions were not immediately taken to bring the AI system in conformity with the obligations" established in the AI Act.[301]

If the claim for damages is against the deployer of the high-risk AI system, the fault of the defendant will be established automatically in two instances: (a) If they "did not comply with its obligations to use or monitor the AI system in accordance with the accompanying instructions of use or, where appropriate, suspend or interrupt its use" pursuant of its obligations under the AI Act;[302] and if the claimant proves that the deployer "exposed the AI system to input data under its control which is not relevant in view of the system's intended purpose."[303]

In the case of non-high-risk AI systems, Article 4(5) establishes that the presumption of causality will apply only if the court determines that it is excessively difficult for the claimant to prove the causal link between damage and fault. This should be assessed given the characteristics of certain AI systems, such as their autonomy or opacity.[304]

Lastly, Article 5 creates a monitoring program to provide the European Commission with information on incidents involving AI systems. This intends to assess whether additional measures would be needed, such as introducing a strict liability regime and/or mandatory insurance.[305]

D. AI Harms (Second Act)

So, what would happen if an AI system causes harm and the Revised PLD and AILD are already in place? How will this new regime help address some of the challenges associated with better accounting for the complex

---

[299] Proposal AILD, *supra* note 276, Art. 4.2(c).
[300] Proposal AILD, *supra* note 276, Art. 4.2(d).
[301] Proposal AILD, *supra* note 276, Art. 4.2(e).
[302] Proposal AILD, *supra* note 276, Art. 4.3(a).
[303] Proposal AILD, *supra* note 276, Art. 4.3(b).
[304] Proposal AILD, *supra* note 276, Explanatory Memorandum, Art. 4.5.
[305] Proposal AILD, *supra* note 276, Art. 5.

human-AI interactions, and getting redress for harm to fundamental rights? This section offers two case studies, to shed some light on those questions.

5. AI and Safety

a. The Case of an Autopilot

Imagine a resident of a European city that gets involved in an accident while using one of Tesla's main competitors' cars, a Gauss.[306] This happens in a part of the city where using Autopilot is allowed. Assume the AI Act, the PLD and the AILD are in vigor (as they were presented in the previous section), and that these are the main EU-law institutions that apply; there are no special liability nor product safety rules for Automated vehicles in these cities.[307] The vehicle swerved into a curb, causing a car accident which resulted in an injury to the driver. Following the facts of the case presented in Part I, Gauss cautions drivers to keep their hands on the wheel, and "be prepared to take over at any moment." In the accident, the driver received a warning to control the vehicle less than a second before the strike, as this was when the software identified it was facing an unknown situation. Gauss says the software worked correctly.[308]

The driver sues Gauss, alleging that the Autopilot feature failed to operate safely and caused the accident. Recall that in the real-life case presented in Part I, the jury found that the Autopilot feature had not malfunctioned and that the driver's negligence caused the accident. The question in this new case would thus be whether given the new EU regulations, something would change. Considering what we now know about human-AI interactions.[309] A positive result for the victim would be that the Gauss is found to be at fault, or the software is found to be defective because it passed the control to the driver less than one second before the strike.

---

[306] Like a Tesla, a Gaus is a unit to measure a magnetic field https://www.britannica.com/science/gauss

[307] Special liability rules for road accidents exist in several countries, as do special safety regulations for AVs.

[308] *See supra* Part I.B.1

[309] *See supra* Part I C

Because this is a claim pertaining to a bodily injury, suffered by a natural person, and caused by a product, this claim is covered by the Revised PLD.[310] This is beneficial for the plaintiff as they would not have to establish fault, they only have to prove that the product was defective, the damage suffered, and the causal link amongst both.[311]

A second legal element is that AV software is the high-risk category under the AI Act.[312] Recall that, in this example, AVs are high-risk systems under the AI Act and have to comply with the relevant security requirements (In real life, the EU is expected to draft specific security rules for AVs and they will be exempt from the core obligations of the AI Act. Let's assume, for this example, that these will be equivalent to those of the AI Act.)[313] Thus, Gauss is obliged to meet safety requirements such as producing technical documentation and record keeping, human oversight, and transparency.[314]

Let's also assume the damage is easily established (the injury). According to Article 6 of the Revised PLD, a product is defective if it does not "provide the safety which the public at large is entitled to expect."[315] This includes the presentation of the product, instructions, etc;[316] the reasonably foreseeable use and misuse of the product,[317] product safety requirements,[318] and the specific expectations of the end-users for whom the

---

[310] Proposal New PLD *supra* note 243, Art. 1 "This Directive lays down common rules on the liability of economic operators for damage suffered by natural persons caused by defective products."

[311] Proposal New PLD *supra* note 243, Article 5; Art. 9.1

[312] Software supporting motor vehicles is under the current high-risk category of the AI Act, but it is also expected that specific regulations will be developed. *See* David Fernandez Llorca and Emilia Gomez Gutierrez, *Artificial Intelligence in Autonomous Vehicles towards trustworthy systems*; European Commission 2022 (JRC128170) Available at https://publications.jrc.ec.europa.eu/repository/handle/JRC128170.

[313] *See* Hacker *supra* note 284 at 21: "Technically, autonomous vehicles will be considered high-risk (Article 6(1) and (2) AI Act), but are exempt from all of the core obligations of the AI Act (Articles 2(2) and 84 and Annex II Section B No. 2, 3, 6 and 7 AI Act), hence rendering the relevant references in Articles 3 and 4 AILD Proposal inapplicable to them."

[314] *See* supra

[315] Proposal New PLD *supra* note 243, Art. 6.1.

[316] Proposal New PLD *supra* note 243, Art. 6.1(a).

[317] Proposal New PLD *supra* note 243, Art. 6.1(b).

[318] Proposal New PLD *supra* note 243, Art. 6.1(f).

product is intended.[319] Article 9, the defectiveness is presumed if, the plaintiff shows that the vehicle (1) does not comply with mandatory safety requirements of the product, or (2) that the damage was caused by an obvious malfunction.[320] Producers are exempt if the defect did not exist when the product was placed on the market,[321] if the defect is caused due to compliance of the product with mandatory regulations,[322] or if the state of scientific-technical knowledge at the time the product was placed on the market was not such that the defect could be discovered.[323]

Following the research on the complexities of AI-Human interactions, one of the arguments the plaintiff could raise is that handing over control less than a second before the accident is not "the kind of safety the public at large expects", nor according to the expectations of end-users who, in this case, is a regular driver (but not a professional car-racing driver, for example). It would thus be a defect. Though, in the case the plaintiff would have to provide evidence to assert this, this seems, indeed, the kind of problematic interface inspired by a MABA-MABA framework and that is being revisited.[324] Though there are, still to date, no clear standards of care about what the expected duty of care is, the human oversight requirement under the AI Act (or the future requirements for AVs in particular), may offer some guidelines: There must be "appropriate human-machine interface tools" so that high-risk AI systems can be "effectively overseen by natural persons." Similarly, it also requires that individuals are aware of the possibility of relying and over-relying on the system, and "be able to intervene in the operation."[325]

With the PLD in place, the plaintiff would be able to request documentation and evidence from Gauss about the system and its design.[326] In this case, this would include information on the technical documentation on the

---

[319] Proposal New PLD *supra* note 243, Art. 6.1(h).
[320] Proposal New PLD *supra* note 243, Art. 9.1 (b), (c).
[321] Proposal New PLD *supra* note 243, Art. 10.1(c).
[322] Proposal New PLD *supra* note 243, Art. 10.1(d).
[323] Proposal New PLD *supra* note 243, Art. 10.1(e).
[324] *See supra* Part I. C.
[325] *See* AI Act, *supra* note 12, Article 14 *see* supra note 238 for full text.
[326] Proposal New PLD *supra* note 243, Art. 8

autopilot, the AI-Human interface but also, if this were a device covered by the AI Act, the conformity assessments with the requirements of the AI Act.[327] The conformity assessment would show whether the human interface meets EU standards, or it doesn't. If it does, it will most likely be uphill for the plaintiff to prove that the interface is not of the kind the public at large expects and reasonable for the end user, as, one would expect, the standards will be developed accordingly, and the standard will most likely be in accordance with the state of scientific and technical knowledge (it may be that, once these standards are in place, handing over control to a non-professional user in less than a second is simply not in conformity). If the conformity assessment is non-compliant with the safety standards, causality will be presumed and Gauss will have to prove that this didn't cause the accident (regardless of complaints that may be filed aside, under the AI Act, for nonconformity).

In all cases, if the plaintiff did not abide by her expected standard of care and, for example, didn't follow instructions, was distracted, or was in breach of a legal obligation, the liability of the manufacturer could be reduced, but most likely not eliminated, both parties would share liability.[328] This is positive, as it would also encourage harm-reducing behavior from AI system end-users.[329] However, if the plaintiff contributed to the accident with her action or omission with no fault - perhaps she did receive control of the car, but given how control was handed it was not reasonable to expect from her that she would control the vehicle - the Revised PLD also establishes that this should not reduce the liability of the producer.[330]

b. Variations to the main theme: Non-high-risk systems, human-AI hybrid systems and the importance of standards and documentation

The example above shows a few advantages of the upcoming PLD regime but also reveals a shortcoming of the regime in place:

---

[327] *See* AI Act, *supra* note 12, Art. 19.

[328] Proposal New PLD *supra* note 243, Art 12.2 this echoes the principle of the contributory conduct or activity of the victim *see* PETS *supra* note 6, Article 8:101.

[329] *See* Buiten et al. supra note 4..

[330] Proposal AILD, *supra* note 276, Art. 12.

First, when an accident involves an AI-powered product that falls outside the high-risk system category defined by the AI Act the level of protection for victims is lower. The level of scrutiny a plaintiff can apply to the AI system's functioning, depends on the existing evidence on the functioning of a system, and the behavior of the producer and provider. Under the AI Act, however, only the producers and deployers of high risk systems and foundational models are required to produce and keep documentation about the functioning, explainability and activity log of AI systems, as well as their compliance with mandatory standards. Thus, even if under the PLD courts are always empowered to order the defendant to disclose relevant evidence, it is less clear that plaintiffs will have access to equivalent evidence than victims of harms by high-risk AI systems. If there is no explainability documentation and conformity assessments, nor pre-established standards, about AI systems, establishing effectiveness can remain significantly challenging for the victim/user.

This prevalence of that information asymmetry is aggravated in the AILD. The AILD's provision providing for disclosure of documentation only applies to high-risk systems. Victims of harms that occur by or with the participation of an AI system that is not high risk, but is still opaque or complex, will thus still face significant hurdles overcoming technical and organizational opacity of AI systems. Phillip Hacker has pointed out before that this is a problem arising due to the EU AI liability regime's excessive reliance on the risk categories defined in the AI Act.[331]

Second, the ease with which victims will be able to bring liability claims will strongly depend on the compliance of the special requirements and standards mandated by the AI Act. This is specially important in the case of hybrid systems under the PLD: When the AI Act is in place, high-risk systems will be very likely to be designed to meet the expectations and standards of the human control requirement. This should improve the interface overall and to a certain degree link the standard of conduct of developers and providers to more clearly-defined industry standards. Thus,

[331] Hacker, *supra* note 284 at 20 (echoing critiques in this sense and because the list of the AI Act is both over and under inclusive).

under a product liability claim, it may be that many of the factors at play in the Boeing 737 MAX accidents would potentially situate liability on the manufacturer: the pilots had not had access to information about the AI system in their training, and even when information was shared it did not seem to be the case that they knew how to disable it.[332] A significant amount of the legal work of proving a defect will thus be focused on proving that the human-AI interface was not fit for purpose. As above, however, if the system at issue is not a high-risk system, less extensive and accurate documentation may be available to prove such claims.

At the same time, recall that in instances where compliance with standards is what led to the harm, developers and deployers will not be held liable.[333] Though this makes sense from the developer and deployers perspective, it shifts attention to how the human in the loop requirement will be developed in the standard-setting process. If these standardization process fails to account for the difficulties discussed in Part I C, then the outcome will be undesirable and victims are likely to remain unprotected under civil liability rules vis à vis victims of harms that occur without a hybrid AI systems: developers will argue that the human was a regulatory requirement, and the human (or their employer) may be able to argue that the system was not fit for purpose.

Third, is the treatment of the human in the loop within the AILD. The AILD only applies to damages where there is no human assessment after the AI system's output. The phrasing of the article does not seem to consider yet the complexities of human assessment after an AI system provides advice. It is unclear how this recital may affect situations where humans and AI are supposed to work together. Recall that, in the Boeing 737 MAX accidents, at least one of the accidents happened because a pilot failed to steer the plane up, while the system was steering it down. In this case, it may be that the AILD will apply in this context if the plaintiffs succeed at arguing that this scenario *is not* an instance where "damage is caused by a human

---

[332] *See supra*

[333] Recall that under the PLD, manufacturers and distributors will not be liable if they are able to prove that the defectiveness is due to compliance of the product with mandatory regulations. Proposal New PLD *supra* note 243, Art. 10(d).

assessment followed by a human act or omission, while the AI system only provides information or advice."[334] It may not apply, however, to instances where control is handed over a second before an accident happens, as in the case of the Tesla autopilot. This is, unless the AILD introduces some of the nuances the newer version of the AI Act has, but plaintiffs will still need to assert and substantiate the likelihood that the human-machine system did not adequately prepare the human for effective control of the situation in order to establish the applicability of the AILD. In what seems like a circular situation, plaintiffs will have to do this to establish the authority of courts to compel AI developers to disclose pertinent evidence, while obtaining a clear understanding of the roles of humans and machines would benefit from examining the documentation of the human-AI interface and system dynamics.

6. AI and Fundamental Rights: The case of data protection

One of the main objectives of the AILD is that it will help protect, and give redress to victims of harm to fundamental rights, such as the right to non-discrimination.[335] As already discussed above, one of the key limitations of the current version of the AILD is that it ends up not applying to high-risk systems that can potentially affect fundamental rights. This is a function of the human oversight requirement of the AI Act for high-risk systems (and may partially depend on the final wording: currently the Commission and Parliament's version differ on whether AI systems must be so designed so that they *can* be effectively supervised, or so that they *are* effectively supervised by a person[336]). Indeed, and as above, the AILD is not supposed to apply to situations caused by a human assessment followed by a human act or omission where the AI system only provides information or advice.[337] Though high-risk systems are not the only type of AI system that can eventually affect fundamental rights, to the extent the list of high-risk

---

[334] Proposal AILD, *supra* note 276, Recital 15.
[335] *See supra*
[336] *See* supra Part III. A. 2
[337] Proposal AILD, *supra* note 276, Recital 15.

systems contains a list of the "usual suspects," it seems like a notable exclusion.[338]

In cases that are high-risk systems and where the damage is not caused by a human assessment followed by a human act or omission because the AI system only provides information or advice,[339] plaintiffs will be able to access information about the system which should make it easier to prove fault (assuming, again, that the EU standardization process will clarify what the standards of behavior are for AI providers and deployers). In cases that are *not* high-risk systems no such documentation will be mandatorily produced nor does the obligation to disclose information apply, which may complicate the victims' work of proving fault.[340]

In the case of an AI affectation to fundamental rights, a fundamental question is, however, what are the requirements for plaintiffs or victims to show the existence of damage, from a civil liability perspective? Recall that damage requires material or immaterial harm to a legally protected interest. Under fundamental rights law in the EU, however, whether the mere affectation of fundamental rights will be associated with non-material damages that give rise to compensation will sometimes depend on national liability laws.[341] When harmonization exists - as in the case of data protection law or antidiscrimination law - establishing an illegal affectation of a right is not enough to succeed in a liability claim; damage must be

---

[338] This is something that I would expect to be corrected as the Directive is discussed, as one of the objectives of the directive is, also, "to lay out common rules on "the disclosure of evidence on high-risk artificial intelligence (AI) systems to enable a claimant to substantiate a non-contractual fault-based civil law claim for damages" as explained in Art. 1(a)

[339] Proposal AILD, *supra* note 276, Recital 15.

[340] Proposal AILD, *supra* note 276, Art. 1.

[341] *See* European Court of Justice, UI v. Österreichische Post AG, C-300/21, at 14 [*hereinafter* UI v. OP]. *See also* Proposal AILD, *supra* note 276, Explanatory Memorandum at 10: "In addition, depending on each Member State's civil law system and traditions, victims will be able to claim compensation for damage to other legal interests, such as violations of personal dignity (Articles 1 and 4 of the Charter), respect for private and family life (Article 7), the right to equality (Article 20) and non-discrimination (Article 21)."

established.[342] How exactly to establish it is in some instances, however, unclear and will vary from Member State to Member State:

Take, for example, the case of the right to data protection and an algorithm used to infer party affinity. In the spring of 2023, the European Court of Justice (ECJ), decided a data protection case where the plaintiff brought an action against the Austrian Post Office seeking, first, an injunction to cease processing their personal data without authorization and, second, an order requiring that company to pay EUR 1000 by way of compensation for the non-material damage which he claimed to have suffered.[343] The Austrian Post Office collects information on the political affinities of the Austrian population and uses an algorithm to define "target group addresses", which it then sells to different organizations for targeted advertising purposes.[344] In the case, the Austrian Post identified that the plaintiff had a high degree of affinity with a certain Austrian political party. The plaintiff, who had not consented to the processing of his personal data, was offended by the affinity attributed to him with that party. As explained by the Court "the fact that data relating to his supposed political opinions were retained within that company caused him great upset, a loss of confidence and a feeling of exposure."[345]

In this case, the ECJ considered the question of the conditions required to exercise a right to compensation for damage as a result of an infringement on an individual's data protection rights. Indeed, article 82.1 of the GDPR establishes that "[a]ny person who has suffered material or non-material damage as a result of an infringement of this Regulation shall have the right to receive compensation from the controller or processor for the damage suffered."[346] The question, in particular, was whether, in such instances, it is necessary for data subjects to have suffered damage that is distinct from the infringement of the GDPR.[347] The Court's answer was affirmative. It

---

[342] *See* supra Part II. A. 1

[343] UI v. OP *supra* note 341, at 13.

[344] UI v. OP *supra* note 341, at 11.

[345]UI v. OP *supra* note 341, at 13.

[346] GDPR *supra* note 136, Art. 82.

[347] UI v. OP *supra* note 341; *see also* Recital 75 of the GDPR also establishes that "The risk to the rights and freedoms of natural persons, of varying likelihood and severity, may

explained that "the separate reference to 'damage' and to an 'infringement' in Article 82(1) of the GDPR would be superfluous if the EU legislature had considered that an infringement of the provisions of that regulation could be sufficient, by itself and in any event, to give rise to a right to compensation."[348] Consequently, the conditions that give rise to compensation for an infringement on an individual's data protection rights require establishing, in essence, similar conditions to any other liability claim, "namely processing of personal data that infringes the provisions of the GDPR, damage suffered by the data subject, and a causal link between that unlawful processing and that damage."[349]

As the Court explained, one of the reasons for this is that the GDPR provides for administrative and judicial remedies before a supervisory authority in case of an infringement of the GDPR, without it being necessary that the data subject must have suffered "damage."[350] Similarly, the GDPR also permits the imposition of administrative fines and other penalties, which have a punitive purpose and are not conditioned by the existence of damage.[351] Thus, even if the damage is not defined in Article 82, the injured party must prove that the consequence of the breach of the GDPR constituted non-material damage, even if it must be interpreted broadly.[352] The Court does not clarify what this means, but Member States' courts have granted small amounts of compensations for immaterial damages for data protection harms, by considering the "anguish produced by the more or less complicated process that the person concerned has had to follow for the rectification or cancellation of the incorrectly processed

---

result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorized reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; (...)" GDPR, *supra* note 136, Recital 75.

[348]UI v. OP *supra* note 341, at 34.

[349]UI v. OP *supra* note 341, at 36.

[350] UI v. OP *supra* note 341, at 39, GDPR, *supra* note 136, Art. 77.

[351] UI v. OP *supra* note 341, at 40, GDPR, *supra* note 136, Art. 83 and 84.

[352] UI v. OP *supra* note 341, at 50 (in the decision it is unclear what is a damage within the meaning of the GDPR).

data."[353] In UI v. Österreichische Post, however, the ECJ says that "[i]t is apparent from the order for reference that no harm other than those adverse emotional effects of a temporary nature has been established."[354]

E. Conclusion to this Part

This Part presented the European Approach to AI liability, as it is framed by the AI Act, and "ran it" through a couple of examples of situations where AI harms occurred: a safety harm, and a harm to a fundamental right.

This revealed that the PLD and the AILD do advance in important ways the objectives of ensuring that victims of AI harms enjoy the same level of protection as people harmed by other technologies. In particular, the strict liability regime in AI-products, and, in general, the measures that facilitate disclosure of measures and key documentation should help victims of AI harms address some of the information asymmetries at issue when seeking compensation for an AI harm.

Additionally, if the more comprehensive requirements of huma supervision proposed in the latest version of the AI Act is adopted, it is likely that in liability cases a lot of the attention will focus less on who had formal "control" and, rather, on whether the human-interface was fit for the purpose of enabling effective human supervision. This may, overall, allow for a more nuanced analysis of the hybrid systems when accidents happen.

As revealed in the last section, the proposed directives, however, still fall short in at least three main ways:

---

[353] In 2014, after the seminal judgement on the right to be forgotten, *Google Spain v. Mario Costeja*, the Barcelona Court of Appeals ordered Google to pay damages to another individual who south to remove links to some old damaging information from search results. In this case, and in a similar vein as the ECJ did recently, the Barcelona Court clarified, that the mere infringement of data protection rights does not imply a right to compensation for damages, and that there is in data protection law a presumption of damage upon an infringement of rights. In the case, the Court ordered Google o pay 8000 Euros in damages, upon a request of compensation for 2 million Euros for material damages, and 338'000 Euros for immaterial damages. The Barcelona Court found that the plaintiff had failed to prove how the data protection infringement had led to the material damages, and in general, to the immaterial damages. *See* Audiencia Provincial Barcelona, SAP B 8246/2014 - ECLI:ES:APB:2014:8246, at 28.

[354] UI v. OP *supra* note 341, at 50.

First, and in general, the liability regime does not seem to address well the challenges victims will encounter when trying to prove the existence of a defect (in the case of "products," or other special regimes), or fault on the side of the producer or deployer (in all other cases), when the AI system at issue is complex, opaque, or autonomous and not a high-risk system or foundation models under the AI Act.

This arises from the fact that the requirements to produce standardized documentation, and make systems explainable and transparent, are, under the AI Act, mainly aimed at the producers and deployers of high-risk systems and foundation models. In cases where these obligations do not extend to producers and deployers, the clarity surrounding the type of documentation that AI deployers and developers should reveal and its comprehensibility becomes less evident. Consequently, there remains a distinct possibility that victims of AI-related harm will still grapple with significant information asymmetries and associated costs when trying to prove the elements of liability.

While prioritizing certain systems over others might align with a risk-regulation perspective, maintaining such differences under an AI liability regime, negates the right to an effective remedy for *all* victims, especially those of low probability or still unforeseeable damage.[355]

Second, the liability regimes treat human-hybrid systems in a somewhat contradictory manner. On the one hand, many of the liability questions related to AI-human hybrid systems will end up being about whether a particular system was designed to actually enable human supervision - as defined by the AI Act and as will be developed in the standardization process. Thus in future cases similar to those of Tesla Autopilot, or the machine-pilot interface in the cases of the Boeing MAX, AI producers will have to show that the interface was designed to effectively allow humans to understand different situations, and assume control if needed. Increased focus on whether a human-AI interface is "fit for purpose" is an important improvement from the status quo, proving this may again be hard in the case of systems that are not subject to higher standards by the AI Act. At the same time, the focus of the AI Act on human control even when full automation would be more desirable, may lead to situations where, under the AI Liability Directive designers may rightfully claim that the defect or situation at issue arose *because* they have to comply with the human supervision requirement. From the victim's perspective, this will lead again to an undesirable scenario where

[355] *See* Hacker, *supra* note 284 at 21 making a similar point.

no one will be responsible.

On the other hand, the AILD also *excludes* from its application systems where AI are only advising humans but not effectively deciding. As we discussed in the previous part, this seems to be a too simplistic assumption about human-AI interactions and of the different ways in which humans can be biased by AI recommendation systems. It may be that many of these issues will be addressed with the adoption of the human supervision requirements in the newer versions of the AI Act, which, recall, requires that AI systems are so developed so that they can be effectively overseen and that the individual in charge has sufficient training and understanding to do so.[356] However, the research on whether this is possible and how to do this is still not conclusive.[357]

This exclusion thus may create incentives for AI designers to design systems so that humans have apparent control, even if more collaborative or even automated systems would be better. It will be important to see how the standardization process addresses this question. In any case, and especially in the cases that have obvious implications for fundamental rights - such as several high-risk systems - it would make sense to extend at least the disclosure of evidence provisions, so that the AI-Human interaction can be better scrutinized.

Thirdly, the AI liability regime includes an important vulnerability for the effective safeguarding of fundamental rights. On the one hand, since the AILD excludes from its applications systems where AI recommends outputs, this may leave out several AI systems used to make decisions that are consequential to fundamental rights. This would be the case, for example, of systems used in educational and vocational settings to determine who can access a certain program, access to essential private and public services, decisions regarding migration and asylum, and decisions assistants in the administration of justice.[358]

On the other hand, even when the AILD applies, EU Law and Member State liability laws typically require that material or immaterial harm for compensation, the mere infringement of a fundamental right is not enough (it is enough for other remedies, such as an injunction or restitution). There are cases in which courts have granted immaterial harms for the affectation of

---

[356] *See* supra

[357] *See* supra

[358] *See supra xx*

fundamental rights. However, the requirements for this vary from Member State to Member State, and, oftentimes, immaterial harms may simply not occur. Of course, victims of AI harm can still seek redress through fundamental rights law; the paradox, however, is that the AI liability regime is the one that will offer a more accessible burden of proof to counteract the challenges posed by the opacity, complexity, and generative nature of AI systems.

The next Part offers some recommendations for remedying these limitations in the current AI Liability framework in Europe, which hopefully can also be useful elsewhere.

IV. REFORM RECOMMENDATIONS AND CONSIDERATIONS FOR THE LARGER CONVERSATION

While AI can do much good, it can also do harm. AI harm might be material, to safety, life and property, for example, and immaterial, to fundamental rights, such as privacy and the right to non-discrimination in access to education and public services.[359] The characteristics of AI, and how individuals interact with AI, however, make it difficult to trace back potentially problematic decisions or outcomes made with the involvement of AI systems, which makes it difficult for victims of harm to obtain redress.

The 2022 directives proposed by the European Commission seek to update the existing liability frameworks in EU Member States, so that individuals who suffer such harm obtain fair compensation, and thus to ensure, in general, that the uptake of AI is done with individual interests in mind. As the EU strategy emphasizes, and to the extent the EU also wants to incentivize the development and adoption of "trustworthy" AI, a fit for purpose liability regime also creates legal certainty for businesses.[360]

The proposals, though certainly advancing in an important direction and part of a broader regulatory initiative. This Part however proposes a few avenues for reform and consideration to the existing proposals from a justice and accountability perspective, without unnecessarily affecting the development and deployment of trustworthy AI in Europe. Based on the key challenges identified in Part II, and the challenges identified in Part III, it proposes a few

[359] *See* White Paper on AI *supra* note 22, at 11.

[360] White Paper on AI *supra* note 22, at 13.

avenues in which the AI Liability Regime can (1) better address the information asymmetries for systems that are *not* subject to special requirements under the AI Act; (2) ensure victims of harms in AI-Human systems are not left worse off than victims of solely automatized or non automatized systems; (3) improve the redress of fundamental rights. Lastly, this Part finishes with some considerations for the specific challenges brought about by autonomous systems, such as generative AI, and some of the limitations to liability law to address AI related risks and harms.

A. Addressing information asymmetries

How does the proposed AI Liability liability regime address the information asymmetries from the victims of an AI harm perspective?

Information asymmetries between plaintiffs and AI developers and producers are importantly a function of AI opacity, because it obstructs effective inspection of AI systems.[361] The EU proposals successfully address organizational opacity, especially for high-risk systems under the AILD and, in general, under the PLD, because developers will no longer be able to assert confidentiality over the evidence. Thus, once the AI Act is applicable, there will also be adequate documentation, which should also diminish the difficulty in scrutinizing the working of a system. Similarly, the strict liability regime under the PLD, and the rebuttable presumption of causality under the AI Act, are positive adjustments to ease the burden of victims of proving causality.[362]

However, it is noteworthy that the proposed regime may better serve the victims' high-risk systems and foundation models, as defined by the AI Act, than the victims of harms of other systems. From an organizational opacity perspective, in the case of the AILD, courts' power to demand the disclosure of relevant evidence extends only to high-risk systems. Even

---

[361] Fraser et al. *supra* note 30, at 1; European Commision, *Report on the safety and liability implications of AI supra* note 9, at 16: In addition, access to the algorithm and the data could be impossible without the cooperation of the potentially liable party. In practice, victims may thus not be able to make a liability claim. In addition, it would be unclear how to demonstrate the fault of an AI acting autonomously, or what would be considered the fault of a person relying on the use of AI.

[362] *See supra*

though the AI Act's high-risk systems list is a good proxy for the systems that are most likely to cause harm, and are complex, they will not be the only systems that cause harm, nor are they the only opaque and complex systems tha may, both now and in the future, cause harm. Thus it is advisable that under the AILD, as in the PLD, courts are *always* empowered to order the defendant to disclose relevant evidence that is at its disposal, upon request of an injured person claiming compensation and when the claimant has presented facts and evidence sufficient to support the plausibility of the claim for compensation.

In the case of technically opaque or complex systems, victims seeking to prove fault under the AILD may again find it easier when the system is high-risk this is considering that explanatory documentation that can be relied upon to provide evidence will most likely be the one produced on the transparency, explainability and record keeping requirements produced under the AI Act for high-risk systems. Additionally, the development of legal and industry standards, will enable plaintiffs to compare a producer or deployer's behavior with other actor's behaviors and standards of care.

Consequently, if the power to request documentation from non-high risk systems be extended, courts should be empowered to request developers and deployers to provide ex-post explanations of how a system operates. This should be done to the extent possible and based on a reasonable justification presented by the plaintiff as to why this is needed.

B. Human-AI Hybrid systems and the role of standards

In instances where liability claims involve human-AI hybrid systems, courts should emphasize evaluating the identity of the human-AI interface. This is particularly crucial when examining cases where the human element in the loop is being considered as the cause or a contributing factor to AI-related harm.

To shift legal processes in this direction, and as the European Union's framework for trustworthy AI reaches completion, it is essential for these considerations to take center stage during the process of establishing industry standards for the human supervision requirement under the AI Act. Indeed, one of the key insights drawn from Part II is the structural role that the standard-setting process will play not only in implementing and materializing the ambitions of the AI Act but, importantly, in creating the baseline

expectations to assess and evaluate liability claims.

It is thus of critical importance that human oversight standards include, for example, the clear definition of roles and responsibilities of each party involved, consider the level of training and automation of the system in place, and account for the competencies possessed by the human actor in question. Professionals such as pilots or machine operators should arguably be held to a more stringent accountability standard compared to everyday consumers.

As in critical safety industries or other industries with experience on human-machine interactions, EU standard setting bodies and judges should pay special attention to the stated goals of the AI-Human system, the reasonability of those expectations, and systems are designed and labeled sufficiently for effective use, and address training and organizational policies.[363] Though from a liability perspective technical standards are different from standards of care, it seems inescapable that at least part of the evaluation of compliance with standards of care will rely on what are defined to be the appropriate technical standards for hybrid systems.

C. Redress to affectations of Fundamental Rights

The third element of discussion is the idoneity of the AILD as a means to seek redress for affectations of fundamental rights.

The first shortcoming of the existing AILD is the exclusion of AI systems that must be so designed that are supervised by humans from its scope of application. This would lead, for example, an algorithm like the one at issue in the Dutch scandal outside of its scope of application, but it is also especially worrisome as human supervisors are increasingly introduced as a means to specifically mitigate the risks posed by AI systems used in different forms of decision-making that can affect fundamental rights. A first key recommendation is, thus, to eliminate this requirement.

The second, more structural shortcoming, is that liability law necessarily requires the occurrence of harms to warrant compensation - the main remedy within liability law. This may constrain the victims ambition for justice, as not all affectations to fundamental rights will necessarily lead to material or immaterial harms that can be evaluated in the terms required by liability law. At the same time, in some instances fundamental rights violations are better

[363] Crootof et. al, *supra* note 6, at 466.

addressed with non-financial remedies, such as injunctions, declarations, or specific performance orders to correct the violation. Thus, several rights-protecting regulations, such as the GDPR or anti-discrimination directives provide for administrative and judicial remedies before a supervisory authority in case of an infringement of the right without it being necessary that the data subject must have suffered "damage."[364]

To be fair, the general framework for trustworthy AI under the AI Act, is centered around the understanding that the protection of fundamental rights isn't only about an individual's right itself - for example, a person's right to freedom of expression -, but it also about societies as a whole, by, for example, promoting political participation and a functioning democracy. fThese societal aspects of the protection of fundamental rights will be addressed via the enforcement of the AI Act and its safety requirements. At the same time, during the discussion process of the AI Act, the Parliament introduced a complaint process to give individuals and groups additional avenues for redress. Under this process, complaints may be lodged with the relevant national supervisory authority if they consider a given system infringes the regulation.[365] This remedy, however, focuses on the infringement of the AI Act but not on the illegal infringement of fundamental rights (and instances are possible where high-risk systems that comply with the AI Act still illegally infringe on fundamental rights. Recall that risk regulation reduces the likelihood of harm and mitigates potential risks, but does not completely eliminate the possibility of unintended consequences or unexpected interactions).

To enhance access to suitable avenues of recourse for individuals who fall victim to illegal infringements of fundamental rights involving AI systems, European and Member State authorities might contemplate the adoption or expansion of certain measures outlined in the AILD. These measures could be applied to mitigate information asymmetries in key regulations dedicated to safeguarding fundamental rights, including directives aimed at preventing discrimination.

D. The Challenge of Autonomous and Generative AI

[Still thinking about this: but the key limitations seem to be that PLD refers to material harms and is still based on the "control" of different parties -

---

[364] UI v. OP *supra* note 341,at 39, GDPR, *supra* note 136, Art. 77.

[365] *See* AI Act – IMCO-LIBE Draft Compromise Amendments, *supra* note 217, Art. 68a.

manufacturer or deployer, and that AILD relies on fault. A lot of the reasonable standard of behavior to be expected from parties, may again rely on what is defined in standards and obligations to conduct risk assessments, etc. and what is the industry practice (OpenAI, for example, does a lot of content moderation). A lot of this will rely on risk assessments and impact assessments.]

CONCLUSION

Adapting liability law to the distinct challenges posed by AI is an important element of ensuring that the vast and fast adoption of AI systems in all facets is done in a way that guarantees the protection of people's rights and interests, but also to provide legal certainty for AI developers and deployers.

Adapting liability to address the challenges of new technologies is, also, a good compliment to risk and safety regulation. Indeed, relying solely on risk regulation has distributive consequences, including the possibility that individual harms and costs will be dismissed if a particular measure makes sense collectively, which may especially harm minorities. When adopting risk regulation, quantifiable harms take precedence over unquantifiable or less-known harms.[366] Similarly, one of the main arguments that were raised when the AI Act was first published was that it didn't include individual rights nor rights of action for affected persons, even if its stated goal is to protect fundamental rights in Europe.[367]

The analysis of the AILD and the PLD reveals that, indeed, liability law can play a significant role in addressing AI harms. At the same time, it is interesting to note that the EU approach to AI liability is closely intertwined with risk regulation and it's tool kits - risk assessments, standard setting and impact assessments -, specially as these *ex-ante* regulatory interventions will often lead to the creation of the documentation, standards and information that will be important to successfully succeed in liability claims *ex post*.

The dyad of risk-regulation and AI liability (including the prohibition, as in the AI Act, of certain systems that pose unacceptable risk) may however still leave gaps, especially when harms are unquantifiable, as are many harms to fundamental rights. These gaps are, in many cases elsewhere already

[366] Kaminsky, *supra* note 11, at. 8.

[367] European Digital Rights et al., An EU Artificial Intelligence Act for Fundamental Rights: A Civil Society Statement, Nov. 31, 2021. Available at: https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf.

addressed by fundamental rights law but victims may face similar challenges in seeking redress as those faced by the victims of AI harms before AI rules are adjusted to account for the characteristics of AI systems. In those cases, and elsewhere, the European approach to AI liability offers important lessons of the procedural arrangements that may help redress AI harms.