24. Adding to the EU AI Liability Directive: degree of autonomy, chain of confidence, inherent flaws of indecent induction, and mandatory insurance

Ronald P. Loui[1]

I INTRODUCTION

Artificial Intelligence (AI) is a broad category and that category continues to expand, so one problem with commenting on AI liability is scope. The disruptions AI has brought sit atop software technology's own disruptions to liability law, and even those have arguably not yet fully settled. These present as much of a policy problem as a legal conundrum. But they present some opportunities, especially where law and legislation meet, for revisiting how societies should control technologies responsible for actual harm.

AI liability is a pressing problem: business innovation based on AI is both exhilarating and horrifying. Software control of systems is now pervasive, and the attempt to apply AI to improve the software ubiquitous. Meanwhile, new services and devices, especially at the consumer level, are invented quickly. AI hype, which has always been a tendency for AI purveyors and the public, complicates the legal landscape because it mixes fraud and misrepresentation, design flaws, implied warranty, and contributory negligence.

The recent EU European Directive on AI Liability (EU AILD) succeeds in tackling the broad scope; it offers focus for discussing what can be done. We may look at what it contains as first steps, and what it could have contained or could yet contain. This focus is helpful even if the AI Liability Directive is not passed, or is passed and is surpassed by other efforts in quick succession.

Briefly, the AI Liability Directive follows as byproduct of the EU AI Act, "the first-ever legal framework for AI, which addresses the risks of AI." It mainly divides AI systems into Unacceptable Risk, High Risk, Limited Risk, and Minimal Risk,[2] proposing "strict obliga-

[1] Thanks are owed to Marialuisa S. Gallozzi, Mary K. Engle, Palig Taslakian, and Connie Lopez, for helpful conversations. Also Quinton Zondervan, Vipin Chaudhary, David Selinger, and Rose Chan for retorts. And the editors for their quick turnaround. The errors are my own.

[2] The definition "(2) 'risk' means the combination of the probability of an occurrence of harm and the severity of that harm;" comes dangerously close to a suggestion of quantifiable probability and severity; usually in risk analysis, these are considered unquantifiable or hard to quantify. Elsewhere "mitigation" is discussed, which is more in line with unquantifiable risk analysis than an expected utility approach. For a deeper dive into this author's approach to risk analysis in contrast with expected utility theory, see my R. Loui, "Against Narrow Optimization and Short Horizons: An Argument-based, Path Planning, and Variable Multiattribute Model for Decision and Risk," *Journal of Logics*, 3(2), (2016).

tions" for the high-risk systems "before they can be put on the market."[3] The AI Liability Directive itself is aimed at making it easier to pursue liability claims by *presuming causality of the software system* rather than placing the burden of showing causality on the plaintiffs. This presumption of causality can be rebutted, but certainly makes civil claims easier to pursue when AI systems do harm.

This chapter gives a technologist's view on how the Directive appears, especially as someone who sees where abuses keep being committed during the software engineering of the product. This should be considered a response to a hypothetical request for general comment on proposed legislation.

The AI Act is complete, passed, and adopted at time of this writing, but the AI Liability Directive is still evolving. In particular, the definition of an AI system has changed, and this solves some problems but introduces others. The main point of the Directive is about presuming AI systems being causal; this is a potentially positive step since the software systems might be nontransparent, dynamic, weakly specified, nondeterministic, technically inscrutable, or trained by a third party as a foundation for further training. Hence, it is hard to show liability due to fault at a particular time in a given situation. There is much to say about causality, background assumptions, technological entrenchment, and legal standards here, even as the Directive seeks to moot that point, or at least to lower the evidentiary hurdle for victims. AI causality is worth consideration in terms of technological novelty, whether proximal or contributory or concurrent, whether counterfactual or probabilistic or actual.

Meanwhile, what is not contemplated by the Directive, which perhaps should be, includes suitability for marketed purpose (including using generative AI as if it were factual), especially for delegation of authority and how degrees of autonomy relate to proportionality of cause or fault. There are also inherent design flaws in machine learning classification (bias and false certainty, nonspecific testing). Finally, the end user has little control or knowledge of the system, by manufacturer's explicit and intentional product design which aims to usurp control and force user's abdication. Acts such as this should contemplate how responsibility can be placed on those who create the impression that the AI system could act with such independence, wide breadth of application, and such loose rein, i.e., what person contributed what claims of adequacy in a chain of confidence across the various deployers of the AI system.

This chapter will praise doctrines of *respondeat inferior* for engineers, and *piercing the corporate engineering veil*. These would be ways of addressing important gaps where there is currently insufficient upstream economic and legal disincentive for producing AI systems that may cause harm under fully autonomous use. Basically, those who hype or pass undue optimism through the product design should be held to account, at least proportionally. Mandatory insurance will be endorsed as a policy aim.

---

[3] European Commission, "AI Act," Shaping Europe's Digital Future, 2024. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

II THE EU AI ACT IN BRIEF

Definition of AI System, Providers, and Deployers

The scope of the EU AI Act is for providers and deployers, "manufacturers placing on the market or putting into service an AI system."[4] Crucially, it defines an "AI system":

> (1) "AI system" means a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[5]

This definition had been Annex I, which was deleted. Annex I had read:

> (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; (c) Statistical approaches, Bayesian estimation, search and optimization methods,[6]

which was referenced in an earlier attempt at defining an AI system:

> (1) "artificial intelligence system" (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.[7]

The earlier definition attempted to define an AI system by its software technology. It ranged broadly, including classic AI techniques such as rule-based expert systems, though the impetus for AI regulation appears to be the rise of neural nets for classification. Otherwise the push for legislation might have occurred in the earlier eras of AI using methods listed in (b) and (c). Even machine learning, listed in (a), supervised and unsupervised, are long-occurring topics in AI. The provocation appears to be the proliferation of highly autonomous systems, made autonomous because of the accuracy and effectiveness of recent deep neural net (DNN) techniques. The high-risk uses are enumerated by area of application: biometric categorization,

---

[4] European Parliament, Article 2, "EU Artificial Intelligence Act," 2024. https://artificialintelligenceact.eu/article/2/.

[5] European Parliament. Article 3, "EU Artificial Intelligence Act," 2024. https://artificialintelligenceact.eu/article/3/.

[6] European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," Brussels, 2021. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

[7] European Parliament. Title I, Article 3, "EU Artificial Intelligence Act," 2024. https://artificialintelligenceact.com/title-i/article-3/.

safety components of critical infrastructure, education, healthcare, financial services, law enforcement, and administration of justice.[8]

Issues with Neural Nets as Predictors: Indecent Induction

If it is the massive spread of black box neural nets, and putative success due to deep multi-layered architectures made possible by voluminous computation, that has spurred action, it is worth knowing that there are specific problems with systems built this way. It is the nature of the neural net that is vexing: To name a few issues widely known and relevant to ascription of fault:[9]

- Neural nets are based on massive training data which may be nontransparent and hard to characterize, poorly cleaned, unverified, possibly lost after training, and dependent on the multidimensional feature set or tokenization of the architect's and data engineers' choosing;
- Neural net training processes are inherently flawed with respect to subclass errors and bias (increasing training set size without increasing specificity and relevance just hides the small sample problem), and their adequacy for nonstable time series data (where the trend changes) depends crucially on the feature set;
- Neural nets can evolve their behaviors in the hands of users based on additional interaction with the environment;
- Neural nets have outputs that can be hard to explain: so hard in fact that explainable versions are current research;
- Neural nets can have nondeterministic and nonrepeatable behavior, especially when interacting with an environment where sensor inputs cannot be recreated exactly;
- Neural nets are trained and tested so as to produce an output network that is a local minimum under an error function; hence, this selection is literally random, even if it is random among a class of candidates that all score better than their explicitly considered alternatives;
- Neural nets make predictions but are themselves somewhat unpredictable, even if one cares mainly about predictions, not explanations of predictions; and
- Neural nets typically do not report the confidence of the result based in particular on relevant training set sample sizes.

I call this kind of neural net abuse "indecent induction" (a reference to David Hume's "scandal of induction."[10] I have private communication from a 2023 Nobel economist saying that the

---

[8] European Parliament, "Regulation (EU) 2024/… of the European Parliament and of the Council, laying down harmonised rules on artificial intelligence and amending Regulations (EC)," 2019–2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf, Annex III.

[9] See also Hutson reporting on Rahimi: Matthew Hutson, "AI Researchers Allege That Machine Learning Is Alchemy." *Science* 360(6388), (2018): 861 and the recent Kapoor-Narayanan: Sayash Kapoor and Arvind Narayanan, "Leakage and the Reproducibility Crisis in ML-Based Science." arXiv preprint arXiv:2207.07048 (2022). https://arxiv.org/abs/2207.07048.

[10] See for example Marc Lange, "Hume and the Problem of Induction." In Gabbay et al. (eds), *Handbook of the History of Logic*. Vol. 10. North-Holland, 2011, 43–91.

"neural net" name obscures the well-known problems of induction and regression, which we both observed as a weakness over 30 years ago during a seminar on predicting the stock market with neural nets.[11]) We will discuss this a bit more under inherent design flaws, where some of these flaws are shared beyond neural nets and DNNs, shared among most machine learning "gradient optimization" methods with training and testing sets. These points apply prior to even considering liability attached to neural net large language models (LLMs) that provide "generative" AI.

There are many other features of this learning-based (implicitly programmed from data) form of AI that distinguish it from logic-based or statistical reasoning. The latter are more familiar and cogent. For the bulk of machine learning methods that are not based on neural nets, there is a set of rules interpretable by humans, not a black box claimed to have excellent and widely applicable input-output behavior. Somewhat ironically, the neural net purports to do what the brain does, but it is the logical and traditionally statistical methods that are human-readable.

Human-Proxying and Autonomy

The definition of AI in terms of systems applied in a human-proxying way (i.e., what heretofore would have required a human for perception, decision-making, or judgment) is typically acceptable among specialists. It is actually insightful. This is indeed the crux of the problem: autonomy and delegation of authority. The definition of AI contains wording about "influence" on "environments."[12] This is a concern with cyber-physical systems. But the wording is inclusive: "can influence" not "must influence" phrasing.[13] The rise of AI in cyber-physical systems is correctly spurring regulation: This is software that interacts with the physical environment, such as AI in autopilots for planes, cars, drones, and nuclear plants, as opposed to stock markets, document search, language translation, and image generation. But the EU AI Act aims broadly and can include systems with autonomous operation going back decades in design and original deployment.

The revised definition of an AI system currently in the Directive's draft obviously makes possible inclusion of other, future AI technologies, and it depends on "varying levels of autonomy."[14] Especially where deployment of control systems is a concern, the proxy of human decision-making, the abdication of control, and the bypassing of the human mind deserve the attention. A definition in terms of autonomy does permit the inclusion of primitive autonomous decision-making devices and software: From thermostats to cruise control, to Boeing's infamously ill-fated 737 Max MCAS; they are included, even though those algorithms use extremely simple decision rules (they would not have satisfied the technical requirements listed in the original definition of an AI system). The criterion based on autonomous deployment is more about AI abuse than system architecture; it is the one who deploys the system for some purpose, who states that this is an acceptable risk, who replaces the humans, unleashes the machine on the world, who is likely the main candidate subject to liability claims.

---

[11] Philip H. Dybvig, personal email, 2023.

[12] European Parliament, *supra* note 7: "… generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with … ."

[13] European Parliament, *supra* note 8, Annex III.

[14] *Ibid.*

Other Parts of AI System Definition: Adaptivity, Environments, Objectives, Any Software

Note that the "adaptivity" of the system is envisioned but not required. This adaptivity turns out to be the source of one of the concerns regarding fault determination and proof of causality. It can be hard to find a snapshot of the system at the time it made the fateful decision presumably causing harm. Even if the adaptivity of the system is insignificant, the fact that the system is adaptive makes more remote any fault of the original designers; it's likely that the remoteness is used to shield upstream providers from causal proximity. It's not our fault; it's how the system adapted. On the other hand, where adaptivity is desired, too much ascription of responsibility to original programmers could hinder innovation and hamper legitimate service providers (e.g., based on inability to foresee consequences of unlimited adaptation; obviously a duty of care might include a limit on adaptation to a reasonable range). We will continue to discuss this under causality and proportionality.

The text in the definition of an AI system[15] about "Environments," being either "physical or virtual," being "influenced" as a possible but not necessary condition, adds little; "objectives" being "explicit or implicit" also adds little. They do show that activities are intended to be covered under the Directive with broad scope. The "objectives" of the system will be relevant when the discussion turns to fraudulent manufacturer claims, or simply discussing testing with respect to due diligence and contributory negligence. What the system is claimed to do, and why anyone thought that it deserves autonomy and independence, is a matter of representation.

Thus the definition of the Act as passed is that "AI system" can include any software system that is given some degree of autonomy. The Directive inherits this definition from the Act. AI regulation and liability depend on the use and abuse, not the specific wiring under the hood so to speak; however, existing software regulation frameworks can be brought into conflict when using such a sweeping definition.

There are obligations, prohibitions, and monitoring requirements set forth in the AI Act, but of possible interest specifically to liability exposure are the risk management norms described in Article 9 of Chapter 2: "Throughout the entire lifecycle of a high-risk AI system" there must be steps for the "identification and analysis of the known and reasonably foreseeable risks," "estimation and evaluation of the risks that may emerge," and "adoption of appropriate and targeted risk management measures designed to address the risks identified."[16]

The definition here is for steps, not specified, which appears to be a low bar. There is additional defeasance: "the risks … shall concern only those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system."[17] This mitigation through design is an appropriate characterization of risk analysis versus cost-benefit analysis. It does beg the question what reasonable mitigations might be when a technology such as DNNs becomes entrenched or widely accepted. This reappears in our discussion of causality.

---

[15] European Parliament, *supra* note 5: "… decisions that can influence physical or virtual environments."

[16] European Parliament. Article 9, "EU Artificial Intelligence Act," Requirements for High Risk AI Systems: Risk Management System https://artificialintelligenceact.eu/article/9/.

[17] *Ibid.*

Operators and Others Involved

Before proceeding, it is worth reviewing the EU AI Act's Title I Article 3 classification of *operators*:

> (2) "*provider*" means a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge;
> (4) "*user*" means any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity;
> (5) "*authorised representative*" means any natural or legal person established in the Union who has received a written mandate from a provider of an AI system to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation;
> (6) "*importer*" means any natural or legal person established in the Union that places on the market or puts into service an AI system that bears the name or trademark of a natural or legal person established outside the Union;
> (7) "*distributor*" means any natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market without affecting its properties;
> (8) "*operator*" means the provider, the user, the authorised representative, the importer and the distributor;[18]

I am mostly concerned with the variety of sub-roles under provider, and the various users making various choices and reasons for making those choices. Obviously liability can attach more broadly, to other operators, but for the engineer the proportionality depends on this chain:

- the nexus of product manager, algorithm decider and scientific officers, data engineers, safety and testing engineers; then
- the internal and external marketing of the system;
- the milieux in which it becomes acceptable to use such a system for a specific purpose, with various degrees of inspection, monitoring, human observation and intervention;
- the users or anyone in the hierarchy of command and control at the endpoint or edge of the service chain, which includes *deployers* (who might also be users, who have sole authority to deploy), who might instruct or command users to accept autonomous use of a system, i.e., who exercise authority over users and can transfer agency from user to system.

Envisioning the hand off of responsibility along this chain, two other definitions in the EU AI Act Title I Article 3 are worth knowing:

> (12) "*intended purpose*" means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation;
> (13) "*reasonably foreseeable misuse*" means the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems.[19]

---

[18] European Commission, *supra* note 6. Italics added.
[19] *Ibid.*

Here, I would note that the purpose includes the context and conditions, in particular, when can the system be presumed to act as intended sufficient for the purpose, under what range of conditions (e.g., In poor weather? For two decades? Even when sensors fail? Even when market volatility is high? Even when fooled by emergency vehicle strobe lights?); and the foreseeable misuse may relate to over-delegation of authority and autonomy to the AI system, not just using the system for a completely different purpose.

III THE AI LIABILITY DIRECTIVE (EU AILD) MAIN POINTS

Presumption of Causality

The Executive Summary describes the need for action:

> Current liability rules, in particular national rules based on fault, are not adapted to handle compensation claims for harm caused by AI-enabled products/services. Under such rules, victims need to prove a wrongful action/omission of a person that caused the damage. The specific characteristics of AI, including autonomy and opacity (the so-called "black box" effect), make it difficult or prohibitively expensive to identify the liable person and prove the requirements for a successful liability claim.[20]

And the proposed fix:

> The AI liability directive would create a rebuttable "presumption of causality", to ease the burden of proof for victims to establish damage caused by an AI system.

The key here is that the presumption is rebuttable, so defendants can show non-causality, at least in a legal sense e.g., with a duty of care defense; it is rebuttable at least for judgment in their particular case. The presumption makes it possible to claim causality without plaintiffs having to acquire access to the internals or past behaviors and hypothetical behaviors of the system.

So the presumption is flipped for AI systems. Is it also flipped for software systems generally? Just the inscrutable ones, or the ones that are self-modifying after interacting with the environment? Perhaps just for the systems that are given some degree of autonomy. The scope remains a problem.

Nevertheless, this does solve a massive problem with modern software systems. One cannot see the engineering. Even if one can acquire the source code of all software and firmware in the stack, the code may remain difficult to understand and analyze. Theoretical results in computation prove this in the abstract, but the practical reality is that software behaviors may be hard to understand and their actions hard to foresee even with a small amount of complexity. This is especially true with systems based on machine learning, where even simple algorithms would

---

[20] Commission Staff Working Document Executive Summary of the Impact Assessment Report, EU 9/28/22 https:// commission .europa .eu/ document/ download/ 80504486 -9244 -4542 -9a65-ea6cb9844353_en?filename=1_5_197611_summary_impact_asse_dir_ai_en.pdf. See also Tambiama Madiega, Briefing EU Legislation in Progress, Artificial Intelligence Liability Directive, Members' Research Service PE 739.342 February 2023. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

likely require all of the data for training and testing, the particularities of the training (e.g., any "boosting," "bootstrapping," and intentional biasing of data in training sets, how data sets were cleaned, how training was initialized, etc.), and a log of the data it may have encountered for dynamic adaptation, reconfiguration or "online" learning. But inscrutability happens even when there is transparency and logging, if only because systems are interdependent, and programs are embedded in a software "stack" making liberal use of software libraries that are themselves not fully understood. Even if the system's embedding in the cyber-physical environment as a human proxy decision-maker is fairly benign, such as a security camera triggered under some condition, those conditions might involve vision algorithms trained on hard-to-explain neural nets.

The issue is broadened beyond presumption, to include disclosure:

> [It would] give national courts the power to order disclosure of evidence about high-risk AI systems suspected of having caused damage.[21]

The motivating concerns are enumerated:

- [there is] legal uncertainty regarding how injured parties can be compensated for damage caused by software;
- new technologies introduce new risks, such as openness to data inputs that affect safety or cybersecurity risks;
- specific characteristics of AI (e.g. opacity/lack of transparency, explainability, autonomous behavior, continuous adaptation, limited predictability) make it particularly difficult to meet the burden of proof for a successful claim. … AI systems have characteristics that make it excessively difficult or even impossible for victims to identify and prove the fault of a potentially liable person or a defect and the causal link between that fault/defect and the damage suffered.[22]

This last point confirms the EU AI Act's focus on neural nets and machine learning; even the lowly but relatively autonomous thermostat is AI under the EU AI Act, but do not typically complicate the causal path by having "specific characteristics" that make it "excessively difficult" to prove fault. One of the "characteristics that make it … difficult … to identify … the fault,"[23] I suspect, is the vague notion of adequacy for autonomous operation, not actually the complexity of the engineered system. So one could focus on testing, warranty, and representations, not manufacturing defect or programming error.

There is additionally comment on a strict liability regime for high-risk autonomous AI systems, with the EU Parliament recommending mandatory insurance.

Who Gives This Autonomy?

The Act has a weakness defining the AI systems to which the flip of presumption of causality (even a presumption of fault) applies. What level of autonomy (with what opacity and muta-

---

21 *Ibid.*
22 *Ibid.*
23 *Ibid.*

bility) suffices to meet the threshold? There is no proposal here for a degree of complexity or a degree of autonomy leading to different burdens to show causality. It simply declares AI systems causally responsible, at least presumptively.

The AI definition in the proposed Liability Directive currently is "designed to operate with varying levels of autonomy and that may exhibit adaptiveness."[24] How much of the autonomy was designed, or intended by the purveyor, versus how much autonomy was given by the operator? There are actually two operators to distinguish: the provider/producer of the system, presumably the defendant, and the person or persons who allowed the autonomy of the system as it was operating, the deployer/user (although both could be defendants if a third party were the injured, e.g., a pedestrian by a fully self-driving automobile).

What should be the focus? The definition of AI relies on autonomy, so the potential defendants should be the ones who give the system autonomy. However, the complication is that the purveyor of the AI system makes claims about the acceptable autonomous operation of the system controlled by the AI. It is exactly what happens when "advanced cruised control" is marketed as "full self-driving." The driver permits the device to have autonomy in a specific environment at a specific time, which may have been poorly advised, but the manufacturer represented the software as adequate for that purpose. Joint and several liability is baked into the definition of AI systems here. It is particularly brutal to hold the user of an AI system negligent, when all the user did was believe what the provider claimed. In the case of MCAS for Boeing airplane accidents, it is mainly Boeing's fault for design defect, not the fault of the airline that did not disable (or did not realize additional training was needed), nor the fault of the pilots who were the proximate operators. There are other cases where the decision to yield decision-making is clearly joint. It seems that the manufacturer would only be fully shielded from responsibility if the warning was made, "do not give this system full autonomy" and the operator actually did. Meanwhile, the user would only be fully shielded from responsibility if the defense were "I did not turn this on" and the system turned itself on.

Cynical businesses even attempt to benefit from the ambiguity of who is responsible in this way. Perhaps they are subject to warranty and misrepresentation civil actions, but the human in the loop who turned on the switch will share the responsibility. Early self-driving exercises even placed a human monitor in the front passenger seat to provide yet another scapegoat were there to be harms incurred.[25] It must be largely the manufacturer's responsibility when AI claims are made: This is malfunction, not just failure to exercise a supervisory duty. Because once the AI system is engaged, the prospect of effective human supervision and monitoring, is typically small and on different time scales and requiring higher levels of abstraction. Controls may be few, short of switching the whole AI system off. The human in the loop is often barely in the loop when AI is doing its decision-making.

Even under a learned intermediary doctrine, such as shielding the pharmacist from the doctor's error administrating a drug, it would be fanciful to claim that users of AI devices are learned intermediaries, though they may be fully knowledgeable of the risks. AI systems are often useful precisely because their users are not learned, they are *unlearned intermediaries*,

---

[24] European Parliament, *supra* note 8, Annex III.

[25] See, e.g., Riess, Rebekah, "Uber Self-Driving Car Test Driver Pleads Guilty to Endangerment in Pedestrian Death Case." *CNN* July 29, 2023. http://cnn.com/2023/07/29/business/uber-self-driving-car-death-guilty/index.html.

cognitively uninvolved users, even going back to "expert systems" where the expert replaces the judgment of the user by design, providing a "virtual" or "remote" expert.[26]

Degrees of Autonomy

As autonomy is a crucial aspect of AI liability, and the determination of fault will often be contributory, I propose a second dimension to the EU AILD that explicitly takes into account degree and kind of autonomy of the AI system. This would produce a two-dimensional table (Table 24.1).

Table 24.1 EU AI Act degrees of risk augmented with degrees of autonomy

Risk	Fully autonomous	Partly autonomous	Human directed
Unacceptable risk	NOT OK	?	MAYBE OK
High risk	NOT OK	MAYBE OK	OK
Limited risk	MAYBE OK	OK	OK
Minimal risk	OK	OK	OK

Notes: More columns could be added.

Obviously the degree of autonomy columns could be refined, just as the degree of risk rows could be; this is intended to be an easy visualization of the idea for clarity. There may be important ways in which partial autonomy may fall short of full autonomy: kinds of controls, rate of supervision, information available for intervention, opportunity for intervention, kind of assistance, etc.

---

[26] Barton writes "As product manufacturers increasingly act as intermediaries between the end users' data and the product, new duties could emerge." Jonathan T. Barton, "Introduction to AI and IoT Issues in Product Liability Litigation." (2019). https://www.stantonbarton.com/cmss_files/attachmentlibrary/WLJ_MED2526_Barton.pdf. Wein considers "absurd" the "notion that mechanical intermediaries might function as responsible 'agents' for their proprietors. The extent to which courts will continue to treat ever more autonomous and accomplished mechanical stand-ins as agents remains to be seen." Hence, *respondeat superior* applies in force where the AI system is considered the subordinate. Leon E. Wein, "Responsibility of Intelligent Artifacts: Toward an Automation Jurisprudence," *Harv. JL & Tech.* 6 (1992): 103. Harned et al.: If "a diagnostic system is designed to take the scan, read it, make the diagnosis, and then present it to the physician who acts merely as a messenger between the system and the patient, then it would seem that the physician is playing a relatively passive role in this provision of treatment. If a faulty diagnosis is made resulting in patient harm, and the patient was not adequately warned regarding this risk, the physician's relatively passive role could end up insulating her from liability, because it could eliminate the manufacturer's ability to utilize the learned intermediary defense. Manufacturers might therefore be more likely to design their medical machine vision systems to actively engage the physician not only because they believe it will increase the likelihood of physician and hospital adoption, but also because it may provide the manufacturer shelter from liability by enabling them to use the learned intermediary defense." Zach Harned, Matthew P. Lungren, and Pranav Rajpurkar, "Machine Vision, Medical AI, and Malpractice," Comment, Machine Vision, Medical AI, and Malpractice, *Harv. JL & Tech.* (2019).

The SAE (Society of Automotive Engineers) levels of autonomous driving are a closely related idea, and these are shown in the list below.

Note that most of these SAE levels have to do with kinds of controls and opportunity for intervention, not degree of human attention or concern, or degree to which the human is informed and knowledgeable. Delegation can take many forms; all delegation incurs liability risks for AI systems, and in proportion to the degree of delegation. Regardless of the risk of the system as deployed and with what purpose, the degree of involvement from alerting and cautionary advice, to full usurpation, must inform degree of liability.

The EU AI Act foresees this dimension in text such as:

> (d) the extent to which the AI system acts autonomously and the possibility for a human to override a decision or recommendations that may lead to potential harm [shall be taken into account] [when assessing high-risk].

The EU AILD should incorporate some of this same thinking.[27]

Levels of autonomy for driving, according to SAE.[28]

- Level 0: no driving automation
  - (automatic emergency braking, blind spot or lane departure warning)
- Level 1: driver assistance
  - (lane centering or adaptive cruise control)
- Level 2: partial driving automation
  - (all of the above simultaneously)
- Level 3: conditional driving automation
  - (traffic jam chauffeur)
- Level 4: high driving automation
  - (local driverless taxi, pedals or steering wheel not installed)
- Level 5: full driving automation
  - (all of the above, everywhere in all conditions)

---

[27] Lior quotes multiple sources. Anat Lior, "AI entities as AI agents: Artificial Intelligence Liability and the AI *respondeat superior* Analogy." *Mitchell Hamline L. Rev.* 46 (2019): 1043: "the AI entities' varying degrees of autonomy, and the absence of complete human control with regards to the potential behavior of AI entities lead to difficulty in establishing a legal nexus of causation between the victim and the tortfeasor as well as a difficulty in reasoning about causation in fact between the damage inflicted and the liable party. This in turn hampers the attribution of legal responsibility to a liable party, including the manufacturer of the AI entity within the product analogy." His note 86, referencing Matthew U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harv. J.L. & Tech.* 29 (2016): 353, 363; Jos Lehmann et al., "Causation in AI and the Law," *Artificial Intelligence & L.* 12 (2004): 279.

[28] Society of Automotive Engineers (SAE), "SAE Levels of Driving Automation™ Refined for Clarity and International Audience," 2021. https://www.sae.org/blog/sae-j3016-update.

IV THREE ISSUES: CAUSALITY, DELEGATION/ASSURANCE CHAINS, ENTRENCHMENT

Legal vs. Probabilistic Causality

Although the Directive attempts to dispose of causality by removing it as an impediment for plaintiff success, legal AI causality yet raises problems of causal claims generally.

Stephenson-Harwood legal analyst Simon Bollans is helpful here:[29]

> [EU AILD lowers] the evidentiary hurdles for victims injured by AI-related products or services and making it easier for victims to successfully establish claims against AI operators, providers or users;

and introduces:

> a presumption of causation between the defendant's fault and the damage caused to a claimant by the AI system. This presumption would apply if … it can be considered reasonably likely, based on the circumstances of the case, that the fault influenced the output produced by the AI system, or the AI system's failure to produce an output; and the claimant has shown that the output of the AI system, or the AI system's failure to produce an output, gave rise to the damage.[30]

It is AI scholarship itself that has shed light on probabilistic causality: The two recent Turing Awards to AI researchers are for neural nets, and before that, for Bayesian network analysis of causality. In the latter work, the attempt to decide which factors are probabilistically causal is not usually successful; it depends on the way that a probability distribution is broken into paths of uncorrelated influence, or equivalently, how conditional independence cascades from causes to effects. The objective in this conditional independence scholarship is to define hidden causes and screening variables, which are important in scientific and medical claims, but perhaps not well represented in legal disputes.[31]

This has not prevented legal systems from relying on causation when assigning responsibility or determining proportional damages. The counterfactual analysis, "but for" the plaintiff or defendant doing (or not doing) some action rightly ascertains involvement, but hardly proportionality. One could imagine counterfactuals that lead to probability claims, such as "but for

---

[29] Simon Bollans, "EU Artificial Intelligence Liability Directive," Stephenson-Harwood Legal, 2023. https://www.shlegal.com/insights/eu-artificial-intelligence-liability-directive.

[30] *Ibid.* This motivation for reversal of presumption is discussed in Bathaee: "This form of causation, which includes doctrines such as reliance and the causation element of Article III standing, is often predicated on the assumption that one can determine whether a defendant's conduct is connected to the alleged harm. When AI is a strong black box, this sort of inquiry is nearly impossible to undertake. … [E]xamination of a version of the AI as it existed at a particular point in time – a snapshot – may be the only way to find facts that can satisfy the required evidentiary burden." Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harv. JL & Tech.* 31 (2017): 889.

[31] See e.g. discussion of Pearl's probabilistic networks. Christopher Hitchcock, "Probabilistic Causation," *The Stanford Encyclopedia of Philosophy*, E. N. Zalta ed., Metaphysics Research Lab, Stanford University, 2021. https:// plato .stanford .edu/ archives/ spr2021/ entries/ causation -probabilistic/.

A's doing X, the probability of F would have remained low." Even qualitative, as opposed to quantitative claims of probabilistic influence would be hard to determine and defend.[32]

Proportionality

One worry about the EU AILD specifically creating a one-sided presumption of causality is that it might make proportionality ascertainment more difficult. In many cases, negligence is joint and several; responsibility is shared among parties. All operators in an AI chain that successively delegates authority and autonomy are likely partly responsible, even if the system itself can be presumed causally effective, even presumed to be the most proximal. The evidentiary burden of a "but for" claim might be reduced, so that AI black boxes need not be understood mechanically, but all operators from providers to deployers to users, are potentially responsible for giving that system autonomy, and in different proportion. This is actually much clearer in non-physical AI; the cyber-physical systems are entangled with the physics of momentum. "But for" the AI system recommending buying a falling stock, money would not have been lost. There is not much causal intricacy here, no physical mechanism requiring causal explanation; what matters is how the AI system got its way, how it came to be persuasive of the equities trader, not how the badly timed purchase produced fiscal harm.[33]

Delegating Authority, Delegating Autonomy, and a Chain of Assurance

So how does the chain of delegating authority work in the case of an AI system? The key here is the same insight that made the EU AI Act change its definition of an AI system: *It is the autonomy, not the engineering, that is new with respect to responsibility under the law.* The operator of a standard vehicle is perfectly responsible for control, but the automated vehicle driver is a complication: not just because the engineering might be complex (it might

---

[32] See e.g. Abbott-Sarch; though there the discussion is specifically about intention, the complications are the same: "One conceivable way to argue that an AI (say, an autonomous vehicle) had the intention (purpose) to cause an outcome (to harm a pedestrian) would be to ask whether the AI was guiding its behavior so as to make this outcome more likely (relative to its background probability of occurring). Is the AI monitoring conditions around it to identify ways to make this outcome more likely? Is the AI then disposed to make these behavioral adjustments to make the outcome more likely (either as a goal in itself or as a means to accomplishing another goal)? If so, then the AI plausibly may be said to have the purpose of causing that outcome. Carrying out this sort of inquiry will of course require extensive and technically challenging expert testimony regarding the nature of the programming – and could thus be prohibitively difficult or expensive." Ryan Abbott and Alexander Sarch, "Punishing Artificial Intelligence: Legal Fiction or Science Fiction," *International Conference on Autonomous Systems and the Law*. Cham: Springer International Publishing, 2022.

[33] In Giuffrida et al.: "the main risk of the increasing reliance on AI, and, therefore, the most difficult obstacle for regulators, does not reside in the technology itself, but rather in the interaction between AI-enabled devices." I would agree with the first part but not the second, preferring to say the interaction between those who can pass authority and autonomy along a chain. Iria Giuffrida, Fredric Lederer, and Nicolas Vermerys, "A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law," Case *W. Res. L. Rev*. 68 (2018): 747.

be very simple), but because the control was delegated. So why is full self-driving different from cruise control over the past decades, if not in the engineering complexity? (Ironically an early misnomer: autonomy only recently arriving in the automobile, even one with automatic transmission.) The list below shows how different chains might be relevant to the liability assignment.

Three different chains relevant to finding upstream liable actors: causal, delegation of authority, and a chain of assurance or persuasion that the system is adequate, i.e., a chain of representations. EU AILD addresses those marked with *.[34]

- CAUSAL CHAIN TO HARM: Background Conditions + Background Engineering → Manufacturing/Development* → Deployment/Environment* → Fact Circumstance of the Case* → Probability/Chance* → Particular Harm*
- AUTHORITY DELEGATION CHAIN TO HARM: Social/Human Hierarchy → Upper/Middle Management → Lower Management → Human Actor → AI Software/Device → Specific Algorithm* → Decision in Particular* → Probability/Chance* → Particular Harm*
- AUTONOMY PERSUASION/ASSURANCE/CONFIDENCE CHAIN TO HARM: AI Paradigm Leaders + Hype/Optimism Contributors → Top-Down Product Design for Purpose → Testing with Limited Scope→ Business Model Requiring/Desiring Autonomy → Users Permitting Autonomous Operation → AI Independence* → Fact Circumstance/Decision* → Probability/Chance* → Particular Harm*

The hand-off from manual or (partly/largely) assisted-manual control to automated control happens for many reasons: All the reasons why AI systems are desirable. Frequently, the machine can do things better, faster, more attentively, almost all of the time. But how do we know this as a user? Ideally, the automation is simple enough to understand. Neural nets trained with undisclosed data and mutable in a dynamic environment take the new automation far away from that ideal. In lieu of such an ideal simplicity and observable adequacy, one would want massive specific real world testing, with full, transparent, accurate, and relevant reports of experience: the testimony of product history. This is hard to imagine, especially given the aggressive speed with which AI systems are appearing. *Caveat emptor* is getting harder to practice. The actual problem we see now are users who are incapable of verifying providers' claims with respect to performance, safety, controllability, situational robustness, etc., where users have already been in the dark with respect to aspects such as longevity, environmental impact, foreign labor sourcing, and others. Not just the users, but tracing back through the product deployment and envisionment, the adequacy for autonomy is hard to adjudge, or at least the range of adequacy.[35]

---

[34] Note that delegation of autonomy and delegation of authority are closely related, but delegation of authority can be proper even if it is unadvised, while the other can be improper delegation of authority, even if it is a good idea.

[35] Gerstner in a very early article calls it "blind reliance," discussing mass-marketed versus R&D systems, even when they provide "great potential for substantial harm." In this context, he is concerned with strict liability, not delegation, "Professional services do not ordinarily lend themselves to the doctrine of tort liability without fault because they lack the elements which gave rise to the doctrine. There is no mass production of goods. This language emphasizes the importance of mass production in applying strict liability. While it is possible to argue that even one-of-a-kind

Representation of Adequacy: Assurance and Entrenchment

The novel liability issue is not therefore actual or probable cause, but misrepresentation and fraud, likely unintentional at all links in the chain. These are warranty issues, not just design defect problems; the issue is how unjustified or embellished, or merely hopeful representations create contributory negligence. The question is not whose agency was at fault, but whose information was bad. Responsibility for misrepresentation can be found at many levels, from engineering paradigm adherents and admirers, to design, development, and data engineering, to marketing and product management, to bloggers and influencers who swear to the efficacy in their experience (truthfully or falsely, widely ranging or narrowly). We consider in the next section how this chain of persuasion can be addressed in an act such as the EU AILD.[36]

There is a commonsense concept of cause: If *A* and *B* jointly cause *C*, in terms of actual circumstances that gave rise to *C*, then *A* might be the cause if *B* is presumed, and *B* might be the cause if *A* is presumed. This happens regardless of the magnitude of probabilistic influence of *A* upon *C* versus *B* upon *C*. This means that if some aspect of the engineering is taken for granted, it is not an easy candidate for causal attribution. For example, if one wants to focus on a novel, non-entrenched engineering mechanism such as four-wheel steering or automatic throttle-reduction, one can freely speculate whether it was causal. However, an entrenched engineering mechanism such as rubber tires or round steering wheels is not impeachable.

Even if the risks were foreseeable *ab initio*, at some point society moves past that decision and questions the decisions that are taken with those risks assumed, and engineering developing from that starting point. Even automatic wipers, automatic transmissions, and anti-lock brake systems are imperfect forms of automation, but are so assumed that causal analysis looks elsewhere. The system could be flawed, outdated, or have its many detractors (all engineering approaches have their critics), but it is part of the background after a few decades of widespread use. In the case of neural net AI or machine learning, how long will it take for the

---

items such as custom-designed software should be held to a strict liability standard, there is little precedent for this application for either conventional or expert system software. Although it may be easy to argue that a particular program, *e.g.* one controlling a nuclear power plant or an air traffic control system, is inherently 'unreasonably dangerous' when used as intended and thus should be subject to strict liability, a better standard is that each program must be considered on a case by case basis." Marguerite E. Gerstner, "Liability Issues with Artificial Intelligence Software," *Santa Clara L. Rev.* 33 (1993): 239.

[36] Villasenor sees but does not delve into the issue, believing it is fully covered by existing UCC: "Another variant occurs when the manufacturer of the AI-based MRI reading system informs customers that, over time, the algorithm will not only learn from its own accumulated experience in analyzing images, but … its accuracy improves less quickly than it otherwise would have. This might lead to a products liability claim asserting that the manufacturer engaged in misrepresentation and the product contains a manufacturing defect. To the extent that a manufacturer makes assurances in the marketing and sale of the MRI reading system that turn out not to true, any resulting harms could give rise to a breach of warranty claim. This would be handled under well-established approaches in accordance with the Uniform Commercial Code, which addresses the explicit and implicit warranties that are created through the sale of goods." J. Villasenor, "Products Liability Law as a Way to Address AI Harms," Brookings Institution. United States of America, 2019. Retrieved from https://policycommons.net/artifacts/4141490/products-liability-law-as-a-way-to-address-ai-harms/4950556/ on June 1, 2024. CID: 20.500.12592/ts6b43.

devices to gain such acceptance that they are no longer candidates for legal causal influence, hence legal causal liability?

Given the rate of technological adoption without the buffer of experience and reflection, machine learning systems can become matter of fact, like mobile phones and electronic signatures, before legislative acts have had a chance to shape them.[37]

V INHERENT DESIGN FLAWS OF MACHINE LEARNING

Flaws of Neural Nets as Design Flaws

I earlier listed a few known and lesser known problems with the machine learning approach to building automatic classification and control systems. To make further the point that the technology deserves to be suspect, consider two easily understood design flaws in more detail. All systems have pros and cons, but these technical observations are a good prelude to the discussion of misrepresentation throughout the product chain: How a paradigm shift that appears magical can in fact be a constant upstream liability due to hidden flaws.

Using a neural net as a classifier (on this input, decide it is a member of a particular class, which includes binary yes-no classification) appears to solve an age-old inductive problem of how to fit multidimensional data to sample data, like fitting lines to points, or planes in hyperspace; the computation discovers automatically which dimensions and in which combinations there are good evidence-based connections. One might call it automated pre-theoretic scientific discovery. To someone who was not already engaged in multivariate nonlinear regression, it's magical; to those who were so engaged, it borders on fraud. Two glaring fundamental and simple flaws are not talked about, that are visible to those who have done a lot of nonlinear regression.

Time Series Variation

The easiest seen inherent flaw worth mentioning specifically here is time series variation and insufficient expressiveness in the tagging of training data. Briefly, if there is a trend over time so that training data is out of date when a system is in use, machine learning by its nature is inherently flawed. There are ways to use the neural net classifier as a time series predictor, but this requires that the data be tagged with time information, and that the specificity of the time suffices for seeing the time-varying nature. This is especially important for cyclical data. The problem also is that frequently the developers, programmers, data engineers, and product managers do not realize that there is a time component that should be taken into account. The neural net can automatically find the pattern, but only if the features describing data points

---

[37] Karnow perhaps is relevant here: "Yet where responsibility and thus liability are so spread out over numerous and distributed actors, the very sense of 'responsibility' and 'liability' is diminished. At that point legal causation seems to fade into the background and tort law, founded on the element of causation, falls away. Assigning legal liability involves discrimination among an infinite number of causal candidates. That discrimination is avowedly based on perceptions of policy, society's collective sense of what is reasonable and who should be blamed for certain injuries." Curtis E. A. Karnow, "Liability for Distributed Artificial Intelligences," *Berkeley Tech. LJ* 11 (1996): 147.

express the pattern well enough to expose it in the first place. If the time-sensitive pattern is monthly but only years are recorded in the data, the cycle is not knowable to the AI system (insufficient granularity in description). Even when trends are spotted in data sets, these might be spurious or irrelevant, e.g., related to a change in how data are harvested, such as newer photos of Labradoodles having more pixels, or more visual recognition data collected under a California sun.

Note that selecting features insufficient to expose time-dependent variations is a design flaw that can be corrected with better descriptions. Relying on machine learning when data are not known to be sufficiently described, that is when pattern recognition is limited but assumed good enough, is an inherent design flaw. Putting an inductive decision-maker in place of a human decision-maker most of the time will be an improvement; but sometimes humans can do good things. The human presumably would have some ability to monitor and recognize insufficient data, inadequate description, or ignorance, and seek competent statistical advice.[38]

Sample Sizes of Relevant Subclasses

The next inherent flaw worth mentioning in detail is the small sample sizes of relevant subclasses. This manifests as bias and overconfidence, both of which are hidden by nondeterministic selection of the predictor network and large testing sets containing largely irrelevant data. It's easy to see what this means nontechnically as follows.

When a neural network is trained, it typically marginally perturbs connection weights over a long process; this process tries to get the predictions of every data set example to match the desired output: features on the input, prediction on the output, and for the training set, the "correct" predictions are known. This process inevitably results in a "local maximum" fit, which is like ascending to the top of a hill, among the many hills that might have been ascended. There are ways to escape the local hill and move to a higher one, but the fact is that the network is not optimal, just selected well relative to others considered. Thus, it is "nondeterministic," meaning that had the training gone differently, started in a different place, or made different random marginal perturbations, a different network would have resulted.

When the quality of fit is scored, the number of training examples that the network correctly outputs is the figure of merit. The more it gets right, the better the network.

Now consider the problem of projecting from a class of similar prior examples, where the sample from that class is very small, say three data points. If all three agree that the output

---

[38] Time series prediction is on of the earliest uses of neural nets, e.g. William Remus and Marcus O'Connor, "Neural Networks for Time-Series Forecasting." In J. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Springer, 2001: 245–56; I. Kaastra and M. Boyd, "Designing a Neural Network for Forecasting Financial and Economic Time Series," *Neurocomputing* 10 (2016): 215–36. It remains a topic of study e.g., Dawei Cheng et al., "Financial Time Series Forecasting with Multi-Modality Graph Neural Network," *Pattern Recognition* 121 (2022): 108218; Zihao Zhang, Bryan Lim, and Stefan Zohren, "Deep Learning for Market by Order Data," *Applied Mathematical Finance* 28(1) (2021): 79–95: "Although a large number of deep learning models have been developed for time-series forecasting, some limitations still exist. Firstly, deep neural networks typically require time series to be discretized at regular intervals, making it difficult to forecast datasets where observations can be missing or arrive at random intervals."

should be the same, perhaps it will find such a network that reproduces the shared value as output. In that case, the network has no training error on those examples, though it might be hastily generalizing, or over-generalizing, because three is not a large sample. It may be that ignoring some features permits a larger sample to bear on each prediction, which is not the problem. In the case where two outputs should be one value, and the third should be another, according to the training set, the network can "memorize" what features the first two examples share in their input features, which the third example does not. This is how it creates an exceptional subclass and "over-fits." It will get all three predictions to accord with its training set. But it is guilty of projecting from too small a sample size.[39] It is in essence producing a prediction from almost zero knowledge, which is the holy grail of induction, and statistically unadvisable. A more traditional statistician would report a *p*-value, a degree of confidence, a sigma value for standard deviation, an "*n*" for the relevant reference class sample size, or some other representation of the nonrobustness or uncertainty of prediction.[40]

Randomness and Brittleness

What the neural network does during training, typically, is make a random choice among local optima and fail to report the quality or variation among those choices; also it makes a decision during prediction, typically, that does not include a report of the brittleness or uncertainty of that decision. This is the "magic" of neural nets as classifiers, essentially random selection hidden under the optimality of fit, as it relates to the idea of "over-fit" or alternatively, generalization not well supported by the data. This haunts all of the machine learning methods that are based on training, testing, and fixed features used to describe each example in a data set. It's just not talked about in this way. The modelers are happy because the network scores well under validation, but in practice, the neural network hallucinates (a different kind of hallucination from what happens with generative GPT LLMs such as ChatGPT3.5, but the same word works well here).[41]

---

[39] Liu-Wei et al.: "[W]hen facing high dimension, low sample size (HDLSS) data, such as the phenotype prediction problem using genetic data in bioinformatics, DNN suffers from overfitting and high-variance gradients. … Deep Neural Pursuit (DNP) … selects a subset of high dimensional features for the alleviation of overfitting and takes the average over multiple dropouts to calculate gradients with low variance. As the first DNN method applied on the HDLSS data, DNP enjoys the advantages of … the robustness to high dimensionality, the capability of learning from a small number of samples, the stability in feature selection, and the end-to-end training." Bo Liu et al., "Deep Neural Networks for High Dimension, Low Sample Size Data," IJCAI International Joint Conference on AI 2017.

[40] See also Kyburg on interval-valued representations of probability, e.g., Henry E. Kyburg, "Levi, Petersen, and Direct Inference," *Philosophy of Science* 50(4) (1983): 630–4.

[41] In the same way, "maximum entropy" methods produced a probability of exactly. 500,000,000 … from queries where sample sizes of 0 bore on the prediction. This is a production of certainty from ignorance. It can be justified in some situations where theoretical symmetries exist, such as physics or chemistry state spaces, but not in a projection from data. Nguyen-Yosinski-Clune discuss the overconfidence in some overprojections: "[I]t is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion)." Anh Nguyen, Jason Yosinski, and Jeff Clune, "Deep Neural Networks Are Easily Fooled: High Confidence

Buchholz is instructive here (even if I am critical and he is supportive of Deep Neural Nets' apparent ability to avoid overfitting, based on a few recent studies, which is itself a small sample projection) in three separate passages:

> [C]hoosing the right reference class for making a prediction reduces to identifying the model with the right set of variables. With respect to the reference class, the criteria of narrowness, reliability, and homogeneity are constitutive for what is "right". With respect to the set of variables, the model should be selected such that it avoids under- and overfitting … As mentioned above, the latter is closely related to a model's complexity and thus, given the model's overall structure (i.e., a linear function, a specific architecture, etc.), to the number of variables it contains: the model should include enough variables to avoid underfitting; yet it should also contain only relevant variables to avoid overfitting. Consequently, when framing the [Reference Class Problem (RCP)] as a problem of statistical model selection, there is a close connection between the goal of avoiding under- and overfitting and the goal of choosing a reference class that is as narrow as possible while also being homogeneous.
>
> [I]n the context of prediction, it is not the goal to investigate reference classes themselves or the predicates by which they are determined. Instead, the goal is to identify those features that determine a reference class for making accurate predictions. Thus, the criterion of predictive relevance alone is discerning suitable from unsuitable features in this context. Whether or not the features and the reference class they determine have an immediately obvious meaning is less important.
>
> Maximizing accuracy alone is therefore insufficient, because apart from all relevant features, the most accurate model might also include irrelevant features. Furthermore, maximizing accuracy alone involves the risk of overfitting.[42]

This sometimes is reported as "bias," though there are many ways that bias can creep into predictors and many kinds of bias. Some forms of bias can be addressed with better data practices, but the relevant sample size problem and accompanying overconfidence in small sample predictions is inherent in the approach. Neural nets do pay attention to larger samples from more inclusive, less specific sets, defined with fewer features, where the larger relevant examples bear on the prediction. Hence, another possibility is that training ignores the specific features that are in disagreement and finds other features on which to differentiate, properly generalizing conservatively, producing the desired output with appropriate confidence (this is a legitimate magic of computationally intensive search through combinations of variations).

What one should worry about besides the predictions being reported with undue confidence and certainty, is how testing further hides this nondeterminism. The defense of the approach is often to add more data to the training set. Thus, even if the one or two errors are still in the mix, because inherently there should not have been a confident prediction based on the over-fit, the score that the neural net receives from verifying its behavior on the training set will rise; it will rise because irrelevant other predictions are added into the mix. The correct idea is to add more data that are relevant data, sample data relevant to the specific queries, to enlarge the sample sizes of the input combinations in question. But the testing regime is focused on individuals with unique features in a large set, not on interesting projectible subsets within the data bouillabaisse: adding vegetables might add to the overall goodness, but does not necessarily make the clams taste better.

Predictions for Unrecognizable Images," Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[42] Oliver Buchholz, "The Deep Neural Network Approach to the Reference Class Problem," *Synthese* 201(3) (2023): 111.

In fact, the best way to hide flaws in small sample subclass projection would be to randomize the test set as a selection from the training set. Thus, the network can be "verified" to a high accuracy and "test predict" to a high accuracy while still missing badly on specific small samples of importance.

Ensembles and Subclass Testing, Certifications

Fortunately there is an implication for duty of care and diligent testing at the development stage: score the neural net for fitness with an emphasis on subclasses that are likely to be queried. Then test for robustness of the predictions on those feature sets as more examples are added with those specific combinations of features. The beauty of the neural net approach to induction is that one need not know in advance which combinations of features will appear in queries, and which features are most relevant, hence likely to require specific attention. This is what gives it power and allows it to be overly certain about its outputs. The beauty of the law is that certain demographic classes are protected, so this kind of bias can potentially be ferreted out by requiring testing for specific queries, in the same way that neural nets and insurance indicators are tested for nondiscrimination.[43] This gives an approach to what reasonable due diligence might be expected. Another approach is to train an "ensemble" of predictors,[44] then report the outputs, agreeing or disagreeing, among the ensemble.

---

[43] Tartaglione-Grangetto: "How to guarantee from a technical point of view that an ANN model is respecting those ethical guidelines remains an open question. In this work we focus on plugging ethical constraints in the learning process, *e.g.* to avoid gender and race biases that have been found in commercial face recognition tools." Enzo Tartaglione and Marco Grangetto, "A Non-Discriminatory Approach to Ethical Deep Learning," 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE, 2020. Raff-Sylvester: "There are many reasons for desiring fair classification algorithms. These include legal mandates to be non-discriminative, ensuring a moral or ethical goal, or for use as evidence in legal proceedings. Despite the longstanding need and interest in this problem, there are few methods available today for training fair networks. When we say that a network is fair, we mean fair with respect to a protected attribute …, such as age or gender. Our desire is that a model's predicted label … given a feature vector … is invariant to changes …. An initial reaction may be to simply remove … from the feature vector … . While intuitive, this does not remove the correlations … that exist in the data, and so the result will still produce a biased model." Edward Raff and Jared Sylvester, "Gradient Reversal Against Discrimination: A Fair Neural Network Learning Approach," 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018.

[44] Inoue considers ensembles of predictors, rather than single predictors, to gain control of degrees of uncertainty in prediction: "Trained neural networks (or any classifiers in general) often make a prediction for a sample near the decision boundary of class A and class B with a low probability; the classifier assigns similar probabilities for both classes, and a class with a higher probability (but with a small margin) is selected as the result. For such samples, ensembling works efficiently by reducing the effects of random perturbations. While ensembling works near a decision boundary that is properly learned by a classifier, some decision boundaries can be totally missed by a trained classifier due to insufficient expressiveness in the model that is used or a lack of appropriate training data … . Typically, such mispredictions cannot be avoided by ensembling predictions from another classifier trained with different random numbers since these mispredictions are caused by the poor expressiveness of the model rather than the perturbations that come from random numbers." Hiroshi Inoue, "Adaptive Ensemble Prediction for Deep Neural

It might appear human-like for AI to act capriciously and decide on less data than prudent in specific circumstances; this may appear spontaneous, hence, intelligently ampliative. To err, after all seems human. But this overconfidence may also be unjustified inference, leading to action that creates a liability nightmare. When these design flaws are understood more fully, heads will roll. The problem of hallucination was immediately understood for generative AI because untruths that make false reference are easier to see than poor predictions.

All of the known flaws of neural nets are relevant to a claim of design flaw in products that rely on neural nets to achieve their AI. An easy example is sensitivity to one mislabeled training input. Some are acknowledged more openly than others, and some are more fundamental, not easily mitigated simply by requiring diligence with data cleaning, and what passes as "testing" and "validation" in the neural net community. Knowing the indecency of the inductive method, the brittle form of testing and validation, I would expect designs to limit the autonomy of AI or at least include ensembles of predictors and out-of-range safeguards, estimates of predictive accuracy, visualization of relevant training data from which the current prediction is projected, visualization of the main features upon which prediction is based, and self-monitoring of all kinds; otherwise, these are flaws, and promulgating the product as adequate for unspecified situations (even under well-specified purpose, but not well-circumscribed situation) looks like fraud.

AI systems could be certified for specific advanced testing, such as testing for extrapolation as opposed to interpolation. In that case, the outliers in each dimension would be withheld from the training set, and used as the test set. A similar certification for validity with recent data could be given, where data that are recent are withheld from the training, and used for the testing. Disclosure of ensemble agreement, or disagreement, would be another kind of explicit assurance, and testing on specific subclasses. Perhaps the careful engineer should check for robustness with respect to model complexity, by adding nodes in successive steps, and noting whether each step largely agreed with the former. Some of these tests do imply a much greater computation, where computation is already expensive, but on the other hand, those who do such testing should be rewarded with labeling attesting to it.

VI HOLDING ENGINEERING ACCOUNTABLE THROUGHOUT THE ENTERPRISE

Upstream Influence

Probably anyone interested in the EU AILD appreciates that with AI products there are a lot of claims about what a product can do, and less disclosure about what it cannot do. One point of the last section's detail is to suggest that the "watering down the concrete" starts far upstream, in the machine learning and neural net paradigm itself, not just in the software and product development teams that envision and embody what is eventually put to market. No doubt it is easy to blame the C-suite of the corporation, especially the marketing. There has always been a lot of hype under the AI marketing bridge.

Networks Based on Confidence Level," 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019.

Vicarious liability, or *respondeat superior* doctrine, receives a lot of attention in the legal commentary on AI (see, e.g., this section's notes). Scholars mainly suggest that this is how an automaton (a "slave") could be deemed an agent of a kind, yet exposes the producers and providers of the AI, or the users (the "masters") to responsibility for harm. This analysis is appropriate.

However, the indemnification of those engineers who produced the systems may be inappropriate, if shielded by corporate officers, for example, Chief Scientific Officer and Chief Marketing Officer. This is because in the ecosystem that produces claims of a system's adequacy as an autonomous actor, many can share blame. There are actually two chains of responsibility reassignment here (in addition to any causal chains familiar to technology-based product liability): First, delegation of authority; second, a chain of persuasion, confidence, and assurance.

The first is familiar to all, and even should remind EU members of Nuremburg history.

The second is about the development of a nontransparent, not-easily-measured artifact that is said to be fit to the purpose of doing a task with a large degree of autonomy.

Who says this product suffices for the task or can be trusted with independence and reduced human involvement, interaction, input, and surveillance? The company, and its public marketing statements say so. Who informs those public statements, since the marketing arm cannot test for robustness and safety in all intended situations? Like any advanced technology, the product managers: those who can test and interrogate. Who informs that level of management? The algorithms and data people. Because of the nature of the "testing" and the faith in the paradigm, they are too easily trusted. It's as if the effectiveness of airplanes were not known in many weather situations such as extreme heat, extreme altitude, leading edges of tornadoes, extreme turbulence, but the aircraft product were pushed by all involved as a miraculous solution for all of the user's flight conditions.

This chain of attesting adequacy can be traced back to its sources, with contributions along the way, when considering how persons responsible for harm, jointly and several, can be identified.

Mechanistic Causal Chain versus Chain of Unchecked Optimism

The main thrust of this chapter in relation to the EU AILD is that the mechanistic causal chain leading to harm is not the only chain to look at; the chains of delegation of authority and chains of certification for autonomous operation are equally important. This delegation happens because users, reasonably or not (under the best of circumstances making a cost-benefit analysis or even a risk-mitigation investment), believe that the system can perform with a degree of autonomy as advertised. It's not so much that the risk was foreseeable, or even that there was a breached warranty of fitness, but that the user was persuaded to abdicate control in some situation, to let go of the steering wheel, to weigh the AI system's medical diagnosis heavily, to go to sleep assuming the phone's alarm would work the next morning. Someone, or some chain of unchecked optimism, is responsible for the persuasion, which is too often an unjustified representation.[45]

---

[45] Claussén-Karlsson refers to a "chain from production to use": here will almost certainly be situations where all humans in the chain from production to use of the AI have done everything

Pity the poor engineer who is told to build the product that the managers have in mind, top-down, and to use the DNN technology that appears to be the key to creating autonomous system. "Find the data to make it happen" is the specification. When the deadlines approach, the pressure on the AI engineer to certify adequacy for a level of autonomy, which is hard to express in terms of novel situations, but easy to express in terms of verification of training (i.e., few errors compared to training examples, and on a randomly drawn testing set, enough data to be confident about behavioral range), is enormous. This happens in any engineering product development, but in most of those situations where harm could be produced, society requires a licensed professional engineer to sign off at crucial stages. There has been no equivalent in software, and certainly no regulation of data engineers and neural net testers, whose art is equally dark. As the AI industry is developing, the data engineers and neural net python programmers are somewhat disposable and exchangeable, relatively weakly accredited; one would not expect longevity of employment that puts the same people in the line of fire when tort actions begin to arise. The engineer who signs off on the autonomy adequacy of the system will be long gone. This is all the more reason to have a doctrine of *respondeat superior*, but if there were *respondeat inferior*, diligence would be incentivized throughout the chain.[46]

"right," yet the AI learns wrong nonetheless, and to a "chain of production": "Concerning narrow AIs, harm may occur either because of a fault in the chain of production or by default." Matilde Claussén-Karlsson, "Artificial Intelligence and the External Element of the Crime an Analysis of the Liability Problem," Juridicum Thesis 2017, Orebro Universitet. Osmani considers "whether responsibility can be allocated to different agents in the production chain, *i.e.* imposing strict liability on operators of the Al agents, such as programmers, designers, owners, and other parties involved in the process of manufacturing," but in that article the focus is on programmers and foreseeability. Nora Osmani, "The Complexity of Criminal Liability of AI Systems," *Masaryk University Journal of Law and Technology* 14(1) (2020): 53–82.

[46] Johnson-Verdicchio are more helpful and I quote these passages at length for their relevance, even though the VW emissions case was intentional fraud, not perhaps a failure of duty of care: "Top management delegated this goal to the advanced AI. As with the artifact and the designer, the user is part of the agency that produces the emission fraud. However, top management is different in that, unlike the artifact and the designer, its members have intentions; they have intentional agency. Because of this, top management would be held responsible, and what they would be held responsible for would be dependent on the nature of their intentionality. That is, depending on the details of the case, they might be considered ethically reprehensible, negligent, strictly liable or to have committed some other complex infraction. As the only entity in the triad with intentional agency, was top management completely responsible for the fraud? Because of its reliance on the advanced AI, instead of jumping to such an ascription of responsibility, the relationship between top management and the advanced AI needs to be explored further. To start, we can ask: what was top management thinking when it delegated the achievement of its goal to the advanced AI? Answering this question brings into focus new relevant entities and a new triad: the designers of the advanced AI, and the triad for the creation of the advanced AI. Suppose that top management hired a team of software engineers to create an advanced AI for car design and manufacture. In the triad for the creation of the advanced AI, top management is again the user, human software engineers are the designers, and any number of futuristic hardware and software tools would be the artifacts that facilitate the production of the advanced AI. Top management still plays the role of the user, this time with the goal of creating an advanced AI that could solve the problem of passing the EPA test. In this triad, triadic agency (the combination of user, designer, and artifact) produces an advanced AI capable of solving the problem of the EPA test." "Top management claims that the

Part of this is the nature of AI software design: they are not built to specification top-down, but programmed bottom-up; behaviors that the system exhibits are generalized in the imagination of the user. It's not like knowing a bridge with a certain material and fit will bear so much stress for so long, based on experience and equations. It's very much: This training set represents all the situations we can include, and we hope that will cover enough of the situations that the product will encounter. It's really more like saying that this vitamin supplement worked well for hundreds of people in the study, but with not much known about biological pathway or pharmaceutical mechanism, we think it will work for you too, the importance of personalized medicine notwithstanding.[47]

Including Low-Level Engineers: Piercing the Engineering Veil

Perhaps there are liability limits or levels of indemnification that would protect low-level engineers from massive class action damages. The point is to put some of their skin in the responsibility game, legally and economically. Because in products where complexity and opacity prevent testing at each hand-off, to make sure that the job was well done, there must be some penalty for doing the work poorly. Or equivalently here, there must be some penalty for misrepresenting what the AI can and cannot do. The misrepresentation is passed up the chain through management, to markets, aided by bloggers influenced by paradigm followers behaving like cultists, through to users who really should be more skeptical of advanced automation on all but the lowest risk tasks. In manufacturing fault, we see Boeing airplane bolts not installed; AI products have this problem at the programming and training stage: handing off to the hopeful product managers, who hand off to the persuasive marketers, who then take a chain of misrepresentation onto a chain of autonomy delegation.[48]

---

decision to use the defeat device was made by the engineers once they realized that the engines on which they were working would never meet the EPA standards without significant improvement (*i.e.*, investments by the company). Allegedly, not wanting to be bearers of bad news to their higher-ups, the engineers handled the problem on their own, keeping the engines as they were, but adding the defeat device (Smith and Parloff 2016). On this account, top management may not have had the intention to break the law, though the engineers did." "Top management acknowledges that they specified both goals on which the engineers acted (to achieve a particular level of performance for the car and to meet the EPA standards). Intentionally setting these goals and intentionally creating a corporate culture in which engineers feared the consequences of failure (and did not want to tell top management that these goals could not be met) can be seen as setting off the sequence of events that led to the fraud. The point is that the issue of responsibility depends not just on causal sequences but on intentions and intentionality." Deborah G. Johnson and Mario Verdicchio, "AI, Agency And Responsibility: The VW Fraud Case and Beyond," *AI & Society* 34 (2019): 639–47.

[47] One of the persons acknowledged in conversation with the author successfully reined in the excessive on-air claims of Dr. Mehemt Oz while an FTC investigator.

[48] Villasenor, *supra* note 36: "Blaming the upstream or downstream supply chain: As occurs with non-AI products, products liability in relation to AI will often involve finger pointing at other places in the supply chain. In the AI ecosystem, there will typically be multiple suppliers upstream of a consumer. To start with, there is the company that sold the product directly to the consumer. That company in turn may have purchased a software component of the product from a separate entity. And that entity may have built some portions of the software in-house and licensed other portions from yet another company. Apportioning blame within the supply chain will involve not

Respondeat Inferior

Respondeat inferior, holding the employees accountable not just the employers, is thus advisable. This term appears in the legal scholarship literature in a few places because *respondeat superior* is so doctrinal.[49] Piercing the corporate veil might also be a useful concept: not just to expose shareholders, or venture capitalists, but piercing the corporate engineering veil as well.

Manda is helpful here:

> This approach of holding a subsidiary responsible for intentional violations of the law that are conducted "on behalf of" (but not "at the direction of") a parent is described in other areas of the law as *respondeat inferior* liability. *Respondeat superior* liability is a risk-shifting exception to the general rule that entities are separate. The principle imputes liability to the parent for a subsidiary's or employee's acts where the parent knew or should have known of the existence of a violation but took no proactive steps to prevent it from continuing. But *respondeat superior* liability "can never be predicated solely upon the fact of a parent corporation's ownership of a controlling interest in the shares of its subsidiary." At the very least, there must be direct intervention by the parent in the management

---

only technical analysis regarding the sources of various aspects of the AI algorithm, but also the legal agreements among the companies involved, including any associated indemnification agreements." In Rachum-Twaig, under a design defect trigger, "a manufacturer will be liable for harms caused by a product 'when the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission, of the alternative design renders the product not reasonably safe.' There are generally two approaches for determining whether a design defect has occurred. … [T]he plaintiff must prove that an alternative design could have reduced the risk imposed by the product using preventive measures that are reasonable in relation to the harm. This is basically a cost-benefit analysis of alternative designs which imports a notion of negligence to the otherwise strict liability regime of products liability. Note, however, that the risks and harms considered under this approach are only foreseeable risks and harms, and that the reasonableness of the design should not be considered only with respect to a specific harm done, but rather with respect to the product's safety at large. A second approach to the design defect trigger is the 'consumer expectations' approach, which basically asks whether the dangers entailed by a product exceed those reasonably expected by the potential consumers." Omri Rachum-Twaig, "Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots," *U. Ill. L. Rev.* (2020): 1141.

[49] Abbott-Sarch, *supra* note 32: "The culpable mental states of AI developers, owners, or users could be imputed to the AI under certain circumstances pursuant to a *respondeat superior* theory." "[F]or cases of Hard AI Crime that is not straightforwardly reduced to human conduct – particularly where the harm is unforeseeable to designers and there is no upstream human conduct that is seriously unreasonable to be found – nothing like *respondeat superior* would be appropriate. Some other approach to AI *mens rea* would be required." Also Čerka et al.: "In accordance with the provision that the AI system is a tool, the parallel with a party's responsibility for the behaviour of children or employees can be used. The following relationships are the best examples of vicarious liability: (a) liability of the principal for the act of his agent or liability of the parents for their child; (c) liability of the master for the act of his servant (example discussed above). This means that for AI's behaviour vicarious liability appears to a person on whose behalf it acts or at whose disposal and supervision AI is. It can be listed as users of AI or their owners." Paulius Čerka, Jurgita Grigienė, and Gintarė Sirbikytė, "Liability for Damages Caused by Artificial Intelligence," *Computer Law & Security Review* 31(3) (2015): 376–89.

> of the subsidiary to such an extent that "the subsidiary's paraphernalia of incorporation, directors, and officers are completely ignored." That is, there must be more than just *the capability of control, even if it was never utilized* before an issuer is required to pay fines and penalties beyond those that its subsidiary pays for the violations the latter commits. In contrast, a *respondeat inferior* standard balances the benefits of limited liability with those of … purpose, i.e. putting an end to corruption by focusing prosecution and deterrence efforts on the source of the evils that are sought to be eliminated.
>
> [R]espondeat inferior, is not a liability-shifting mechanism like its mirror opposite, *respondeat superior*. Rather, *respondeat inferior* is a find-the-blame-in-the-right-place mechanism that allows for liability to attach at the source of the violation. …, *respondeat inferior* subsidiary liability solves the injustice of parent issuers being called upon to assume liability for actions they can only control in hindsight and them then passing blame to *subordinate 'wayward' employees* who act as scapegoats.[50]

VII MANDATORY AI INSURANCE

Future Strict Liability

The EU AILD has the following longer term goal:

> Parliament stressed that, while high-risk AI systems should fall under strict liability laws (combined with mandatory insurance cover), any other activities, devices or processes driven by AI systems that cause harm or damage should remain subject to fault-based liability. The affected person would benefit from a presumption of fault on the part of the operator, unless the latter is able to prove that it has abided by its duty of care.[51]
>
> The AI liability directive therefore proposes to leave the door open for future legislative development. In particular, that review should examine whether there is a need to create no-fault liability rules for claims against the operator combined with a mandatory insurance for the operation of certain AI systems, as suggested by the European Parliament resolution of 20 October 2020 on a civil liability regime for artificial intelligence … .[52]

Regulation Too Slow

This author agrees with the aim that AI systems be subject to a defeasibly no-fault mandatory insurance requirement. Regulation, certification, licensing, credentialing are all good ideas but the criteria will be too slow to evolve, perhaps also not sufficiently specific per industry given the scale of legislative burden. The solution in the case of automobile accidents was to move to no-fault mandatory insurance. As many have pointed out, this allows for innovation, internalized incentive, specificity, and immediate protection of those harmed. We understand how to regulate insurers, and insurers understand how to regulate risk.

---

[50] Peter Manda, "Bringing Fairness to FCPA Settlements: Protecting the Corporate Form Through *Respondeat Inferior* Subsidiary Liability," *Int'l. In-House Counsel J.* 8 (2014): 1 (emphases original).

[51] Tambiama Madiega, Briefing EU Legislation in Progress, Artificial intelligence liability directive, Members' Research Service PE 739.342 – February 2023 https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

[52] *Ibid.*

In the context of medical AI systems, Stern et al. say:

> Well-designed AI liability insurance can mitigate predictable liability risks and uncertainties in a way that is aligned with the interests of health care's main stakeholders, including patients, physicians, and health care organization leadership. A market for AI insurance will encourage the use of high-quality AI, because insurers will be most keen to underwrite those products that are demonstrably safe and effective. As such, well-designed AI insurance products are likely to reduce the uncertainty associated with liability risk for both manufacturers – including developers of software as a medical device – and clinician users and thereby increase innovation, competition, adoption, and trust in beneficial technological advances.[53]

The problem for insuring where there is fast innovation is that a market of insurance may be slow to develop and rate, and this may hinder technological progress. However, governments can provide insurance or backstop insurers' losses if the interest is that compelling. Would taxpayers contribute to the indemnification of Roomba vacuum cleaner providers? No, but perhaps to AI-assisted medical diagnostics, or even airplane autopilots, alerting and advising, because of the obvious society-wide desirability.

Zech is succinct: "In scenarios where a clear allocation of risk control is no longer possible, social insurance might step in."[54]

Given that the EU can immediately create a mandatory insurance market, at least partially covering losses and partially imposing strict no-fault liability, I am surprised that this is not a part of the current proposal.

---

[53] Ariel Dora Stern et al., "AI Insurance: How Liability Insurance Can Drive the Responsible Adoption of Artificial Intelligence in Health Care," *NEJM Catalyst Innovations in Care Delivery* 3(4) (2022): CAT-21.

[54] Zech: "In order to legally cope with the risks associated with artificial intelligence or current digital systems, the introduction of strict liability rules is being discussed, especially for high-risk AI systems. Strict liability not only influences the level of care but also the activity level by fully internalising economic risks of AI, thereby activating private risk knowledge. It also incentivises the further development of existing technologies and, arguably, helps public acceptance. Strict liability may also be used as an instrument for risk-distribution, especially when combined with compulsory liability insurance (third party insurance). However, like any liability rule, it only works when a proof of individual causation is possible. Additional influence on the level of activity: Strict liability assigns the economic risk to the injurer regardless of whether the injurer behaves in accordance with existing duties or not, *i.e.*, it internalises the risk completely (at least to the extent that the damage is compensable). The risk controller must therefore consider whether the expected benefit of an activity exceeds its risk. If it is not worthwhile, the risky activity will not be carried out. By delegating the assessment to the technology developers and users (*i.e.* the manufacturers and operators), private risk knowledge is made available." Herbert Zech, "Liability for AI: Public Policy Considerations," *ERA forum*, 22(1) (2021): 147–58. Maliha et al.: "[S]ome automobile insurers have already sponsored data-gathering efforts for new AI technologies such as autonomous-vehicle-guidance software. Insurers could reward users with lower rates for selecting certain more-effective AI programs, just as insurers already reward drivers for selecting safer cars and avoiding accidents. Thus, insurers would facilitate AI adoption through two methods: 1) blunting liability costs by spreading the risk across all policyholders, and 2) developing best practices for companies looking to use AI." George Maliha, Sara Gerke, Ravi B. Parikh, and I. Glenn Cohen, "To Spur Growth in AI, We Need a New Approach to Legal Liability," *Harvard Business Review* (2021).

The fact that EU has brought attention to AI regulation and has such a specific proposal for liability is laudable. It remains to be seen how chilling the effect on AI products and services will be simply by reversing the causality presumption (defeasibly).

The EU AILD also contains language, not discussed here, concerning corporate transparency in litigation of high-risk system failures. As we have seen elsewhere, the threat of transparency itself is often sufficient incentive to exercise more care.[55]

VIII CONCLUSION

The EU AI Liability Directive in its current form inherits an insightful definition of an AI system from the now passed EU AI Act, basing the definition on autonomy. The degree of autonomy might interact with degree of risk, and the neural net predictors on which many new AI systems depend might be inherently flawed. The flip of presumption of causality raises a few questions about technological entrenchment, proportionality, and proximity. This is because jointly negligent actors may be behind the marketing and found throughout the engineering and product design, as confidence in the system is built, but cannot easily be tested further along the chain. The Directive therefore misses an opportunity to consider how confidence in the adequacy of the system in an autonomous setting is passed through a chain of designers, developers, and deployers, ultimately leaving the user at the mercy of potential misrepresentation. The author shares the Directive's objective of moving quickly to a mandatory insurance regime for addressing harms of AI systems.

[55] A former student, Quinton Zondervan, counts as his great legislative achievement the allowance of self-driving cars on the streets of Cambridge, MA, in accordance with state law, while adding the possibility of documents disclosure when the activity falls short on safety and performance promises.