THE EXPLANATIONS ONE NEEDS FOR THE EXPLANATIONS ONE GIVES—THE NECESSITY OF EXPLAINABLE AI (XAI) FOR CAUSAL EXPLANATIONS OF AI-RELATED HARM: DECONSTRUCTING THE 'REFUGE OF IGNORANCE' IN THE EU'S AI LIABILITY REGULATION

Ljupcho Grozdanovski*

Abstract: This paper examines how explanations related to the adverse outcomes of Artificial Intelligence (AI) contribute to the development of causal evidentiary explanations in disputes surrounding AI liability. The study employs a dual approach: first, it analyzes the emerging global caselaw in the field of AI liability, seeking to discern prevailing trends regarding the evidence and explanations considered essential for the fair resolution of disputes. Against the backdrop of those trends, the paper evaluates the upcoming legislation in the European Union (EU) concerning AI liability, namely the AI Liability Directive (AILD) and Revised Product Liability Directive (R-PLD). The objective is to ascertain whether the systems of evidence and procedural rights outlined in this legislation, particularly the right to request the disclosure of evidence, enable litigants to adequately understand the causality underlying AI-related harms. Moreover, the paper seeks o determine if litigants can effectively express their views before dispute-resolution authorities based on that understanding. An examination of the AILD and R-PLD reveals that their evidence systems primarily support *ad hoc* explanations, allowing litigants and courts to assess the extent of the defendants' compliance with the standards enshrined in regulatory instruments, such a the AI Act. However, the paper contends that, beyond *ad hoc* explanations, achievin fair resolution in AI liability disputes necessitates *post-hoc* explanations. These shoul be directed at unveiling the functionalities of AI systems and the rationale behind harmful automated decisions. The paper thus suggests that 'full' explainable AI (XAI) that is, both *ad hoc* and *post hoc*, is necessary so that the constitutional requirements associated with the right to a fair trial (access to courts, equality of arms, contradictory debate) can be effectively met.

Keywords: AI; Causation; Explainability; Fair Trial; Procedural Fairness; Equality of Arms; Effective Participation; AI liability; Product Liability; AI Act; AI Liability Directive; Product Liability Directive

* National Foundation for Scientific Research (FNRS); Faculty of Law, Political Science and Criminology, University of Liège, Belgium.

Table of Contents

INTRODUCTION

A. The Limits of Causal Knowledge and the Refuge of Ignorance Metaphor

In his *Ethics*,[1] 17-century philosopher Spinoza discussed what he termed the 'reduction to ignorance' method, citing an incident of an unfortunate passerby fatally struck by a stone dislodged from a roof. To causally explain the bad timing of the fall, God-fearing dogmatics would, no doubt, ask an endless string of 'why-s': "perhaps you will reply that it happened because the wind blew and the person was walking along that way. But they will press: why did the wind blow at that time? Why was the person going that way at that very time? (…) And so on and so on, and they will not stop asking for causes of causes until you take refuge in the will of God, which is the *refuge of ignorance*."[2]

Our ambition is not to explore the depths of Spinoza's philosophy, but to draw attention to his stance when discussing the construction of knowledge: one would spare oneself from knowing 'proper' if they relied on the *belief* that all worldly occurrences had, as *causa prima*, a metaphysical, omniscient designer of reality. Even pious jusnaturalists like Grotius and Pufendorf hypothesized that if God did not exist (as *the* authority decreeing oughts and ought-nots), Nature would continue to function according to its inherent rationality.[3] Although Spinoza's philosophy is deist - his concept of 'God' coinciding with that of 'Nature' (*Deus sive Natura*)[4] - his work is reflective of the 17-century rationalist rebellion against naïve religiosity, aiming to uncover the dividing line between (true) knowledge and non-knowledge.

---

[1] Baruch Spinoza, *Ethics. Proved in a Geometrical Order*, (ed. by Matthew J. Ksiner, CUP, 2018).

[2] *Id*., at 37 (emphasis added).

[3] Grotius, arguably, pioneered the hypothesis that moral normativity is irrespective of religious affiliation, going counter the Medieval *zeitgeist* according to which, moral normativity was divinely ordained, as opposed to derived from - because inherent to - Man's (rational) nature. Pufendorf later espoused the same view. See, namely, T.J. Hochstrasser, *Natural Law Theories in the Early Enlightenment* (CUP, 2000) at 84: "Pufendorf was entirely correct to identify Grotius and Hobbes as his crucial predecessors, since both had forced their opponents to fight them on new ground of their own choosing: Grotius by insisting that the source of natural law must be located in a principle to which all nations could assent irrespective of religious affiliation; and Hobbes, by his contention that the individual is capable of creating his own moral world from his personal psychological calculations."

[4] Summarizing Spinoza's philosophy is not our point of focus here. May it suffice stressing that he synonymizes God and Nature, asserting that from the infinite attributes of God (Nature), only two are knowable to us: thought and space. All of what is knowable can be understood as a particular expression either of those attributes. On the issue of gaining knowledge of the *essence* of knowable objects, Winch gives an excellent and pedagogical account of Spinoza's epistemology: "Spinoza distinguishes between '*essentia formalis*' and '*essentia objectiva*' (…) the sense of 'objective' doesn't at all lie in a contrast with 'subjective'; it highlights the relation of an idea to its object, to what it asserts or represents to be the case. The 'formal essence' on the other hand is, as it were, the idea as a distinct mental existent, considered in abstraction from its relation to an object." See Peter Winch, *Spinoza on Ethics and Understanding* (CUP, 2020), at 6. Spinoza, much like other philosophers such as Descartes or Kant, tackled the issue of 'knowledge' and 'representation' of reality, the former being traditionally thought to be 'objective' while the latter 'subjective'. The interrelationship between the two, as analyzed in Spinoza's philosophy, will not be further discussed here. However, this is a useful point to keep in mind as we explore the construction of knowledge *tout court* and of causal knowledge because we find, in the backdrop of the relevant theories, the objective/subjective dilemma which has indeed 'tainted' millennia-long traditions of erudite philosophical thought.

Is believing antinomic to knowing? For early-day rationalists, the answer would likely be 'yes.' Modern-day epistemologists are not as quick to dissociate the two, namely because our ability to know is limited. When we are called to causally explain portions of reality that are, to some extent, unknowable to us (e.g. why did a stone mysteriously fall off a roof?), there will invariably come a point where the explanation we give is based, not on 'what *we know to be true'* but on 'what *we believe to be true*.' In many ways, scientific communities today play a role similar to that of religious institutions in Spinoza's time: they nurture normative belief systems that serve as benchmarks for distinguishing valid, trustworthy information from 'false' counterparts. Since science operates largely without relying on faith, the convictions comprising the body of scientific knowledge, including those related to causality, are embraced only when verifiable and verified. Merely asserting claims without substantiation is typically insufficient for justified rational acceptance.

While modern epistemology has eased its skepticism toward beliefs, it has not yet resolved its inner conflict of striving for absolute certainty or truth, alongside the necessity to make internal epistemic compromises in determining what might qualify as *acceptable* knowledge. The pursuit of perfect, permanent, universal, agnostic, and context-independent knowledge alas remains practically unattainable. This - in many ways tragic - realization is at the core of Spinoza's refuge-of-ignorance metaphor: as we endeavor to understand the world causally, we are driven by an ideal (of absolute truth) while being entangled in the constraints of reality (where our capacity to know is limited). The million-dollar question is then: '*how do we decide what is true, if the attainment of perfect knowledge of causation is impossible*?' Probabilists suggested the notion of *necessity*: there comes a point where, by virtue of experience, we detect repetitive, regular associations which we taxonomize as reliable or stable causal phenomena (in the sense of 'X necessarily causes Y'). In their highest expression, these infallible causalities are labelled as (natural) *laws* or normative, 'universal causal regularities.'[5] However, to further our investigation of necessity in connection to causality, it is essential to bring forth one of the 18th century luminaries: David Hume.

B. The Concept of Necessity in Causation

In his Treaties of Human Nature,[6] Hume wrote:

"Probability… must in some respects be founded on the impressions of our memory and senses, and in some respects on our ideas. Were there no mixture of any impression in our probable reasonings, the conclusion wou'd be entirely chimerical: And were there no mixture of ideas, the action of the mind, in observing the relation, wou'd, properly speaking, be sensation, not reasoning.... The only connexion or relation of objects, which can lead us beyond the immediate impressions of our memory and senses, is that of cause and effect. ... The idea of cause and effect is deriv'd from experience, which informs us, that such particular objects, in all past instances, have been constantly conjoined with each other: And as an object similar to one of these is suppos'd to be immediately present in its impression, we thence presume on the existence of one similar to its usual attendant. According to this account of things, ... probability is founded on the presumption of a resemblance betwixt those objects, of which we have had experience, and those of which we have had none; and therefore 't

[5] Max Kistler, *Causation and the Laws of Nature* (Routledge, 2006), at 77.
[6] David Hume, *A Treatise of Human Nature* (ed. by L. A. Selby-Bigge, Clarendon Press, 1888).

is impossible this presumption can arise from probability. The same principle cannot be both cause and effect of another."[7]

An idea that transpires from the cited gloss is Hume's *assumption of uniformity* of Nature. The repetitiveness of observable events (say, rain does not fall when the sky is clear) justifies the associative reasoning Hume referred to. Our past experiences are the cognitive benchmark against which we interpret and explain any new experience. This type of reasoning can be explained by our all-too-human need to somehow make the new familiar. Repetitive events are ultimately what allows us to make causal generalizations which become our *nomological interpretations of reality*:[8] if the weather is cloudy, we may expect rain, snow or nothing at all, but we can be sure not to expect sunshine. The cause/effect link between 'clouds' and 'no sun' enters our arsenal of so-called *background knowledge*, which we mobilize whenever we encounter causal interrelationships we experience as novel.

His brilliance and insight notwithstanding, Hume's Achilles' heel is precisely his assumption that Nature is casually regular. Based on experience, clear skies consistently indicate the absence of rainfall and this we take to be a 'universal given,' a sort of intuitive law by virtue of which rain is generally not expected on a sunny day.

Reality is of course 'messier'[9] than our perceptions thereof, as modern scholarship pointed out in its critique of Hume's work. Kistler e.g. criticized Hume's disregard of *exceptional situations i.e.* cases where real-world occurrences deviate from what we view as nomological causations (*i.e*. causations characterized by a level of predictability).[10] Quantum physics is frequently referenced as an instance of epistemic departure from Newtonian physics: at the sub-atomic level, the behavior of particles appears to deviate from the laws governing supra-atomic behavior.[11] In the context of these 'exceptional situations,' Hume seems to have also omitted *accidental causation i.e.* cause/effect links that we explain in reference to so-called universal laws of Nature. Here again, Kistler cautioned against 'universalizing' the truth of causal phenomena that are due to coincidence[12] and not some unwavering, universal law of Nature.

With Kistler's criticism in mind, it follows that in a perfectly ordained, predictable world, events would, indeed, be causally linked by *necessity*: specific causes would *reliably* yield specific effects and only those. Of course - and again - arriving at a stable universal causal knowledge is a tricky business, for the reasons Kistler outlined in his excellent study.[13]

---

[7] *Id*., at 89-90 *cit. in* Henry W. Johnstone Jr., "Hume's Arguments Concerning Causal Necessity" 16-3 *Philosophy and Phenomenological Research* (1956), 331-340, at 337.

[8] For Kistler, 'nomologcial' is understood as 'normative' within the meaning of the laws of Nature. In the context of causality, we will use 'nomological' to refer to normative representations of necessary cause-effect interrelationships. See Max Kistler, *Causation and the Laws of Nature*, *cit. supra*, at 5.

[9] We paraphrase F.H. Bradley, "Epistemology Legalized: Or, Truth, Justice, and the American Way" *in* Susan Haack, Evidence Matters (CUP, 2014), 27 at 30.

[10] Max Kistler, *Causation and Laws of Nature, cit. supra,* at 76.

[11] For a comprehensive analysis of Newtonian mechanics and Quantum mechanics, see Albrecht Lindner, Dieter Strauch, A Complete Couse on Theoretical Physics. From Clasical Mechanics to Advanced Quantum Statistics (Springer, 2018), at 69 seq. and 275 seq.

[12] Max Kistler, *Causation and Laws of Nature*, *cit. supra*, at 75.

[13] *Id*.

Nevertheless, there is some virtue in epistemic and cognitive stability. Be it in the discovery of causation in science or in law, we cannot consider that all causal relations are a matter of chance. Generalizations about the world (such as clouds usually, though not always mean 'rain') are necessary for our every-day decisions and predictions. Hume's philosophy may be flawed, but it expressed the right intuition: *we need* to consider some causal interrelationships as true. The alternative - a perpetual state of uncertainty and doubt - would simply be untenable. As a result, we choose to assign truth values selectively to specific representations of causality (like 'dark clouds *ergo* rain').

Epistemologists have heavily reflected on the concepts of truth and falsity in causal contexts. Special focus has been placed on the *conditions* under which we decide to designate something as true. This point will be discussed further[14] as we explore the interrelationship between experience, belief and knowledge in explaining causal phenomena. At this stage, we shall stress two points which will frame our further discussion. First, Hume's concept of *causal necessity*, though debatable, has shaped the ways in which we approach knowledge of causation in both 'hard' science and law. Indeed, we often construct such knowledge in terms of necessity (as in 'X necessarily causes Y') because our aim is to ultimately distinguish *correlation* from *causation:* an event can be correlated (positively associated) to several other events but it will be *causally linked* to only one or a few of them. Causes are, in essence, *conditions that appear to be necessary* for specific effects to occur. In law, it is this Humean understanding of 'cause' that underlies the but-for test, which we will discuss further in this paper.[15]

Second, AI poses a challenge to our Humean (and human) inclination to *a priori* perceive reality as relatively stable. To begin with, AI systems exhibit a profound departure from Humean principles, since they are not natural entities, subject to the governance of natural causality. Put differently, we cannot resort to the laws of physics to, say, uncover the origins of algorithmic biases. If AI systems operate outside the jurisdiction of physical laws (as far as causality is concerned) they - intelligent as they are - are, in principle, governed by the laws of (human) reason. In this regard, AI systems align with Humean principles because their decisions and predictions result from associations between existing knowledge (represented by sets of training data) and new information. Just as humans explain new experiences by drawing connections to familiar ones, AI systems create associations between variables in a new situation (unseen during training) and the variable connections already established in the training data. However, the 'laws of reason' do not work as predictably as the laws of Nature, which is inconvenient when we are asked to causally explain the real-world consequences of AI. We thus find ourselves in a conundrum: we are and will increasingly be pushed to causally explain AI 'behavior' without any real possibility of mapping out, if not the 'laws' at least some *consistent trends* regarding the effects that behavior might cause. We know that recruitment AI systems *can be* discriminatory, but they can also be perfectly skill-based…

In dealing with such unpredictability, European and global regulatory reactions were in a manner of speaking, Humean that is, *stability seeking* (as will be argued). They chose to view the uniqueness and novelty of AI rationality through the lens of

[14] See *infra*, Sub-Section 2.2.
[15] See *infra*, Sub-Section 2.2.2.

agency, the referent for stability here being the role of *human* agency in causation. The regulatory verdict was clear: while causal phenomena might involve AI systems, *causal responsibility* will always fall on humans. In light of this, the *causal knowledge* involving AI should allow the identification of a responsible human agent, without it being necessary - or even desirable - to determine if a specific consequence (like harm) was caused by an AI system *having acted alone*. End of the story.

C. AI Output as the Object of Inquiry

Fast-forward a few centuries from Spinoza and Hume: our explanatory abilities have no doubt improved, only nowadays, it is not falling stones but Artificial Intelligence (AI)[16] that pushes us to the edge of what is knowable and explainable. In particular, in the field of AI liability, Spinoza's 'reduction to ignorance' method seems to be far from *dépassé*: just as, centuries ago, divine volition and action were assumed to be the original cause of all worldly occurrences, human intent, action or inaction (in other words, human agency) are now assumed to be the root cause of all harm occasioned by the use of AI systems. Not because we have conclusive evidence that this is always true, but because such is our millennial, *normative belief*: people harm people, even if the causing of harm is made possible through the use of sophisticated, smart technologies.

Our collective preference to uphold an anthropocentric view of causality is perhaps a 'healthy' reaction to the realization that AI systems can work in mysterious ways. Examples of recruitment AI, automated vehicles and credit-scoring AI, to name a few have shown that intelligent systems may not always offer the possibility for their decisional processes to be scrutinized. To compensate our lack of causal knowledge in such instances, we turn to our 'nomic' causal representations, *seeking refuge in the human agency postulate*, as cornerstone of longstanding liability doctrines.[17] But those doctrines date from a time when non-human intelligence and agency were inconceivable… In recent decades, part of scholarship reflected on whether the concept of agency ought to be reconceptualized in order to extend to non-human entities who reason (and therefore, act) in similar ways as humans. The consensus has fallen on the

[16] For the purpose of this paper, we will refer to the definition of AI included in the AI Act. See, Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (AI Act), and amending certain Union legislative acts, COM(2021) 206 *final*, art. 3(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." The 'techniques and approaches' mentioned in Annex I are Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning (Annex I, a)); logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems (Annex I, b)) and statistical approaches, Bayesian estimation, search and optimization methods (Annex I, c)).

[17] See *inter alia* Ljupcho Grozdanovski, «L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale? Réflexions sur l'avenir du régime de responsabilité du fait de produits défectueux dans l'Union européenne» (2022) 232/233 2 *Réseaux,* 99.

fact that, their levels of intelligence[18] notwithstanding, AI systems form a class of commodities[19] meaning that, when harm is causally associated with those systems, the culprit will invariably be the human having either programmed or used them. But in doing so, are we not choosing a *causal belief* over *causal knowledge*? Are we not (re)creating a 'refuge of ignorance'?... It certainly seems so. In lieu of looking to design discovery methods through which litigants could uncover the *actual causal power* of AI systems we, as a collective, seem to prefer the safety of what we have always known to be true *i.e*. that rational and moral agency can only be a human prerogative.

The postulate of the 'human puppeteer' - discrete but always present behind the scenes in opaque AI decision-making - could perhaps be tenable, had we remained in the early days of Artificial Narrow Intelligence (ANI). In the stone age of AI - dating to only a few years ago - we mostly dealt with hyperspecialized "idiot savants,"[20] very good in performing one task or a set of tasks, useless at anything else. Since then, technological innovation has developed at a galloping pace, resulting in more generally intelligent systems. Generative AI like ChatGPT gives an illustration of this. We have not yet reached the stage of Artificial General Intelligence (AGI)[21] and certainly not that of Artificial Super Intelligence (ASI)… But we are getting there. Of course, 'general intelligence' is a multifaceted concept which includes - under the 'general' label - several types of intelligence.[22] For the sake of simplicity, we will consider the level(s) of AI intelligence as correlating to level(s) of *cognitive* and *decisional*

---

[18] Though there are many possible ways to define intelligence *tout court*, it is possible to argue that it translates to a series of abilities that allow an agent to *autonomously arrive* at a solution or make a prediction in a context where all the variables are not known. See Kristin Thorisson, Helgi Helgasson, "Cognitive Architectures and Autonomy: A Comparative Review" (2012), 3-2 *J. Gen. AI*, 1, at 3. Intelligence in connection to (artificial) agency has raised issues on whether AI's autonomy can warrant the recognition of some form of agency. We have argued in our previous work that AI's autonomy is similar to human autonomy *functionally* in that AI systems are able to simulate human skills which, when exercised - and as a general rule of thumb - aim for efficiency and accuracy. AI's ability to replicate human intelligence has not yet extended to *human ontology*, placing in the core of what it means to be 'intelligent' the (autonomous) ability for empathy and more generally, the ability to distinguish right from wrong. On the distinction between functional and ontological aspects of human and non-human intelligence, see Ljupcho Grozdanovski, «L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale? Réflexions sur l'avenir du régime de responsabilité du fait de produits défectueux dans l'Union européenne», *cit. supra*, at 9.

[19] Commoditization of advanced technologies is not recent. One of its oldest expressions can be found in American caselaw which interpreted robots as mechanical devices "a mere automation, that operates through scientific or mechanical media" but is not "a living thing; it is not endowed with life." See *Louis Marx & Co. and Gehrig Hoban & Co., Inc. v. United States* case (40 Cust. Ct. 610, 610 (1958)). For a comment on this and other US cases in the field of robotics, See Ryan Calo, "Robots in American Law," *Legal Studies Research Paper N° 2016-4* (University of Washington - School of Law), available on: http://euro.ecom.cmu.edu/program/law/08-732/AI/Calo.pdf (last accessed on 20 Jan. 2024), at 14.

[20] Matt Paisner, Michael T. Cox, Michael Maynord, Don Perlis "Goal-driven autonomy for cognitive systems", Proceedings of the Cognitive Science Society (2014), available at <pdfs.semanticscholar. org/2c9c/2bb5381a0e094d80b2095dbedbbe6546911e.pdf>, 2085–2090, at 2085.

[21] AGI includes AI systems able to perform most, if not all, cognitive functions as good as humans, Gonenv Gurkaynak, Ilay Yimaz, Gunes Haksever "Stifling Artificial Intelli- gence: Human Perils" (2016) 32-5 *Comp. L. & Sec'y Rev.*, 749, at 751.

[22] The taxonomy of intelligence is a delicate issue, in the sense that clear-cut categories or types of intelligence are difficult to establish. There are, however, several types of 'abilities' which scholars have associated with types of intelligence. They include, namely, so-called fluid intelligence, crystallized intelligence, visual intelligence, auditory intelligence, cognitive processing speed etc. See Wan Nurul Izza Wan Husin, Angeli Santos, Hazel Melanie Ramos, Mohamad Sahari Nordin, "The place of emotional intelligence in the 'intelligence' taxonomy: Crystallized intelligence or fluid intelligence" (2013) 97 *Procedia - Soc. & Behav'l Sci.*, 214, at 215.

autonomy in reaching a preassigned goal and, in some cases - like those of Deep Learning (DL) systems[23]- even selecting the goal(s) to be achieved. As we will argue further, the more generally intelligent the system, the greater its level of autonomy and the more accurate its outcomes but also, the less scrutable the reasoning patterns through which those outcomes are arrived at.

In sum, we seem to be caught in a thug of war between, on the one hand, imminent technological evolution which promises to emancipate AI form any realistic form of 'panoptic' human control and oversight and, on the other hand, a regulatory *penchant* for stability and continuity, characterized by AI commoditization and the sacrosanct human agency principle. This, of course, has an important impact on the *design of the systems of evidence* used in the adjudication of disputes dealing with AI liability.

D. The Possibility for Evidence and (Causal) Explanation Pertaining to AI Output

The concept of legal evidence[24] is a curious beast, because it simultaneously answers to two sets of validity criteria: those of truth and those of fairness. The realm of truth is that of discovery and epistemology[25] which, in the field of procedural law, find a specific expression in legal rules and principles of evidence. The *raison d'être* of those rules and principles is to *epistemically frame* the process of fact-finding and fact-assessment under an *independent* (impartial) standard of accuracy. Of course, in adjudicatory contexts, fact-accuracy is not sought for accuracy's sake: 'accurate' knowledge of the disputed facts is a factor that impacts the fairness of a dispute's *outcome*.[26] This accuracy/fairness interplay is precisely what marks the specificity of legal evidence as a concept: fairness is both the expected *outcome* from an institutional - most commonly, judicial - law-to-fact application and the *epistemic constraint* of the process through which knowledge of the disputed facts is construed. The longstanding normative creed is, indeed, that only fair procedures (*i.e.* designed to create conditions of fair adjudication) can be conducive to fair outcomes.[27]

Concretely, this means that the parties in a dispute should have *equal procedural abilities* to access and give the evidence they view as relevant and probative. This

---

[23] DL systems are models with multilayered neural networks that are trained with large data sets of data and able to solve highly complex information processing tasks. For an analysis of DL models in fields like medicine, see Christopher M. Bishop, Hugh Bishop, *Deep Learning. Foundations and Concepts* (Springer, 2024).

[24] According to Wigmore, 'evidence' can be understood as any knowable fact or group of facts, considered with a view to its being offered for the purpose of producing conviction as to the truth of a proposition. See John Henry Wigmore, *Evidence in Trials at Common Law* (Little, Brown, 4th ed., 1961).

[25] Epistemology will be understood as the field of study focused on the theorizing and structuring methods of knowledge and beliefs construction. See, *inter alia*, Jaakko Hintikka, *Socratic Epistemology. Explorations of Knowledge-Seeking by Questioning* (CUP, 2012) at 11 seq.

[26] In some strands of evidence scholarship, accurate representations of fact are needed to give way to a correct application of the law, the belief here being that - as Grando put it - "accurate decisions are usually fair." See Michelle T. Grando, *Evidence, Proof, and Fact-Finding in WTO Dispute Settlement*, (OUP, 2009) at 11.

[27] The fair procedures/fair outcomes parallelism derives from Rawls' idea(l) of so-called *perfect procedural justice* model by virtue of which fair procedures, if correctly followed, yield correct and fair results. See John Rawls, *A Theory of Justice* (revised ed.) (Harv. UP, 1999), at 75 seq.

procedural parity, typically expressed in the fair-trial safeguards,[28] is meant to define a level of *baseline equality*, placing the parties on an equal procedural footing when they make their views known before an adjudicating authority. From the evidentiary debate thus organized - and conceptually akin to Habermas's discursive ethics[29] - 'truth' is expected to surface, giving courts the information necessary to answer two cardinal questions: '*what and who caused the dispute*?' and '*what is the most adequate (or fair) legal solution to that dispute*?'

With the truth/fairness interplay in the backdrop, let us turn to the *evidence of causation*. Two issues can be flagged as relevant. First, there is the already discussed (Humean) issue of *necessity*, which invites us to reflect on the evidence and corresponding explanations litigants should *be able to access* to effectively argue how an event was causally linked to another event (typically, a harm). Second, there is - again - the issue of *fairness*: how should systems of evidence, in the EU, be (re)designed so that the *evidence flagged as necessary* under point (1) can be adduced in conditions of procedural parity? To answer both questions - as this paper's chief ambition - we must address a more fundamental issue, characteristic of AI liability: *what exactly are we seeking to explain when we give evidence on the casual link between an AI system and a harm*? Two roads diverge[30] here: the one, more travelled, asks us to explain causality from the vantage point of human agency; the other, less travelled, asks us to engage in proper discovery of the causal chain between a harm and a harm-causing conduct (possibly of a non-human, intelligent entity).

We already alluded to the first alternative earlier: in lieu of engaging in Byzantine debates on whether harm can be imputable to an AI having acted alone, we seem to prefer the *belief* that the authorship of (and by that, the responsibility for) that harm is incumbent to a human (programmer, user), without this warranting an in-depth demonstration of whether that human's actions *actually* contributed to the harm-causing automated decision. Taking human agency as a presumed (as opposed to established) cause of such harm is, of course, reassuring because it maintains conceptual continuity, but it barely holds in the scenario where there is *no evidence of human involvement*, and yet harm was somehow occasioned by an AI's use.

The second alternative is the one where the evidentiary debate on causality would include discovery proper, yielding explanations on *how* a given system made a harmful decision or prediction. Part of AI scholarship supports this view. For example, Barredo Arrieta *et al*. made a point on the nature of causal knowledge, by making the distinction between *causality* and *causation*. Causality, the authors argue, requires a

---

[28] In the EU, the fair trial safeguards are currently enshrined in Article 47 of the EU Charter of Fundamental rights (EUCFR). Those safeguards include the right to a fair and public hearing within a reasonable time by an independent and impartial tribunal previously established by law (Art. 47(2) EUCFR).

[29] We refer to Habermas' 'ideal speech situation' based on three (participatory) equality-enhancing rules namely, the rule of participation, the rule of equal opportunity and the rule against compulsion. See Jürgen Habermas, *The Theory of Communicative Action: Reason and Rationalization of Society*, Beacon Press (1984).

[30] This is an expression drawn from Robert Frost, "The Road Not Taken" (2021) 4 *The Objective Standard*, at 79.

"wide frame of prior knowledge to prove that observed effects are causal."[31] Causation "involves correlation, so an explainable ML model could validate the results provided by causality inference techniques, or provide a first intuition of possible causal relationships within the available data."[32]

When a court seeks to determine 'what happened' in an AI liability case, the knowledge that they would normally seek is that of causation, as defined by Barredo Arrieta *et al*. The practical problem here is that the discovery of causation may not be feasible because the evidence thereof may not be - reasonably - within the litigants' reach. As mentioned earlier, the variable-correlations an AI system may have made prior to the occurrence of harm often remain partially or fully unknowable to human agents (sometimes, including the programmers). For example, how could a loan applicant *even suspect* that an AI, used to preapprove loan applications, was racially biased? That applicant would presumably have no access to the applicants the system had approved, nor would they have information of how the bank usually assesses applicants' credit. In this context, to make their argument, the claimant would require access to two types of evidence. First, they would need to establish that the system's output was indeed racially biased, which implies that they should, somehow, understand and explain that the outcome of a specific variable association (e.g. place of residence *cum* ability to repay the loan) was a key factor in the occurrence of racial discrimination. Second, to causally explain that discrimination, they would need to establish and explain *what actually caused* it (*i.e.* explain if the bias was embedded in the programming data or machine learnt.). If there is evidence showing that the bias was machine learnt, who should then be held as liable?...

We have examined the issue of allocating liability elsewhere.[33] Our suggestion was that, when the human authorship of AI-related harm is not proven, the liable agent (*i.e.* held to compensate the harm) should be the *one having accepted the risk* of the harm occurring. That agent can be either the *programmer*, having released in the market a system that has, in the past, been prone to certain types of malfunctions (e.g. developing unfair biases) or the *user* who, aware of the harms a system may typically cause, had chosen to nevertheless use it.

In this paper, our focus will be more on the *evidentiary causal explanations* needed to determine the *locus* of AI liability, under the European Union's (EU) regulatory framework. In this context, we will explore what can and should be established and explained, when the chain of causality is fully or partially unknowable - possibly more so than in 'ordinary' causal scenarios (*i.e*. those that do not include intelligent systems).

---

[31] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, *et al*. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (2020) 58 *Information Fusion*, 82, at 86.

[32] *Ibid*.

[33] Ljupcho Grozdanovski, «L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale? Réflexions sur l'avenir du régime de responsabilité de produits défectueux dans l'Union européenne» *cit. supra*.

E. A Shift in Perspective: From Causal Explanations Required by Law to Causal Explanations Asked for (and Given) by Litigants

Bearing in mind our 'two roads diverge' metaphor, the legislator of the EU - much like the legislators of several countries around the world - was faced with the difficult task of regulating AI liability against the backdrop of two competing principles: that of discovery (causal knowledge) and that of human agency (belief). In the field of procedure, the guiding principle in choosing the one over the other should, no doubt, be that of (procedural) *fairness*.

If we draw on standard liability doctrines and consider that fair outcomes *always* call for accurate knowledge of causation, then AI liability should not be viewed as an exception, meaning that the culprit should be identified through evidence, not presumptions. If, however, a standard of fairness is thought to be best upheld when the law's postulates remain unshaken, then the level of causal accuracy in AI liability will be required *to the extent that it coheres* with the presumption of human agency… But this is AI liability viewed from the heights of the conceptual tower that is (standard) liability law. It is perhaps more relevant to inspect what happens in the trenches *i.e.* in the *already adjudicated* and/or *forthcoming* AI liability disputes. These invite us to set aside the deontic stance of the law and take on a more down-to-earth, fact-based and, dare we say, humanist perspective by addressing the oft-forgotten '*what do the people need*?' question. Do the *litigants themselves* consider that, to argue causation, they need to understand *how* an AI system caused harm or is this knowledge procedurally irrelevant to them?...

Fundamentally, this paper seeks to conceptualize *procedural fairness* in the face of AI and to do so, it will follow a *bottom-up approach*. It will depart from court practice - mainly North-American - and will seek to induce the features of a concept of 'AI fairness' based on the *procedural needs* expressed by litigants in AI liability cases. Against this backdrop, this paper will critically assess the EU's AI liability regulatory framework, sketching out ways in which that framework ought to be applied, in view of better supporting the litigants' so-called *effective participation*[34] in the resolution of future AI liability disputes.

The doctrinal strand that we will take as a key analytical referent is the doctrine of so-called *procedural abilities* - basic entitlements litigants ought to have to effectively make their views known before a court. This school of thought developed as the procedural 'spinoff' of the so-called *capabilities* approach, as conceptualized in the seminal work of Sen[35] and Nussbaum.[36] Unlike previous - say, Rawlsian[37] - justice theories, aimed at distilling normative, universal understandings of fundamental principles of justice like 'the right,' 'the equal' and 'the good,' the capabilities strand is more interested in the *entitlements* individuals should enjoy to live 'meaningful' lives, the real-world injustices notwithstanding. In a taxonomical *élan*, Nussbaum seminally suggested ten fundamental capabilities which, she argued, are the universal prerequisites for a thriving human existence. These are: life, bodily health, bodily

---

34 Lawrence Solum, "Procedural Justice" (2004) 78 *Calif. L. Rev.*, 181, at 305.
35 Amartya Sen *The Idea of Justice* (Harv.U.P, 2009).
36 Martha C. Nussbaum, *Frontiers of Justice. Disability, Nationality, Species Membership* (Harv.U.P., 2006).
37 John Rawls, *A Theory of Justice*, *cit. supra*.

integrity, senses, imagination and thought, emotions, practical reason, affiliation, play and control over one's environment. The capabilities approach has also been the object of criticism. However, one of its merits is that it offers, if not a perfect, at least a *workable understanding of fairness*, acting more as a general guideline for regulatory action, than a mandatory ethical precept. This is no doubt the reason why Sen's and Nussbaum's scholarship laid the theoretical foundation for the United Nations' (UN) Sustainable Development Goals (SDGs).

In the field of procedure, the capabilities approach was echoed in the so-called procedural abilities - basic procedural entitlements that parties in adjudicatory contexts should have to 'meaningfully'[38] participate in adjudicatory processes. Mirroring Nussbaum's decalogue, Awusu-Bempah[39] suggested a taxonomy of procedural abilities which are also ten: 1) understand the nature of the charge; 2) understand the evidence adduced; 3) understand the trial process and the consequences of being convicted; 4) give instructions to a legal representative; 5) make a decision about whether to plead guilty or not guilty; 6) make a decision about whether to give evidence; 7) make other decisions that might need to be made by the defendant in connection with the trial; 8) follow the proceedings in court on the offence; 9) give evidence; 10) any other ability that appears to the court to be relevant in the particular case.[40] The choice of the procedural abilities strand as the 'intellectual compass' of our analysis is justified by our preoccupation with *effectiveness* translated in, what we previously labelled as, our bottom-up approach to conceptualizing AI (procedural) fairness.

As a matter of *personal conviction* of this paper's author: litigants should feel that the law gives them a discursive space where they can speak their truth.

As a matter of *factual accuracy of AI causation*: litigants should feel that important decisions like those on responsibility or guilt are not arbitrary but informed, based on accurate information.

As a matter of *procedural fairness* in the face of AI: litigants should feel that a system of procedures and remedies provides them with the abilities *they need* to discuss matters like innocence and guilt.

In this context, rather than investigating how (procedural) law should align itself concerning the proof of causality or the presumption of human responsibility, it may be more prudent to contemplate what litigants engaged in discussions about AI-related harm *should be capable of proving and explaining* to ensure a fair resolution to their dispute. This shift from the 'procedural ought' to the 'procedural need' naturally pushes us to raise the issue of the *access* to evidence: if an AI's inner workings are unknowable, how can non-expert litigants access the information *they need* to provide an explanation on who or what caused the harm? Should law provide a procedural right to access such evidence?... These and other questions will be raised in our analysis of the interrelationship between 'what is' explanation in connection to AI, and 'how' that explanation ties (or not) into causal explanations of harm given in AI liability disputes.

---

38 Lawrence Solum, "Procedural Justice," *cit. supra*, at 305.

39 Abenaa Owusu-Bempah, "The interpretation and application of the right to effective participation" (2018) 22-4 *The Int'l J. of Evidence & Proof*, 321.

40 *Id*., at 330.

However, before we outline the structure of our arguments on this point, it is necessary to say a few words on the EU's regulatory frameworks of AI.

F. The EU's Regulation of AI

1. The Substantive Regulation - the AI Act

AI regulation in the Union essentially evolved in two stages. First came substantive law in the form of a proposal for a Regulation laying down harmonized rules on AI (AI Act).[41] We have extensively explored the history and content of this instrument elsewhere and will not offer a detailed account thereof here. We will but mention the aspects of the AI Act that we view as relevant for the remained of this paper.

On the *type of regulation*, the AI Act can, in essence, be thought of as an instrument that transposes product safety logic to risks of fundamental rights violations. The operative assumption is that, like 'ordinary' products, AI systems can be safely used if their programing and use comply with a number of predefined technical standards. This of course is debatable, but we will refrain from further commenting on whether the tried-and-true method of standardized product manufacturing is a good fit for regulating products which are not automated but intelligent. This was *inter alia* a point raised in one of our recent studies.[42]

---

41 Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final.

42 Ljupcho Grozdanovski, Jérôme de Cooman, "Forget the Facts, Aim for the Rights! On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union" (2023) 2 *Rutgers Comp. & Tech'y L. J.*, 207.

More importantly, the AI Act includes a four-level taxonomy of risks: non-high,[43] limited,[44] high and unacceptable.[45] The so-called high-risk AI systems are the most relevant for this paper because the evidentiary frameworks included in the EU's procedural regulation following the AI Act were specifically designed to enable proof of causation in cases involving those systems.

High-risk AI is a class of intelligent systems assumed to pose threats of fundamental rights violations and yet their commercialization is allowed: "rather than being altogether prohibited, they are subject to mandatory requirements, chiefly transparency (art. 13) and human oversight (art. 14)."[46] The AI Act - we argued in our study - distinguishes between two categories of high-risk AI: "the first category includes systems intended to be used as safety component of products covered by EU sectorial product legislations listed in Annex H (art. 6(1)(a)) and that are subject to third party *ex-ante* conformity assessment (art. 6(1)(b)), bearing in mind that a safety component is 'a component of a product or of a system which fulfils a safety function for that product or system or the failure or malfunctioning of which endangers the health and safety of persons or property" (art. 3(14))."[47] The second category "includes stand-alone Al systems with mainly fundamental rights implications that are explicitly listed in Annex III (art. 6(2))."[48]

Annex III of the AI Act lists *eight key areas* where high-risk systems are most likely to be used: biometric identification and categorization of natural persons; management and operation of critical infrastructure; education and vocational training; employment, workers management and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law

---

43 *Id*., at 243: "*non-high-risk AI systems* are defined in opposition to high-risk systems. As high-risks Al systems are exhaustively enumerated, non-high-risks Al systems form a residual (and presumably the largest) category. The regulatory principle for those systems is the absence ofa duty to comply with the mandatory requirements which target the high-risks systems (Art. 8). Developers and users of non-high risk AI systems are, however, encouraged to voluntarily apply these requirements through codes of conduct (Art. 69)."

44 *Id.* at 243-244: "*Limited risks AI system* are, similarly, not subject to mandatory requirements set up in the AI Act (art. 8). However, the Al Act does establish an obligation of transparency for systems which, though formally qualified as non-high risk, interact with natural persons (art. 52(1)), perform emotion recognition or biometric categorization (art. 52(2)). Such systems ought to be designed in a way that natural persons know they interact with or are exposed to an Al system. In a similar vein, users of so-called deepfake technology - *i.e.,* hyper- realistic videos using face swaps that leave little trace of manipulation - are required to disclose that the content has been manipulated or artificially generated (art. *52*(3))."

45 *Id*., at 244: "Al systems that pose *unacceptable risks* are subject to an *ex officio* ban (art. 5). It should be stressed that military applications are excluded from the scope of the Al Act (art. 2(3)). With this exception in mind, AI systems that either use subliminal manipulation of natural person's consciousness (art. 5(1)(a))) or exploit vulnerabilities of a specific group of persons due to their characteristics, *e.g.,* age, physical or psychological disability (art. 5(1)(b)) in order to distort people's behavior in a way that is likely to cause physical or psychological harm are prohibited. The ban also extends to AI systems used by public authorities that score natural persons based on their personal and social behavior, known or predicted (art. 5(1)(c)) as well as those that may lead to detrimental or unfavorable treatment of certain natural persons or groups either "in social contexts which are unrelated to the contexts in which the data was originally generated or collected" (art. 5(1)(c)(i)) or that is "unjustified or disproportionate to their social behavior or its gravity" (art. 5(1)(c)(ii))."

46 *Id*., at 244.

47 *Ibid*.

48 *Ibid*.

enforcement; predictive policing and migration, asylum and border control management. For the systems used in these sectors, the AI Act defines technical standards for compliance such as risk-management (Art. 9), data and data governance (Art. 10), technical documentation (Art. 11), record-keeping (Art. 12), transparency and provision of information to users (Art. 13), human oversight (Art. 14), accuracy, robustness and cybersecurity (Art. 15).

The European Commission's initial proposal for the AI Act underwent several modifications from the EU's legislative bodies *i.e*. the Parliament and the Council. A provisional agreement was eventually reached on 9 December 2023.[49] However, as of that date, a definitive consolidated version of the AI Act was not released; only a document compiling the specific agreed-upon amendments was disclosed. On 22 January 2024, an unofficial version of the AI Act was leaked by EurActive editor Luca Bertuzzi.[50] For the remainder of this paper, we will refer to this leaked version when citing specific provisions from the AI Act.

By identifying the sectors where the risk of AI-related harm is 'high,' the AI Act is, without a doubt, a laudable first since it transcends the congenital diversity of AI as a class of new technologies. However, this instrument relies on a somewhat fallacious assumption: that compliance will somehow suffice for harm to be prevented. Because of this, the AI Act contains virtually no provisions on the *ex post* protection of human agents when a harm eventually ends up materializing. To fill this gap, in September 2022, the EC published a Directive Proposal on adapting non-contractual civil liability rules to AI (AI Liability Directive - AILD).[51]

2. The Procedural Regulation

a. The AI Liability Directive - AILD

The AILD echoes the regulatory principles enshrined in the AI Act and, much like this instrument, it seeks to strike a balance between increasing market gains (by fostering competitiveness and investment in research and innovation), and the safeguard of - what we may call - *non-waivable fundamental rights and democratic values*. In this context, the AILD explicitly states that "to reap the economic and societal benefits of AI and promote the transition to the digital economy, it is necessary to *adapt in a targeted manner certain national civil liability rules to those specific characteristics of certain AI systems*."[52] According to this Directive, the point of reconciliation between market efficiency and procedural fairness is *trust*.[53] The 'adaptations' of national civil liability rules the AILD mentions are assumed to

---

[49] See https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf (last accessed on 23 Jan. 2024).

[50] See Jedidiah Bracy, "EU AI Act: Draft consolidated text leaked online," available on: https://iapp.org/news/a/eu-ai-act-draft-consolidated-text-leaked-online/ (last accessed on 23 Jan. 2024).

[51] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence (AI Liability Directive - AILD) COM (2022) 496 final.

[52] *Id*., Preamble, pt 5 (emphasis added).

[53] This echoes the key objectives highlighted to frame the EU's Regulation of AI, namely - what the EC called - the ecosystem of trust and the ecosystem of excellence. See European Commission, White Paper, *On Artificial Intelligence - A European Approach to Excellence and Trust*, COM (2020) 65 final.

contribute to "*societal and consumer trust* and promote the roll-out of AI"[54] but they are also assumed to "*maintain trust in the judicial system*, by ensuring that victims of damage caused with the involvement of AI have the same effective compensation as victims of damage caused by other technologies."[55] A 'workable equilibrium' between these two 'pillars of trust' can - the Directive states - be achieved through the harmonization of certain non-contractual *fault-based* liability rules, aimed at ensuring that persons who claim compensation for harm caused by AI systems "enjoy a *level of protection equivalent to that enjoyed by persons claiming compensation for damage caused without the involvement of an AI system*."[56]

The AILD carries the imprint of the initial regulatory impulse given by the early-day EU instruments on AI (namely the HLEG's Guidelines on Ethics[57] and the White Paper on AI[58]): the achievement of market gains is not pursued in parallel to a 'pedagogical' protection of fundamental rights; rather *the realization of market gains is framed by the protection of those rights*. This is a relevant point of comparison with the AI Act which, following a logic of prevention of AI-related risks, defines the notion of 'risk' precisely as a violation of fundamental rights and values.[59] That understanding of risk has largely shaped the design of the system of evidence contained in the AILD.

Two main features of this Directive will be highlighted, at this stage. First, it creates a *fault-based* - as opposed to strict - liability regime. This means that the compensation of harm occasioned by an AI system will require *proof of fault*. In this regard, the link between the AI Act and said Directive is salient, given that the notion (and therefore, evidence) of fault is defined as a behavior consisting in "the non-compliance with a duty of care laid down in Union or national law directly intended to protect against the damage that occurred" (Art. 4(2)). The notion of 'fault' is therefore not defined as one might typically expect *i.e.* as the result from a *wrongful* act (*i.e*. a violation of a duty of care, regardless of whether that duty is recognized in a legal provision).[60] Rather, 'fault' is a failure to comply with the standards explicitly laid down in the AI Act's provisions.[61] Fault is therefore understood as *unlawful conduct* (non-compliance with the law) which, as we will subsequently argue, has an important impact on the types of evidence that litigants are authorized to ask for and adduce on the grounds of the instrument considered. Under certain conditions - also discussed further - it is the proof of this type of 'fault' that provides the grounds for a *presumption of causality*.

The justification for this (overly legalistic?) understanding of fault is the fact that the AI Act creates *full harmonization* of the technical requirements pertaining to

---

54 AILD (COM (2022) 496 final) *cit. supra*, Preamble, pt 5.
55 *Ibid* (emphasis added).
56 *Id*., Preamble, pt 7.
57 High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI* (2019), available on: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed on 20 Jan. 2024)
58 European Commission, White Paper, On Artificial Intelligence - A European Approach to Excellence and Trust.
59 Article 2 AI Act, *cit. supra*.
60 See our discussion on wrongfulness and unlawfulness *infra*, Sub-Section 2.2.1 (A).
61 AILD (COM (2022) 496 final) *cit. supra*, Preamble, pt 26: "This Directive covers the fault constituting non-compliance with certain listed requirements laid down in Chapters 2 and 3 of [the AI Act] for providers and users of high-risk AI systems."

the programming and use of high-risk AI systems.[62] Against this backdrop, "and in *full consistency with the logic* of the AI Act,"[63] one of the ambitions stressed in the AILD is to provide steps for providers to adopt or not risk management measures as *relevant evidence* for the purpose of determining whether there has been a case of non-compliance.[64] Further in this paper, we will be critical of the notion of fault, as defined in the AILD and the system of evidence designed around it. At this stage, may it suffice stressing that fault is the point where the AI Act and the AILD intersect: to prove fault under the latter, one would need to prove non-compliance with the former.

Second, the AILD introduces *minimal harmonization* which means that national courts will apply their national rules of procedure and evidence in areas not covered by this harmonization. However, the Directive provides some important procedural guidelines. Two of its key advancements are the *right to request disclosure of evidence* (Art. 3) for victims, and the *right to rebut the so-called presumption of causality*, for respondents (AI providers or users). These are arguably the main source of value of the instrument under consideration. By recognizing the right to request disclosure of evidence, the latter gives a procedural expression to the twin principles of transparency and explainability: after all, only a transparent automated decision can 'open' the access to facts, thus providing grounds for plausible arguments of fault and causation to be presented before a court. However, our analysis of these rights will reveal several inconsistencies in the way the right to request disclosure of evidence is exercised.

Regarding the allocation of the burdens of proof, the AILD defines those - albeit in general terms - by canvassing the main requirements that claimants should meet when arguing and proving fault and causation.[65] The types of relevant facts (*facti probandi*) vary, depending on whether the respondent is an AI provider or an AI user. When the respondent is a provider, the claimant is held to prove the latter's failure to comply with the requirements, listed in the AI Act, that target the so-called 'high-risk' systems. These requirements include transparency (Article 13 AI Act); effective oversight (Article 14 AI Act); accuracy, robustness and cybersecurity (Articles 15 and 16 AI Act); the taking of necessary corrective actions (Articles 16 and 21 AI Act). Alternatively, when the respondent is a user, the claimant is held to prove the former's failure to comply with instructions of use (Article 29 AI Act) and/or exposure of the system to input data which is not relevant in view of the system's intended purpose (Article 29(3), AI Act).

---

62 *Ibid*.
63 *Ibid*.
64 *Ibid*.
65 It should be stressed that the Member States' courts are not deprived of their discretion in defining the relevant facts. However, this discretion notwithstanding, said Directive provides guidelines on the issue of relevance, as regards the proof of fault. Art. 3(1), AILD, "Member States shall ensure that national courts are empowered (…) to disclose relevant evidence (…) about specific high-risk AI systems that is suspected of having caused damage, but was refused, or a claimant, to order the disclosure of such evidence from those persons."

b. The Revised Product Liability Directive - R-PLD

Dating back to 1985, the Product Liability Directive (PLD)[66] naturally did not have the foresight of including AI in its scope of application.'[67] According to the European Commission (EC), the PLD's shortcomings warranting revision mainly had to do with the *design* of the system of evidence the Directive created. In particular, the proof of defectiveness and its link to a harm had shown to be challenging for claimants, especially in complex cases like those involving pharmaceuticals, smart products or AI-enabled products.[68]

Unlike the AILD - which creates a fault-based liability system of evidence - the PLD establishes a *strict liability* system, not requiring proof of fault. The relevant fact (*factum probandum* or the fact for which evidence is sought) that litigants are called to establish within the PLD is *defectiveness*. The proposal for a revision of the PLD (R-PLD) did not change this aspect of the original PLD. The 'new' product liability framework also integrates the strict liability logic, stating that, when AI systems are defective and cause physical harm, property damage or data loss "it is possible to seek compensation from the AI-system provider or from any manufacturer that integrates an AI system into another product."[69]

Prior to submitting the R-PLD proposal, the EC launched a public consultation, during which 77% of participants underlined the procedural challenges faced by litigants in cases involving technically complex products.[70] Pushed to revisit the system of evidence from 1985 - in view of lightening the burden of proof for victims - the EC considered several lines of revision, but ultimately decided on two. First, regarding the *types of products* included in the 'new' Directive's scope of application, the EC chose to include, in the 'new' Directive's scope of application, manufacturers and providers of intangible digital elements, as well as 3^{d} parties providing software added to a product. Second, regarding more specifically the *design* of the system of evidence centered around defectiveness, the EC opted for a system that would ease the burden on consumers by harmonizing the rules on the disclosure of technical information to the victims and the conditions under which defectiveness can be presumed.[71]

To achieve the ambition of 'lightening' the burden of proof especially for

---

66 Council Directive 85/374, of 25 July 1985, on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, OJ L 210, 7.8.1985, p. 29.

67 In Art. 2 PLD, 'product' is defined as "all movables, with the exception of primary agricultural products and game, even though incorporated into another movable or into an immovable (…) 'Product' includes electricity." In the proposal for revision of the PLD (R-PLD), the 1985 definition is broadened. Art. 4(1) R-PLD, states that 'product' means "all movables, even if integrated into another movable or into an immovable. 'Product' includes electricity, digital manufacturing and software."

68 EC, Proposal for a Directive of the European Parliament and of the Council on liability of defective products, COM (2022) 495 final, at 1.

69 *Id*., at 3.

70 *Id*., at 8. The percentage was considerably higher among consumer organisations, NGOs and members of the public (95%) than among business and industry organisations (38%). Industry stakeholders were more open to information disclosure obligations and easing the burden of proof in complex cases than to reversing the burden of proof, which they considered a radical option that would harm innovation.

71 *Id*., pt 9.

claimants, the R-PLD sought to mend the asymmetry of information between the parties in cases characterized by technical or scientific complexity.[72] To do so, it used a well-known procedural 'trick': presumptions. Rebuttable presumptions of fact - the R-PLD states - are "a common mechanism for alleviating a claimant's evidential difficulties, and *allow a court to base the existence of defectiveness or causal link on the presence of another fact that has been proven*, while preserving the rights of the defendant."[73] Indeed, national courts can presume the defectiveness, causation, or both where, "notwithstanding the defendant's disclosure of information, it would be *excessively difficult* for the claimant, in light of the technical or scientific complexity of the case, to prove its defectiveness or the causal link, or both. In such cases, requiring proof would undermine the effectiveness of the right to compensation."[74]

Though the AILD and the R-PLD differ regarding their *facti probandi* (respectively, fault and defectiveness), they converge in two important ways. First, both instruments recognize a right to request a disclosure of evidence (in an *élan* to make evidence more feasible) for claimants. Second, in both instruments, the defendants' refusal to disclose 'relevant evidence' - whatever that might be - generates presumptions (of fault, of defectiveness and/or of causation).

Following up on the 'lightening the burden' idea, Article 9 R-PLD establishes the basic tenets of the upcoming evidentiary regime in product liability. The right to compensation under this instrument depends on the claimant's ability to prove the defectiveness of the product,[75] the damage suffered and the casual link between the two.[76] The defectiveness of the product "shall be presumed" in three cases, discussed further in this paper, but one of the three stands out: the case where there is evidence of the defendant's failure to comply with an obligation to disclose relevant evidence at their disposal.[77] The causal link between the defectiveness of the product and the damage "shall be presumed, where it has been established that the product is defective" and the harm caused is "of *a kind typically consistent* with the defect in question."[78] If due to technical or scientific complexity, the claimant experiences difficulties in proving defectiveness, the causal link or both, they can be presumed if the claimant

---

[72] *Id*., pt 30.
[73] *Id*., pt 31 (emphasis added).
[74] *Id*., pt 34 (emphasis added).
[75] *Id*., Art. 6 defines the notion of 'defectiveness' as failure to provide the safety which the public is entitled to expect, considering: the presentation of the product, including the instructions for installation, use and maintenance (a); the reasonably foreseeable use and misuse of the product (b); the effect of the product of any ability to continue to learn after deployment (c); the effect on the product of other products that can reasonably be expected to be used together with it (d); the moment in time when the product was placed on the market or put into service or, where the manufacturer retains control over the product, the moment when the product left the manufacturer's control (e); product safety requirements, including safety-relevant cybersecurity requirements (f); any intervention by a regulatory authority or by an economic operator (g), the specific expectations of the end-users for whom the product was intended.
[76] *Id*., Art. 9(1).
[77] *Id*., Art. 9(2). The duty to disclose evidence is enshrined in Article 8 R-PLD which states that national courts are empowered, upon request from the claimant "who has presented facts and evidence sufficient to support the plausibility of the claim for compensation, to order the defendant to disclose relevant evidence that is at its disposal" (Art. 9(1)). To determine if the disclosure is proportionate, national courts shall "consider the legitimate interests of all parties, including third parties concerned, in particular in relation to the protection of confidential information and trade secrets within the meaning of Article 2, point 1, of Directive 2016/943" (Art. 8(3)).
[78] *Id*., Art. 9(3), emphasis added.

gives "sufficiently relevant evidence" which shows that "the product contributed to the damage"[79] and "it is likely that the product was defective or that its defectiveness is a likely cause of the damage, or both."[80]

On the surface, these provisions do seem to lighten the burden for the claimants by conveniently setting out presumptions of defectiveness and/or causality. However, they are not - what we called in previous work[81] - *prima facie* presumptions *i.e.* facts held as established without prior evidence (like the presumption of innocence, for example). For the presumptions in the R-PLD to be established, the claimants carry the burden of establishing the basic facts (*indicia*) which if sufficient may, indeed, warrant the presuming of defectiveness and/or causality.

G. Structure and Outline of Main Arguments

To determine if and how explanations pertaining to AI output (as examined in connection to Explainable AI - XAI) can or should support explanations pertaining to the causal links between AI systems and harms suffered, we will follow, as *fil rouge* throughout this paper, the notion of *accuracy*. The inquiries that will frame our analysis are the following: does the accuracy of causal explanations in the field of AI liability *depend on* the accuracy of explanations pertaining to AI outputs? In the affirmative, which components should those explanations have, in order to be viewed as 'accurate'?

With accuracy in the backdrop, **Section 2** will lay down the analytical framework for the remainder of this paper, by focusing on the *type of knowledge* that explanations (*tout court*) provide and the standards that they respond to, in view of achieving accuracy or - at least - plausibility. Against the backdrop of various strands of history and philosophy of science, we will argue that, unlike 'scientific' knowledge (or 'knowledge proper'), explanations are held against *lower standards* of verifiability and accuracy, *believability* (in the eye of the explainee) being the criterion that truly sanctions - what scholars have called - the *goodness* of explanations (**Sub-Section 2.1.**). We will then go on to explore the 'goodness' conditions applied to explanations pertaining to causality in law (**Sub-Section 2.2.**). Since the purpose of *causal explanations* is to allow a competent authority (usually a court) to induce causation from series of correlations (*i.e*. positive associations between a conduct and a harm), the *evidence given*, as well as the criteria applied in its assessment are of utmost importance. Indeed, in cases where the cause-harm link is not self-evident or easily discernable, the type, probative value and relevance of the items of evidence given will play a major role in the mapping out of the stages that form the chain of causality which connects a wrongful and/or unlawful act to a damage.

With explanatory accuracy (*tout court* and in liability law) thus canvassed, **Section 3** will analyze how that concept relates to AI output. To do so, it will examine two sets of accuracy conditions: those applied to automated decisions/predictions and those applied to explanations of automated decisions/predictions. The first series of conditions will be our point of focus in **Sub-Section 3.1**. Bearing in mind the scholarship - explored in Section 2 - on the conditions for valid knowledge-construction, we will seek to determine if the 'knowledge' produced by non-human 'knowers'

---

[79] *Id*., Art. 9(4)(a).
[80] *Id*., Art. 9(4)(b).
[81] Ljupcho Grozdanovski, *La présomption en droit de l'Union européenne* (Anthémis, 2019).

presents any specificity (in terms of how it is formed and when it is 'accurate') in the context of traditional epistemology. Against this backdrop, we will raise the issue of the human knowability of AI output and critically address (and assess) the assumptions we make when we seek to explain automated decisions and predictions which are, partially or totally, inscrutable (and therefore, unknowable) by humans. Based on our exploration on the 'epistemic status' of AI output, **Sub-Section 3.2.** will examine the second series of the abovementioned conditions *i.e.* the accuracy standards for explanations pertaining to AI output. Going by the object of those explanations (*i.e*. the thing explained), we will focus our attention, first, on so-called *ad hoc* explanations (relative to the standards observed *a priori* in the inception of AI systems), second, on so-called *post-hoc* explanations (relative, in essence, to the reasoning patterns underlying harmful automated decisions and uncovered *ex post i.e*. once those decisions have been made). Our analysis of the 'accuracy' criteria applied for each of those two types of explanations will then allow us to critically examine the EU's regulation on AI liability and (finally) address the following issues: 1. whether said regulation - seeking to define systems of adjudication that would not leave litigants without effective judicial protection - allows for the adducing of evidence which can support *ad hoc* explainability, *post hoc* explainability or both; 2. whether the type of explanation required under said regulation takes into account what the *litigants themselves flag as necessary* for the purpose of making their views known and effectively participating in the adjudication of their disputes. Since the EU's AI liability regulation is not yet binding, there is no caselaw which can allow us to map out the procedural needs (in terms of evidence and explanations) that litigants have in disputes dealing with AI-related harm.

In **Section 4**, an examination of the evolving caselaw on AI liability, predominantly in North America, is presented to highlight pertinent procedural (and explanatory) needs. This analysis aims to delineate the types of understanding sought by litigants and courts in disputes related to AI. Drawing insights from specific, relevant cases, it becomes evident that the sought-after understanding in these disputes primarily revolves around two aspects. First (covered in **Sub-Section 4.1.**), there is a focus on the *accuracy of a given AI output*. The procedural concern here centers on whether there is sufficient evidence to ascertain the accuracy or inaccuracy of an automated decision. Second (explored in **Sub-Section 4.2.**), the attention shifts to the rationale justifying human reliance on the - potentially inaccurate and harmful - AI output. This raises the question of the *motives* having led a human agent to believe that an AI decision was accurate and, consequently, trustworthy. Our analysis of the emerging AI liability caselaw will allow us to identify *two trends* on the components of causal explanations: 1. XAI *is* integral to those explanations; 2. XAI should - ideally - be 'full' *i.e*. *ad hoc* and *post hoc*. Based on these conclusions, we will critically assess the EU's AI liability regulation which, from a procedural perspective, seems to restrict the scope of evidentiary debates in AI liability cases to *ad hoc* explainability only.

In our examination of the AILD and R-PLD, with a specific focus on the claimants' entitlement to seek evidence disclosure, the primary finding, highlighted in **Sub-Section 4.3.**, is that the evidence authorized under said instruments mainly supports *ad hoc* explanations. It reveals whether technical standards, particularly those outlined in the AI Act, were complied with in advance. Notably, there is an absence of provisions allowing litigants to receive *post-hoc* explanations, *i.e*. explanations on how a system *concretely* made a given harmful decision.

This will allow us to, in **Section 5**, express our criticism of the AILD and R-PLD. Based on our analysis of the emerging AI liability caselaw, our critique will translate to a plea to interpret (or amend) these instruments so that they may include the *procedural abilities* (to give and receive evidence and explanations) that the litigants in said caselaw flagged as necessary. For victims of AI-related harm, we will argue that *ad hoc* explainability is not enough: what claimants aim to understand when seeking compensation are the components of the causal link between an AI system's functionalities, reasoning patterns and output, on the one hand and a harm suffered, on the other hand. For that purpose, *post-hoc* explainability is paramount. Alternatively, for defendants, we will inquire if the AILD and R-PLD allow them to effectively exercise their right to defense. This inquiry is motivated by the fact that both the AILD and R-PLD charge the defendants with providing evidence (to claimants), but neither raises the question of whether the defendants themselves might need evidence to be disclosed (say, expertise) so that they may defend themselves more successfully against the claimants' allegations.

Drawing from the aforementioned points, our final remarks regarding the correlation between evidence, explanation, and procedural fairness are presented in **Section 6**.

I. ACCURACY OF EXPLANATIONS *TOUT COURT*

In adjudicatory contexts, evidentiary explanations are meant to provide *understanding* of the disputed facts, which explainees (such as courts and juries) are likely to view as accurate or at least, convincing. The epistemic question here is, of course, that of the *criteria* that ought to be met for explanations to qualify as 'accurate.' The ambition of this Section is to uncover those criteria, determine how they translate into law and lay the conceptual framework within which the remainder of this paper can take shape. To do so, **sub-Section 2.1.** will go back to the 'source' and delve into the concept of accuracy in connection to *scientific knowledge*, as the epistemic template (*genus*) for the concept of *explanatory* knowledge. Against the backdrop of our analysis of the knowledge/explanation kinship, **sub-Section 2.2.** will zoom in on *causal explanations* sought for the purpose of adequately representing and proving causality in law.

A. Scientific Knowledge, A Model for Explanatory Knowledge

Aspiring to be agnostic, epistemology abhors bias, one of its longstanding battles having been to 'cleanse' knowledge from preconceptions, beliefs and representations residual in the knowers' minds.[82] Knowledge - Bonderup Dohn insists - should "not just be employed as a black box term or be characterized only in its correlation with, for example, psychological states or social relations without being given an explicit analysis as regards its nature."[83]

But what is 'objectivity'? Minazzi posits that the term 'objective' "refers, in the first instance, to what exists as an object or to what possesses an object or, again, to

[82] Cartesian doubt is echoed here which consists in insisting that "the existence of a thought does not in itself guarantee the existence of what it purports to be a thought of." See Peter Winch, *Spinoza on Ethics and Understanding*, *cit. supra*, at 4.

[83] Nina Bonderup Dohn, "Epistemology in Investigating Knowledge: 'Philosophizing with" (2011) 4 *Metaphilosophy*, 431, at 431.

what belongs to an object."[84] In a similar vein, 'objective' - he argues - indicates both "everything which appears to be valid for everyone, and what appears to be independent of the subject, as well as everything which is 'external' with respect to consciousness of thought and, last but not least, everything which is found to comply with certain rules or methods."[85]

Minazzi's observations allow us to assert that, in the heart of any knowledge-construction endeavor lies the question of whether objectivity translates to *a*-subjectivity *i.e.* subject-independence (the 'subject' here being the person who acquires knowledge, not the person to whom knowledge is communicated).

A brief historic overview of epistemology reveals an original *penchant* for objectivism, characterized by the search for methods meant to 'cut off' (as it were) the knowledge of the outside world from the knower's inner world (**Sub-Section 2.1.1.**). This current was, however, contrasted by subjectivism (**Sub-Section 2.1.2.**) which was eventually - and, in some cases, reluctantly - accepted as unavoidable. Scholars came to realize that, try as they might, 'valid' knowledge could never be fully divorced from belief; an observation which applies *a fortiori* to explanations for one simple reason: their accuracy does not solely depend on the explainor's epistemic and communicative competence. It also, if not mostly, depends on the explainee's *ability to understand* the gist of the explanation given.

1. The Ideal(ized) Objectivism

Enlightenment philosophical traditions - namely Newtonian physics and Kantian transcendental philosophy[86] - gave us the analytical benchmarks we turn to, in order to develop our normative understanding of what 'pure' science is. The methods of scientific discovery and the criteria used for the validation (or invalidation) of knowledge began in the natural sciences, subsequently shaping the epistemology in the social sciences,[87] including law (especially, the law of evidence).[88]

Throughout its evolution and the constant fine-tuning of the criteria for 'true' or 'valid' knowledge, epistemology maintained its original posture of agnosticism, the idea being that knowledge ought to include *belief-independent accounts* of the world and not be 'corrupted' by the knower's *representations* thereof (**A**).

This *penchant* for objectivism is particularly visible in the verificationist strands on explanations. However superficial they might seem - compared to the cognitive depths that knowledge proper aspires to reach - explanations remain *fact-correspondent* that is, pertain to an *object of explanation* that is material, tangible and verifiable (**B**).

---

[84] Fabio Minazzi, *Historical Epistemology and European Philosophy of Science* (Springer, 2022), at 3.
[85] *Id*., at 3-4.
[86] See e.g. Michael Friedman, "Newton and Kant on Absolute Space: From Theology to Transcendental Philosophy" in Michel Bitbol, Pierre Kerszberg, Jean Petitot (ed.), *Constituting Objectivity. Transcendental Perspectives on Modern Physics* (Springer, 2009), 35-50.
[87] For the translation of the 'scientific method' in sociology (the seminal figure of which is, of course, Durkheim), see Enzo Di Nuoscio, "L'individualisme méthodologique comme méthode scientifique: théorie de la rationalité, explication causale, herméneutique" (2020) 70-1, *L'année sociologique*, 129.
[88] The process of giving and assessing legal evidence has been labelled as 'courtroom epistemology.' See Baosheng Zhang, Jia Cao, David R.A. Caruso, "The Mirror of Evidence and the Plausibility of Judicial Proof" (2017) 21 *Int'l J. of Evidence & Proof*, 119, at 123.

a. The Belief-Independence of Knowing

'Proper' knowledge - often synonymized with 'scientific' knowledge[89] - is meant to somehow capture the essence of the portions of reality it pertains to. In its purest, most idealized flavour, it is meant to uncover the "exceptionless laws"[90] that govern the phenomena that fall in the scope of our experiences of reality. In the backdrop of this ideal, it is not surprising that objectivity has been historically fetishized, fostering hostility toward the belief-ridden persona of the knower, 'belief' being usually seen as an irrational creed, held "for a reason which is preposterous or for no reason at all."[91]

Is it possible for someone to pursue knowledge of reality without being emotionally and cognitively tainted by their beliefs? Isn't knowledge itself a set of - as Keynes put it - *rational* beliefs[92]? The belief/knowledge interplay has been a recurring theme in savant circles, which offered varying views on the *posture(s) of the knower* and the *models of reality*. Putnam seminally argued that three important (meta) traditions addressed these issues: "the extreme *Platonist position* which posits non-natural mental powers of directly 'grasping' forms (…) the *verificationist position* which replaces the classical notion of truth with the notion of verification or proof and there is the *moderate realist position* which seeks to preserve the centrality of classical notions of truth and reference without postulating non-natural mental powers."[93] Each meta tradition has given way to numerous sub-strands, the detailed accounts of which fall - alas - outside of the scope of this paper. For the sake of brevity, let us refer to Minazzi's work on *epistemic objectivity*,[94] a point on which he sought guidance in Kant's work.

Kant's brilliant philosophy arguably made two major contributions to the ways in which we understand and construct (objective) knowledge. First, that discovery of

---

[89] Considering traditional epistemology is characterized by three central notions namely knowledge, belief and doubt, securing a level of stability of knowledge appeared as a process of responding to skepticism while, at the same time, creating models of 'valid' epistemic models (*i.e.* models able to reliably deliver knowledge). See Vincent F. Hendricks, John Symons, "Where's the bridge? Epistemology and Epistemic Logic" (2006) 128 *Philosophical Studies*, 137, at 138-139.

[90] We borrow here Putnam's expression used in her comment of Quine's (post-Kantian) view on - what she called - *analyticity*, essentially derived from Kant's concept of analytic judgments. See Hilary Putnam, *Realism and Reason. Philosophical Papers* (vol. 1, CUP, 2010), at 89.

[91] John Maynard Keynes, *A Treaties on Probability* (Macmillan & Co., 1921), at 10.

[92] "Knowledge of a proposition - Keynes writes - always corresponds to certainty of rational belief in it and at the same time to actual truth of the proposition itself. We cannot know a proposition unless it is in fact true." John Maynard Keynes, *A Treaties on Probability*, *cit. supra*, at 11. Keynes' observation is interesting for mainly two reasons. On the one hand, he views knowledge as propositional (knowledge consists of propositions about reality). On the other hand, he dissociates certainty and truth as if to distinguish a justified belief of accuracy (certainty) from accuracy *tout court* (truth). We will not further discuss this distinction, however interesting and relevant it may be for our discussion on the epistemic ideal of objectivism and the epistemic 'tolerance' of subjectivism. Keynes wrote a seminal (though a bit dated) work on probability and defined knowledge-as-certainty so that he could then delve into the concept of probability. However brilliant, he is not in the forefront of strands on explanation, which are the main focus of this paper.

[93] Hilary Putnam, *Realism and Reason. Philosophical Papers*, *cit. supra*, at 1 (emphasis added).

[94] Minazzi raises the longstanding and complex question of whether scientific knowledge can really be value-free? The assumption here is that science is apolitical and acultural knowledge production activity. However, Minazzi ultimately concludes that scientific knowledge and its production are rooted in "a stratified social reality that may produce different images of the human knowledge itself. See Fabio Minazzi, *Historical Epistemology and European Philosophy of Science*, *cit. supra*, at 121.

knowledge is usually not serendipitous, but the result of highly protocolized epistemic processes.[95] Second, under Kant's influence, we place the inferences drawn from our discoveries on a "new heuristic plane of *transcendentality*, by which Kant constructs the overall theoretical framework of his epistemological meta-critic reflection, deeply innovating not only the whole concept of knowledge, but also the style and modes of human rationality."[96] Scientific knowledge is 'scientific' because science is "always capable of thinking its object by constructing it through a plastic critical interplay of continuous comparison with the experimental dimension." [97] In other words, knowledge is produced *within* the confines of already existing conceptual frameworks, where well-established (and constantly perfected) sets of epistemic competences are deployed.[98]

In addition to Minazzi, other contemporaries expressed similar intuitions. Latour seminally expressed the view of trials and experimentation as being "ritual frameworks" with value hierarchies that 'actants' (which include humans as well as, say, microbes) obey in the fabrication of 'scientific facts.'[99] In this context, knowledge proper can be understood as an *adept belief*[100] - the word 'belief' again! - for which an epistemic community considers there to be *sufficient reasons* to hold it as true, at least until more conclusive, belief-dispelling evidence is brought forward.[101]

Though Kant - and others - molded our modern understanding(s) of scientific epistemology, pushing us to sharpen our intuition on what *true* science is, the subjectivism/objectivism dilemma was not altogether effaced from epistemic discourse. To this day, an opposition remains between *materialists* who view facts as the sole gateways to agnostic truth and *mentalists* for whom, knowledge formation carries the

---

[95] Duede writes: "scientists do not *design* the physical processes. Rather, they, as it were, *discover* them. With theory mediated instruments, nothing is out of our hands." Eamon Duede, "Instruments, agents, and artificial intelligence: novel epistemic categories of reliability" (2022) 200 *Synthèse*, 491, at 501.

[96] Fabio Minazzi, *Historical Epistemology and European Philosophy of Science*, *cit. supra* at 11 (emphasis added).

[97] *Id*., at 12.

[98] See e.g. Susanne Mantel, ''Acting for reasons, apt action and knowledge'' (2013) 190 *Synthèse*, 3865, at 3873.

[99] See Kyle McGee, *Bruno Latour: The Normativity of Networks* (Routledge, 2014), at 4.

[100] John Turri, ''Manifest Failure: The Gettier Problem Solved'' (2011) 8 Philosophers, 11 *cit. in* John Greco, "A (different) virtue epistemology" (2012) 1 *Phil'y & Phenomenological Res.*, 1, at 9.

[101] This type of belief-forming epistemic practices (and the virtues or values associated with those) were examined by Sosa in his study on reflective knowledge (essentially focused on the reliability and criteria used to label something as 'knowledge') as opposed to 'animal' knowledge,' which is mostly perceptual, experiential with no ambition to systematize a set of protocols and procedures meant yield Sosian 'apt' beliefs. See Ernest Sosa, *Reflective Knowledge: Apt Belief and Reflective Knowledge* (vol. II, OUP, 2009), at 135 seq.

imprint of the knower's 'mental states' (social contexts, background knowledge and preexisting values and beliefs).[102]

Epistemology's aspiration to cut the umbilical cord between knowledge proper and psychology has transpired into modern evidence scholarship.[103] Of course, the knowledge derived from legal evidence has never been held to the validity standards of science. Nevertheless, the requirement for objectivity has woven through schools of thought on evidence, which can be perceived as unrealistic. When litigants give evidence in a trial, they do so as adversaries, confronting their *versions of the disputed facts* with the goal of winning the case. The answer to the question 'what happened in a dispute?' is, in a way, doomed to be subjective since "it's not about truth, it's about who tells a better story."[104]

However, we mentioned earlier that legal evidence is a peculiar beast,[105] namely because the adducing of evidence should both *be fair* and serve *the purpose of fairness* (*i.e.* a fair resolution of a dispute). It is precisely because of this 'fairness constraint' that the epistemology of legal evidence has - heavily! - drawn inspiration from scientific discovery methods. The idea is that 'adequate' (*i.e.* impartial, politically, culturally and socially neutral and therefore fair[106]) administration of justice requires *some level* of objectivity in the ways in which facts are given and assessed. In this context, a law of evidence is typically meant to (at least minimally) define basic epistemic conditions under which litigants can debate facts and do so before an unbiased authority.

By defining the features of various types of evidence (admissibility, probative value, standards of proof) and the requirements for fact-appraisal (impartiality, legal expertise, fairness), a law of evidence does not establish a scientific discovery-type proceduralism conducive to measurable, verifiable and reproducible results. It does, however, provide a set of *procedural guarantees* meant to preclude evidential truths from depending on the whims of litigators, judges and juries. Those guarantees (mainly linked to the parties' equal opportunity to plead and the courts' independence) warrant,

---

[102] Scientific truths are essentially beliefs held as true or beliefs for which there are good, or valid reasons to accept as true. Beliefs stand so long as they are justified which, of course, begs the question of the conditions that warrant justifiability. In an outline of the main schools of thought within epistemology, Bishop and Trout distinguished three: foundationalism, coherintism and reliabilism. The first two are internalist theories of justification, in the sense that the 'justifiers' for holding a belief as true are accessible to the believer. Foundationalists - Bishop and Trout argue - hold that "many beliefs are justified in terms of their relations to other beliefs." This presupposes a set of basic beliefs that act as 'normative justifiers' of sorts and in reference to which subsequent beliefs are assessed. Coherentists are a spin-off from foundationalists: they also consider that what beliefs can be justified in terms of their relations to other beliefs, but coherentists deny the existence of basic beliefs. For reliabilists, the justifier is external: a belief is justified in case it is produced by a reliable belief-forming mechanism. See Michael A. Bishop, J. D. Trout, "Epistemology's search for significance" (2003) 15 *Journal of Experiential & Theoretical Artificial Intelligence*, 203, at 205.

[103] Modern evidence theory roughly includes the past 200 years of scholarship. See namely Douglas Walton, *Legal Argumentation and Evidence* (Penn. State. Univ. Press, 2002), at 106.

[104] Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication" (2018) 26 *AI & Law*, 345, at 347.

[105] See *supra*,

[106] As May put it, "procedural justice conveys the idea that everyone will be subject to and protected by the same rules. Each person is to be seen as equal before the law." Larry May, *Global Justice & Due Process* (CUP, 2011), at 13.

if not the certainty, at least the *expectation* that the law, impersonal and fair, will *deliver justice*.[107]

Our brief *exposé* on subjectivism and objectivism as debated among epistemologists and as taken over by procedural lawyers, sets the tone for our analysis of *evidentiary explanations*, the accuracy of which is also characterized by a quest for balance between independent (impersonal) standards and subjective beliefs. Against this backdrop, we can - finally - raise the questions we wish to address in this sub-section: 1. what is explanation?; 2. what is an accurate (good) explanation? Though straightforward answers are hardly possible, we will - in a pedagogical *élan* - distinguish two definitions, one we will call static (the act of explaining) the other, dynamic (the process of explaining).

b. The Fact-Correspondence of Explaining

In a static sense, an explanation is, in essence, an *interpretation of experience*: to explain is to provide meaning of specific objects[108] in understandable terms.[109] Of course, for explanatory interpretation to be possible, the object of explanation should be interpretable, interpret-*ability* being the feature of the object explained to acquire concrete meaning.[110]

In AI jargon, interpretability and explainability are often used interchangeably though Hauque *et al.* view them as conceptually distinct. Explainability, they argue, "means explaining the decisions made by machine models in a human-understandable form."[111] Alternatively, interpretability "is the explanation of how or why a model resulted in a particular prediction."[112] However plausible this distinction may be, we will consider, in the remainder of this paper, that *any* explanation (in the field of AI or not) is inherently interpretative. Indeed, to interpret an AI system's decisional process (what Hauque *et al.* call 'interpretability') is to provide the basis for an explanation of its output (explainability *stricto sensu*). Though there might be a semantic or a theoretical interest in distinguishing the two, for the purpose of this study, we will consider interpretability (*i.e*. a system's aptitude to be interpreted) as an epistemic

---

[107] There has been much debate on whether a law of evidence (as a consolidated *corpus* of rules framing evidentiary epistemology) can legitimately exist only if it is codified or it can also emerge from court practice. In an reductionist attempt, Wróblewski argued that a law evidence can be viewed as existing if it includes rules and principles which answer *four essential questions*: how does law distinguish between facts that require evidence from those that do not?; which evidence is admissible?; how is evidence assessed?; what is the role of evidence in the performance of (judicial) review?. See Jerzy Wróblewski, « La preuve juridique : axiologie, logique et argumentation » *in* Chaïm Perelman, Paul Foriers (ed.), *La preuve en droit : études* (Bruylant, 1981), 331, at 338.

[108] William Franz Lamberti, "An overview of explainable and interpretable AI' in Feras Baratesh, Laura Freeman (eds), *AI Assurance. Towards Trustworthy, Explainable, Safe and Ethical AI* (Elsevier, 2022), 55-123, at 57.

[109] Ricardo Guidotti, Anna Monreale, Dino Pedreschi, Fosca Giannotti, "Principles of Explainable Artificial Intelligence" *in* Moamar Sayed-Mouchaweh (ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications* (Springer, 2021), 9, at 12.

[110] *Ibid*.

[111] Bahalul Haque, Najmul Islam, Patrick Mikalef, « Explainable Artificial Intelligence (XAI) from a user perspective : A synthesis of prior literature and problematizing avenues for future research" (2020) 186 *Technological Forecasting & Social Change*, 1, at 2-3.

[112] *Ibid*.

precondition for explainability proper (*i.e*. interpretation given on the system's functionalities and decisional/predictive processes).

As with any type of interpretation, the risk with explanations is that of *mis*interpretation. Badea and Artus[113] call this the interpretation problem (IP). The threat of IP calls for caution because virtually any real-world occurrence can be interpreted in infinite and unspecifiable ways. In the field of AI, the IP arises - Badea and Artus argue - because of the possibility that "a highly advanced machine may find novel interpretations of the rules that we give it, interpretations which are not incorrect, in that they can be seen as valid interpretations of the rule, but which are inappropriate in that we do not approve of them."[114] As a mitigation strategy to the IP (in the field of AI, and in general), explanation theorists sought to define *basic accuracy criteria* which can be clustered in two families: those - leaning toward objectivism - that support explanatory *fact-correspondence* or *facticity* and those - subjectivist-prone - that support *understandability*.

Regarding facticity, the theoretical referent we will use is the so-called *correspondence* theory of truth, which upholds the view of "agreement or correspondence between a statement and the so-called facts or reality."[115] It should be mentioned that correspondence theory does not eradicate subjectivism altogether; its gist consists in preferentially using perceptive reality as 'the' referent for the validity of propositions made about that reality. For instance, if one wishes to know if they may justifiably assert that snow is white, they ought to see the color of snow falling (*i.e*. turn to reality to verify the truth or falsity of the 'snow is white' statement).

That explanations should be in accord with tangible facts does not raise any particular controversy: if this was not the case, there would be next to no difference between explaining and the "narrative techniques of imaginative writers."[116] In adjudicatory contexts, explanations' *rattachement* to reality is paramount precisely because it *enables verification*: when courts are called to resolve disputes, they strive to acquire, from the parties, *accurate* knowledge of facts so that they may draw relevant conclusions on important legal (and by that, social and political) issues like guilt or liability.[117]

---

[113] Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents," *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (2022), SGAI-AI, AI XXXIX, 124.

[114] *Id*., at 125.

[115] Carl G. Hempel, "On the Logical Positivists' Theory of Truth," *in* Richard Jeffrey (ed.), *Selected Philosophical Essays* (CUP, 2012) 9, at 9. Hempel opposes the correspondence truth theory to the coherence theory of truth, according to which "truth is a possible property of a whole system of statements." See *ibid*. Exploring the relevant ways in which correspondence and coherentism are similar, complementary or opposed is beyond the scope of this paper. We refer to the correspondence-theory as a theoretical referent allowing us to make the following point: if explanations are taken as statements *about facts*, they ought to relate to those facts, 'facts' being understood as tangible, perceptible, verifiable *objects of experience*. The choice of correspondence theory is important because it accepts that facts are facts, regardless of whether there are propositions made about them (this is the coherentist thesis according which - if we were to vulgarize it - there are no facts *per se*, only propositions about facts). Whether the explanation-fact correspondence is well-established (adequate or credible) is an issue of assessing the *conditions* under which truth-as-correspondence can stand as acceptable (and therefore accurate). These conditions will be discussed further in this paper.

[116] Simon Stern, "Factuality, Evidence and Truth in Factual Narratives in the Law," *cit. supra*, at 391.

[117] *Id*., at 392.

Referring to Di Bello's probabilistic analysis of criminal trials,[118] Urbaniak stressed that "the relationship between evidence and (evidentiary) narratives goes both ways: from the evidence to the narratives and from the narratives to the evidence."[119] There is something intuitively convincing about this interplay: evidence is both the *foundation* for a narrative about facts and the *standard* against which the validity of the truth of that narrative is assessed. While narrations - Urbaniak writes - play a "crucial role in the account, their relation to evidence and their factual support is also in the focus, hopefully susceptible to a more precise probabilistic analysis."[120] In laymen's terms, when we say that something is true or false in an adjudicatory context, 'something' is usually a state of fact.

The need for explanations to be fact-correspondent implies that they are *context specific*[121] and *factive*.[122] The role of context in assessing the explanatory goodness (understood here as a 'thin' concept of accuracy) will be discussed further. At this stage, may it suffice stressing that facticity is, indeed, the unavoidable but not exclusive referent for the assessment of said goodness. We do not *explain* gravity simply by advising someone to drop a pen. By dropping the pen, they *experience gravity*, but do not gain understanding on what it is and why it works the way it does. All explanatory contexts include an actor who ultimately says 'yay or nay' on the accuracy/plausibility reached (or not) by the explanation given. Enter the figure of the explainee.

2. The Unavoidable Subjectivism

As mentioned earlier, epistemology's aversion to belief has been somewhat 'diluted' in contemporary scholarship. Rarely does a fact speak for itself, declaring - as it were - a truth about the world irrespective of an observing knower's perceptions and beliefs. For example, regardless of how one feels about water's boiling point, it will always be reached at 100 °C. Even propositions (like, 'the sky is blue') which we as laymen view as uncontroversial, have triggered erudite debates on the conditions under which those propositions could be held as true (obviously, because the sky is not always blue)… Our point is the following: any type of knowledge is to some degree belief-dependent: a proposition (hypothesis, theory, explanation…) about a state of facts is true to the extent, and so long as relevant expert and/or non-expert communities believe it to be. In explanatory contexts, *believability* does, indeed, appear to be the apex standard for explanatory accuracy (**A**), assessed by the explainee in reference to the context in which they receive a specific explanation (**B**).

a. Believability as Proxy for Explanatory Accuracy

Delivering understanding, as the purpose of any explanation, allows us to tackle the above-mentioned *dynamic definition* (*i.e*. explaining as a process). Explanations are

---

[118] Marcello Di Bello, *Statistics and probability in criminal trials*, Ph.D. Thesis (University of Stanford., 2013).
[119] Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication" (2018) 26 *Artif. Intell. Law*, 345, at 348.
[120] *Ibid*.
[121] Michael Ridley, "Explainable Artificial Intelligence (XAI)" (2022) 2, *Information technology and libraries*, 1, at 3.
[122] Andrés Paez," The Pragrmatic Turn in Explainable Artificial Intelligence" (2019) 3 *Minds and machines*, 441, at 454.

communicative acts: in most cases, something is explained by someone, to someone.[123] Bearing in mind standard knowledge-construction theories, we tend to place our focus on what the explainer ought to do to deliver clear and accurate information. But communication is a two-way street, the level of comprehensibility of the explanation given depending also (if not predominantly) on the explainee's *ability to understand*.

This ability is largely shaped by the explainee's prior knowledge and experience. For instance, a flat-earther will likely discard the many photos taken from space showing the Earth's spherical shape. Those photos would presumably be dismissed as untrustworthy evidence in the face of a person's unwavering belief that the Earth is, in fact, flat. The point we seek to make through our flat Earth example is the following: though anchored in facts, explanations will always be viewed through the lens of their addressees' beliefs and because of this, they will likely fall on biased ears.[124] In the context of AI, the trustworthiness of explanations (pertaining to, say, the probability that a system develops a gender bias) will largely depend on whether explainees look favorably on AI to begin with.[125]

In examining the explanatory process through the vantage point of communication, Ridley[126] highlighted three features all explanations share. First, they are *contrastive*: when people want to know the 'why' of something, they "do not ask why event *p* happened, but rather why event *p* happened *instead* of some event q."[127] Second, they are *selected*: people are adept at "selecting one or two causes from a sometimes infinite number of causes to be *the* explanation."[128] Third, explanations are *social*: "they are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs."[129]

The fact that an explanation is given *to* someone places, on the explainer, the duty to deliver, to the best of their ability, adequate understanding of the object explained, 'adequate' explanations being, in essence, those that manage to warrant *believability*. It can even be argued that believability is for explanations what accuracy *strico sensu* is for knowledge proper. According to Paez, this believability-as-proxy-

---

[123] Denis J. Hilton, "Conversational processes and causal explanation" (1990) 1, *Psychological Bulletin*, 65, at 65.

[124] This is of course not the least bit surprising, considering that comprehensibility is, typically, a matter of making associations between what a person views as true and what is, to them, new information. As Moehring *et al.* stress, the 'comprehension construct' is a process of developing mental representations, by which prior long-term knowledge is incorporated with the available information." See Anne Moehring, Ulrich Schroeders, Benedikt Leichtmann, Oliver Wilhelm, "Ecological momentary assessment of digital literacy: Influence of fluid and crystallized intelligence, domain-specific knowledge, and computer usage," (2016), 59 *Intelligence*, 170, at 171.

[125] In Vered *et al.*'s excellent work, we find interesting empirical studies (and corresponding inferences) on the interrelationship between explainability of AI and reliance on automated decisions. Based on several empirical studies in radiology, the authors conclude that local and global explanations tend to decrease over-reliance, decreasing the explainees' automation bias. See Mor Vered, Piers Douglas Lionel Howe, Tim Miller, Liz Sonenberg, "The effects of explanations on automation bias" (2023) 322 *AI*, 103952.

[126] Michael Ridley, "Explainable Artificial Intelligence (XAI)," (2022) 41-2, *Inf. Tech'y & libraries*, 1, at 4.

[127] *Ibid*.

[128] *Ibid*.

[129] *Ibid*.

for-accuracy is due to the fact that understanding is not strictly speaking knowing[130] which *a fortiori* suggests that explaining is not strictly speaking discovering.

Paez's intuition is on point. As mentioned earlier, (scientific) knowledge is 'knowledge' because it is "supported by protocol statements"[131] and accepted as such by communities who share the same epistemic competences. Echoing verificationist[132] and Latourian views on knowledge-construction, scientific experiences are often highly proceduralized, the threshold of accuracy (that is, justified acceptance of beliefs) being usually quite high.[133] Because of this, scientists can test accepted beliefs on a continued basis, constantly revisiting the reasons why a theory should remain acceptable.[134] The 'knowledge' explanations provide is of a slightly different kind. They are not issued from discovery *per se*. They rather provide "a kind of packaged summary of the relevant events; and if successful, this summary allows us to make appropriate inferences of the situation."[135] For an explanation to be qualified as 'good,' the golden rule seems to be '*know thy audience*.'[136]

But here, an interesting question emerges: how do facts support explanatory believability? If believability on the side of the explainees is, indeed, the workable variant of accuracy applied in explanatory contexts, do we still need evidence to support the explanation's fact-correspondence? In other words, are explanations 'accurate' only when the explainees believe them to be so, regardless of the interpretations warranted by facts? Subjectivism again rears its ugly head and its 'threat'[137] should not be underrated, given that - as mentioned earlier - explanations, like any form of knowledge, are not pulled out of thin air, but must have *some* anchoring in reality. We thus circle back to the debate previously canvassed on subjectivism and objectivism as the two points of oscillation of modern conceptions of epistemic accuracy. Explanations have not been spared from this debate, as confirmed by representatives of several '-ism' strands.

---

[130] Andrés Paez, "The Pragmatic Turn in Explainable Artificial Intelligence" (2019) 29-3, *Minds & machines*, 441, at 453.
[131] Carl G. Hempel, "On the Logical Positivists' Theory of Truth," *in* Richard Jeffrey (ed), *Selected Philosophical Essays* (CUP, 2012), 9, at 9.
[132] Hilary Putnam, *Realism and Reason. Philosophical Papers*, *cit. supra*, at 89.
[133] A nuance and a clarification should be provided here. The nuance is that no knowledge, be it scientific of explanatory, is absolutely accurate. As Hempel rightly put it, "nowhere in science will one find a criterion of absolute unquestionable truth." See Carl G. Hempel, "On the Logical Positivists' Theory of Truth," *cit. supra*, at 16. The just like the process of scientific discovery is protocolized, the accuracy is as well, in the sense that, what is to be viewed as valid (in the sense of accurate) knowledge is a matter of convention among the members of epistemic communities. This is the effect of (conventionally agreed upon) epistemic norms, which "identify the conditions under which someone should or should not believe, do, or feel something." See Clayton Littlejohn, "Objectivism and Subjectivism," *in* Veli Mitova (ed.), *The Factive Turn in Epistemology* (Cambridge University Press: 2018), 142, at 142.
[134] This is what Minazzi calls "radical critical discussion." See Fabio Minazzi, *Historical Epistemology and European Philosophy of Science*, *cit. supra* at 154.
[135] Laura Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference from Explanation" (2022) 151-7, *J. of exp'l psy'y*, 1481, at 1482.
[136] Michael Ridley, "Explainable Artificial Intelligence (XAI), Information technology and libraries," *cit. supra*, at 3.
[137] Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication," *cit. supra*, at 347.

Evidential *explanationism* is associated with Allen and Pardo[138] who opposed the prevailing probabilistic current in contemporary evidence scholarship,[139] Bayesian probabilities[140] being a prominent school of thought within that scholarship. For probabilists, accuracy is usually function of: 1. the quality of the items of evidence presented in support of a given claim and 2. the individual (and numerically represented) probative value given to each item (e.g. A is true with probability of 0.52/1).[141] In contrast, explanationists view evidence as discursive and "inherently comparative - whether an explanation satisfies the standard depends on the strength of the possible explanations supporting each side."[142] Because of this, explanationist accuracy is, indeed, synonymous with believability: it is not about meticulously measuring probabilities, but about the (non-quantifiable) levels of persuasiveness an explanation can generate. This is understandable, considering how impractical it is to expect factfinders to "actually attach probabilistic numbers to each probability at issue in litigation."[143]

Evidentialism also leans toward a more subjectivist view of explanatory accuracy. Seminally represented by Conee and Feldman,[144] the gist of this current is that "the epistemic justification of a belief is determined by the quality of the believer's evidence for the belief."[145] In truth, Connee and Feldman (re)state longstanding strands in epistemology on the conditions under which beliefs can be justifiably held as valid (*i.e*. taken as true until rebutted). Thinkers like Locke, Hume, Reid and Bentham have long ago "championed or at least anticipated evidentialism."[146] Much like explanationists, evidentialists do not suggest some metric system that would allow us to numerically represent the truth value of explanations. They remain 'subjectivists': Conee and Feldman even called themselves mentalists,[147] positing that the justification

---

[138] Ronald J. Allen, Michael S. Pardo, "Relative plausibility and its critics" (2019) 23-1/2, *The Int'l J. of Evidence & Proof*, 5. The authors argue that the explanationist approach to legal evidence presents advantages that the probabilistic approach does not offer, namely "(1) the need to assign number values to compare the standard of proof; (2) lack of fit between the probabilistic theory and how fact-finders actually evaluate and reason with evidence; (3) inconsistency with legal doctrine and jury instructions (the conjunction problem); and (4) inconsistency with regard to the policy goals underlying standards of proof." See *id*., at 17.

[139] Under this probabilistic view, evidence is represented as *measurable* assessment of the likelihood of the disputed facts. See, for instance, Paul Horwich, *Probability and Evidence* (CUP, 2016), at 100 seq.

[140] Urbaniak provides a concise summary of Bayesian theory. Standard Bayesian epistemology "represents degrees of belief (also known as credences) by real numbers. Degrees of belief of an ideally rational agent, on the standard view, should satisfy the standard axioms of probability: probability should take values between 0 and 1 inclusive, logically impossible events get probability 0, logically certain events have probability 1, and the probability of the union of finitely many disjoint events is the sum of their individual probabilities (in the context of this paper, whether this holds also for infinite unions will not come up)." Rafal Urbaniak, "Narration in judiciary fact-finding: a probabilistic explication," *cit. supra*, at 353. For an analysis of the application of Bayesian theory in the field of legal evidence, see, *inter alia*, Terence Anderson, David Schum, William Twining, *Analysis of Evidence* (CUP, 2009) at 246 seq.

[141] Johan B. Gelbach, "It's all relative: Explanationsim and probabilistic evidence theory" (2019) 1-2 *The Int'l J. of Evidence & Proof*, 168, at 171.

[142] Ronald J. Allen, Michael S. Pardo, "Relative plausibility and its critics," *cit. supra*, at 15.

[143] *Ibid*.

[144] Earl Connee, Richard Feldman, *Evidentialism: Essays in Epistemology* (Oxford University Press: 2004).

[145] *Id*., at 83.

[146] Philipp Berghofer, *The Justificatory Force of Experiences* (Springer : 2022), at 69.

[147] Earl Connee, Richard Feldman, *Evidentialism: Essays in Epistemology*, *cit. supra*, at 99.

of a belief that evidence is true largely depends on the "totality of one's mental states"[148] which are "drawn from *our experiences* as points of interaction with the world."[149] Conscious awareness - they write - is how "we gain whatever evidence we have."[150] Consequently, "much of what we know about the causal structure of the world we infer from directly observing and interacting with it."[151]

In sum, 'accurate' explanations are believable based on a *level of coherence* between the evidence given by the explainor and the explainees' residual beliefs. What reinforces this coherence is the *context* within which explanations are given. Indeed, as in most real-life situations, for explanations too, context is everything.

b. The Benchmark for Believability: Context is Everything

Is there an independent standard against which explanatory believability can be assessed? A definitive answer is next to impossible to give. Generally, evidentialists allude to *shared or common experience* or - to be more exact - *conventional interpretations of reality*.[152] Scholars have called this the *justificatory role of experience*.

Common experiences form - to paraphrase Aristotle - the realm of *doxastic knowledge*:[153] not knowledge *per se*, but a form of 'common wisdom' derived from experiences shared within given communities. Doxa gives people a sense of normalcy, a state of affairs where certain facts (e.g. children born in wedlock are fathered by their mothers' spouses) are accepted as true because they are perceived as 'normal.' In the case of AI, *no one* would ask for an explanation on how an AI system became gender-biased, if that bias was not viewed as a deviation from what the explainees view as a normal state of reality. Such a bias would be perceived as an error, the conventional belief - though often dispelled - being that unfair biases have no place in a world where equality should be the social and legal norm.

The concept of normality is a can of worms in its own right, usually defined through two main versants: *descriptive* (normality derived from the repetition of events) and *prescriptive* (state or conduct resulting from convention).[154] In causal contexts, Kirfel *et al*.[155] confirm through empirical data what Hart and Honoré[156] had previously claimed in legal theory - people tend to designate *abnormal events* as causes of harm:

---

148 Philipp Berghofer, *The Justificatory Force of Experiences*, *cit. supra*, at 70.

149 Earl Connee, Richard Feldman, *Evidentialism: Essays in Epistemology*, *cit. supra*, at 87.

150 *Ibid*.

151 Lara Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference From Explanation" (2022) 7 *Journal of Experimental Psychology*, 1481, at 1482.

152 We allude here to Michalski's definition of experience as the totality of information generated in the course of performing some actions. See Ryszard S. Michalski, "Inferential Theory of Learning as a Conceptual Basis for Multistrategy Learning" (1993) 11 *ML*, 111, at 116.

153 Doxa, as a form of conventional wisdom or a realm of 'truisms' (but not capital 'T' truth) has been correlated with common sense, as a baseline knowledge derived from common experience. See e.g. Georges Molinié, "Doxa et légitimité" (2008) 2 *Langages*, 69. Pietsch also, evoked common intuitions about causality, referring to causal mechanisms thought to be relatively well understood and unambiguous. See Wolfgang Pietsch, *On the Epistemology of Data Science*, *cit. supra*, at 127.

154 Elsa Bernard, *La spécificité du standard juridique en droit communautaire* (Bruylant, 2010), at 37.

155 Laura Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference from Explanation," *cit supra*.

156 H.L. A. Hart, Tony Honoré, *Causation in the Law* (OUP, 1985).

"when two causes are each necessary for producing a certain outcome (conjunctive structure), people judge the abnormal event as more causal."[157]

What is 'normal' and 'abnormal' in the context of AI is open for debate. As we will argue further, the EU's substantive and procedural regulation of AI refers to 'normalcy' by using expressions like 'reasonable foreseeability,' 'intended purpose' (of an AI system), 'foreseeable use (of an AI system) etc. May it suffice stressing, at this stage, that in searching for 'the normal' in connection to AI, scholars' and regulators' reflex was not to focus on descriptive normalcy, but to explore the tenants of a 'new' prescriptive or axiological normalcy. In this context, a 'normally functioning' AI would be one whose output would comply with a given community's foundational axiological framework.[158] In AI jargon, value-conformity is a component of AI accuracy: AI output is 'correct' if it is both statistically accurate (efficacious) and compliant with values labelled as unwavering or norm-setting (effective).

As a flourishing AI scholarship confirms, this stats-meet-values approach to AI accuracy is not the least bit surprising: "new technologies and new forms of human action are always creating moral dilemmas which didn't exist before, which force us to make judgments about how such rules as 'do no harm' apply, and how we interpret or apply the rule in any novel case can only be determined by values external to our rule, values which our rule is in principle incapable of embodying unambiguously."[159] Values,[160] Badea and Artus argue "should be explicit and efficacious, that is, be directly present in the agent's reasoning, and have a material impact upon the decision making of an agent in any relevant situation it acts in. We could then have the agent prioritize these moral goals over practical goals, ensuring that the former are not overruled by the latter."[161] In light of this, the authors suggest that "we adjust the causal power we build into an agent in the design process to the amount which we believe our reasoning mechanisms can successfully handle."[162] If only it were that simple…

AI explainability (and the possibility thereof) are a tricky matter which we will discuss at a later stage in this paper. At this juncture, and after having explored - albeit in broad brush strokes - the objectivist and subjectivist views on explanatory accuracy, a few observations should be made on the importance of explanatory contexts. Indeed, to deliver *good* explanations, explainers should be aware of the intellectual and

---

157 Laura Kirfel, Thomas Icard, Tobias Gerstenberg, "Inference from Explanation," *cit. supra*, at 1489.

158 Axiology is a (vast) field of study with various currents and views on what values are. For the purpose of this paper, the operative understanding of 'value' will be that suggested by Brey, who argued that values correspond to "idealized qualities or conditions in the world that people find good." See Philip Brey, "Values in technology and disclosive computer ethics," *in* Luciano Floridi, *The Cambridge Handbook of Information and Computer Ethics* (CUP, 2012), 41-58, at 46.

159 Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents" (2022) XXXIX International Conference on Innovative Techniques and Applications of Artificial Intelligence (SGAI-AI), 124, at 127.

160 Badea and Artus defined values as "high-level concepts that are relevant considerations during decision making. These could be virtues, character traits ('honesty'), or concepts that are of moral importance ('property') or even morally neutral practical considerations. We argue that values are the tether to the external point of the game, crystallizing what we want from the behaviour of the agents in the game, or in the moral situation. This is supported by arguments from Virtue Ethics." See Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents," *cit. supra*, at 135.

161 *Id*., at 133.

162 *Ibid*.

axiological space in which the explainee operates. The full expression of the '*know thy audience*' rule is in fact, '*know thy audience - in the context where the explanation is received,'* 'context' being understood as the realm of possible experiences which can occur when favorable factors are present.[163]

Why is context so important for explanatory accuracy? Several reasons can be highlighted: because it includes the object of the explanation (facticity); it informs of the explainees' 'background'/conventional knowledge (doxa); it contains the values the explainees look to when deciding if they should believe or not. Above all, context justifies the *inquiries* explainers are called to address. *Explanatory relevance* (the 'why' of an explanation) dictates explanatory *salience* (the 'what' of an explanation), meaning that the answers explanations provide should somehow be *meaningful* in connection to a purpose or interest of importance for the explainee.[164]

To refer back to the example of automated gender bias: the issue of 'how did a system become biased?' naturally calls for informed knowledge (of the system's functionalities) and a capacity to deliver that knowledge (to the satisfaction of the explainee). To provide a 'good answer,' the explainer should exercise so-called *explanatory virtue* which Steel says is "a proxy for probability."[165] The explanatory criteria they should meet are thought to include "the extent to which the hypothesis explains more and different kinds of evidence (*consilience*); the *simplicity* of the explanation, understood as measuring the number and kind of assumptions underpinning it; the extent to which the hypothesis coheres with background beliefs, and the extent to which the hypothesis is *ad hoc*."[166]

Against the backdrop of those criteria, the explainee plays the role of assessor, evaluating whether the explanation given is the *best possible one*.[167] This evaluation essentially takes into account the context in which the explanation is given, the trustworthiness of the explainer and the nature and value of the evidence they bring forward - all factors that may (or not) support the explainee's belief that the information given is *reliable*[168] to a point where it can be seen as accurate, believable or acceptable.

Our general - and for lack of space, lacunary - overview of the epistemology of explanations sets the stage for analyzing this concept's translation in *legal liability* contexts. In those, explanations appear as *instrumental concepts* (means to an end).

---

[163] For an analysis of Boolean probability, in connection to context, see P.D. Bruza, L. Fell, P. Hoyte, S. Dehdashti, A. Obeid, A. Gibson, C. Moreira "Contextuality and context-sensitivity in probabilistic models of cognition" (2023) 140 *Cognitive Psy'y*, 101529.

[164] John Greco, "A (different) virtue epistemology" (2012) 1 *Phil'y & Phenomenological Res.*, 1, at 9.

[165] Sandy Steel, *Proof of Causation in Tort Law* (CUP, 2015), at 79.

[166] *Ibid* (emphasis added).

[167] This echoes the so-called 'best evidence rule', famously coined by Morgan who stated that "the highest degree of probability must govern [courts'] judgment; and it necessarily follows, that they ought to have before them *the best evidence of which the nature of the case will admit*." John Morgan, *Essays upon the Law of Evidence, New Trials, Special Verdicts, Trials at Bar and Repleaders* (Johnson, vol. 1, 1779), at 2-3 (emphasis added).

[168] Steel ties reliability with frequentist probability, the gist being the following: if an explantation stems from frequent occurrences (or causal structures), it will likely be viewed as plausible. See Sandy Steel, *Proof of Causation in Tort Law*, *cit. supra*, at 65: "the evidential probability that p should, plausibly, be influenced by the relevant frequentist probability and (in appropriate contexts) the classical probability that p. The case for p is stronger if there is a very high frequentist probability that p. The point made earlier was only that it cannot be reduced to these."

They are not meant to deliver understanding for understanding's sake; they deliver understanding for the purpose of reaching a verdict, appearing as a crucial component in the exercise of a key public function: the administration of justice.

B. The Accuracy of Causal Explanations

In his study of epistemology in data science, Pietsch raised the question of the function of causal knowledge. Why is it important, he asks, "to identify a relationship as causal rather than as a mere correlation?"[169] The author lamented how misguided we are in believing that *knowing* the causal story is equivalent with being able to explain it: "allegedly, without causal knowledge, one can merely describe how things are, but one cannot explain *why* they are as they are."[170]

The scholar made a criticism and gave a hint. He criticized the 'fundamental mistake' of often confusing causation and correlation: we tend to equate causation with theoretical explanation, while overlooking the much more important function of causation to establish *reliable prediction* and *effective intervention*.[171] Prediction and intervention are, according to Pietsch, what causal knowledge is about:[172] *to know* of a harm-causing fact or event is *to know* how to prevent that fact or event from materializing. Forewarned is forearmed!

Pietsch's hint is one already discussed: as imperfect as they may appear compared to the ideal of scientific knowledge, explanations are, nevertheless, a species of the knowledge-genus. As such, *causal* explanations in law do not translate to an exercise in creative narration but unfold in legally defined procedural frameworks, specifically designed to support reasoning about facts (and the causal links they harbor).

Since explanations deliver understanding (as opposed to 'knowing'), the big question in connection to causal explanations is: what does a court *expect to understand* from an explanation on causation? To answer this question, we will use a distinction, suggested by Le Morvan,[173] between 'knowledge of' and 'knowledge that.' The former is propositional, positing that something is true (e.g. the Earth is a sphere), until proven otherwise. The latter is justificatory, referring to the reasons that justify holding a proposition as true (e.g. there is evidence showing that the Earth is a sphere). Le Morvan's knowledge of/that dichotomy is a useful methodological tool to explore two aspects of causality in law: first, the ways in which causality is *represented* (the 'knowledge of' dimension, explored in **Sub-Section 2.2.1.**); second, the ways in which causality is explained under legally defined standards (the 'knowledge that' dimension, explored in **Sub-Section 2.2.2.**).

---

169 Wolfgang Pietsch, *On the Epistemology of Data Science* (Springer, 2022), at 110.
170 *Ibid* (emphasis added).
171 *Id*., at 112.
172 *Id*., at 111: "(causal knowledge is indispensable) not only for effective intervention but also for reliable prediction. In the absence of a causal connection between different variables, including especially the absence of an indirect connection via common causes, any existing correlation between those variables, no matter how strong, cannot establish reliable prediction."
173 Pierre Le Morvan, "On the ignorance, knowledge, and nature of propositions" (2015) 192 *Synthèse*, 3647.

1. Causality Represented Ex Ante (the 'Understanding of')

As a question of fact, legal causality is first and foremost an issue of evidence. The nature and probative value of the evidence given to support an explanation on causality will, to a large extent, allow a court to distinguish the correlative from the causal, the merely 'possible' from the 'probable.'[174] Think of Spinoza's falling stone. There are at least two plausible explanations for the fall, 1. the wind tilted the stone; 2. God willed the stone into falling. Of course, the evidence supporting each explanation will neither be equally available, nor equally probative: it might be within an inquirer's reach to measure the wind's speed, but it will be far more challenging to elucidate the divine intention behind matters like life-threatening falling objects.

Like in most explanatory contexts, in law, causal explanations are not the products of guesswork or wishful thinking; as factive statements, they need to be backed by evidence, the assumption being that the evidence is, indeed, within the explainor's reach (**A**). However, even when this is the case and evidence is within reach, error is still possible regarding the ways in which causal explanations are given. Two risks in particular are noteworthy: causal *underdetermination* (translating to a narrow view of the causes underlying certain effects) and causal *overdetermination* (translating to a much too broad view of cause-effect interrelationships) (**B**).

a. Da Mihi Facti:[175] the Causal Links Revealed by 'Bare' Facts

Causes - Pietsch writes - can serve as "answers to why-questions even though such answers often do not yield deeper explanations."[176] For example, a layperson might no longer have a headache after taking an aspirin, though they could give only a superficial explanation as to why aspirin cures headaches. Deeper explanations "generally refer either to unifying theoretical laws or to causal mechanisms linking the circumstances with the phenomenon."[177]

Pietsch's view of explanatory superficiality is understandable. There are marked differences in the requirements on 'how far should the discovery of facts go' to meet the standards of, respectively, explanatory and scientific accuracy. The reasons for these differences were outlined in the previous Section. At present, we will focus on the features of the standard of fact-accuracy *required by law*: how 'deep' should the knowledge of causal phenomena be for an explanation thereof to allow the reaching of a fair verdict?

---

[174] By employing the terms 'probable' and 'provable,' we in fact allude to an inductivist theory of legal probability pioneered by L.J. Cohen. Astutely observing (and demonstrating) the occasional absurdity of mathematically calculating the truth value of legal evidence (of innocence or guilt) - as if evidential truth was a measurable property - Cohen suggested a method of *inductive probability*, which departs from an empirical foundation, but is nonadditive and therefore not measurable. This (more 'organic') method of fact assessment is arguably closer to how courts already reason about facts, the example being that of inductive (generalizable) conclusions made based on circumstantial (probabilistically 'weak') evidence. See L. Jonathan Cohen, *The Probable and the Provable* (OUP, 1977).

[175] This adage, in its complete version, is *da mihi facti, dabo tibi jus* - give me the facts and I will give you the law.

[176] Wolfgang Pietsch, *On the Epistemology of Data Science*, *cit. supra*, at 111.

[177] *Ibid*.

The evidentiary and explanatory depth required by law varies, depending on the complexity of the causal constellations the law is called to address. Two uncontroversial statements can be made in this regard. First, causal explanations are usually required *in the presence of harm* having resulted from the violation of a preestablished (typically, legally prescribed) duty of care. Second, the explanations required for the purpose of compensating a harm seek to causally link a conduct, or the reasons underlying it with the harm suffered. As a general rule of thumb, the greater the distance - as it were - between a harmful act and a harm, the greater the 'depth' of the fact-digging enterprise aimed at uncovering the causal chain between the two. While probing evidence is necessary when any causal explanation of a harm is given, its importance is arguably greater in cases of AI-related harm because, in those, discerning the actual causal link is often evidentially challenging in the sense that it is not *directly knowable*. AI use can appear as a 'conduct' having instantiated a harm. However, to understand if a human was involved in that instantiation, it is necessary to understand how the AI system functioned *specifically when the harm occurred* (therefore not generally). In other words, in causal contexts where human and non-human intelligence appear as *plausible candidate-causes of harm*, there is a need for a more in-depth discovery and understanding of the relevant facts.

What does standard liability scholarship tell us about the features of causal explanations? In their seminal work on liability, Hart and Honoré point to two types of causal problems: *explanatory* and *hypothetical*.[178] The former - they argue - "arises when it is not clear how certain harm came about or for what reasons a person did a certain act."[179] The latter arises "when a court, in order to determine whether a wrongful act was in the appropriate sense a necessary condition of the harm inquires whether compliance with the law would have averted the harm."[180] Both deal with the issue of cause, as a precondition for the proof and explanation of causation. Indeed, most debates and evidence in liability cases revolve around uncovering *the* (f)act that can be positively and decisively associated with a harm.

It goes without saying that the concept of cause is relational. Facts - Moore says - are "causal relata"[181] but an *isolated* fact has no causal power. It becomes a cause when, in relation to other facts, it leads to a specific consequence. Hart and Honoré observe that in legal language, the cause-effect dyad is often expressed as 'due to', 'owing to', 'result', 'attributable to', 'the consequence of', 'caused by.'[182] For some purposes - they say - it is important to distinguish between these expressions, "though their similarity on many vital points justifies grouping them together as examples of causal terminology (…) sometimes liability or its extent depends on the proof that a wrongful action, or some other contingency, was the cause of harm: this may be so even where common sense, left to itself, might wish to describe the situation by saying that there were several causes of the harm so each was only a cause."[183] From the perspective of evidence, facts offered as proof of causation seek to establish that an act was indeed 'wrongful' *precisely* because it produced a morally or legally reprehensible

178 H. L. A. Hart, Tony Honoré, *Causation in the Law*, *cit. supra*, at 407.
179 *Ibid*.
180 *Ibid* (emphasis added).
181 Michael S. Moore, *Causation and Responsibility : An Essay in Law, Morals, and Metaphysics* (OUP, 2009), at 33.
182 *Id*., at 87.
183 *Ibid*.

consequence. The important question is - again - that of the criteria used to qualify conduct as 'wrongful' (that is, causally necessary for harm to occur).

The most straight-forward scenario of wrongfulness is that of *unlawfulness*, as exemplified through *legal labelling*. Hart and Honoré alluded to this when referring to hypothetical evidence of causation,[184] the goal of which is to establish that a harm had occurred *because* a legally prescribed duty had not been complied with.[185] However useful, legally prescribed causation can be criticized on two points. First, it tends to synonymize *wrongful* and *unlawful* conduct: harm-causing acts tend to be wrongful, regardless of whether there is a legal rule to confirm that they are. Manslaughter would still be morally wrong, even in the absence of a legal rule to confirm that it was. Alternatively, not all unlawful conduct warrants compensation. Suppose a person got a speeding ticket or was not covered by mandatory health insurance: both are unlawful acts but none creates a duty of compensation, in the sense of liability law. 'Wrongful' acts are therefore a generic category of causal relata which include, but are not limited to 'unlawful' acts, also because no legislator is providential to a point where they can lay out an map of all possible real-world causes and their harmful, compensation-worthy consequences.

This brings us to the second criticism mentioned above: causes (and causations) are vague concepts precisely because no one can have full knowledge of all causal phenomena. Save in rare cases, it is often difficult to *a priori* predict that a specific act has the potential of causing a specific harm. For a swears-by-the-code lawyer, it must be anxiogenic to view the world as an ocean of mostly unforeseeable causal mechanisms which is why law, with its manifest *penchant* for stability, aspires toward *causal invariance*.

b. The Risk of (Mis)representing Causality

To 'represent' or exemplify causality is to have a starting point, a template, an intuition on relevant and repetitive causal connections. However useful, legally exemplifying causality calls for a cautious approach: the 'right' causes should be linked to the 'right' effects. The caution is noteworthy because, as mentioned earlier, reality is causally complex: a cause can have several effects, several causes can converge into producing a single effect, an effect can itself be the cause to some other effect… Causal knowledge is therefore an issue of properly connecting or *fitting together* two or several events, the two obvious risks being that of *underfitting* (tying a cause to one specific effect or set of effects) and *overfitting* (where everything can be the potential cause of everything else). *Adequate* causal knowledge no doubt lies midway between casual underdetermination (***a***) and causal overdetermination (***b***).

i. Causal Underdetermination

Causal invariance is a typical example of underdetermination. It presents itself as an "indispensable navigation device within the infinite space of causal

---

184 H. L. A. Hart, Tony Honoré, *Causation in the Law*, *cit. supra*, at 407.

185 *Id*., at 413: "in the absence of reliable evidence about the hypothetical course of events, a court is naturally inclined to *give effect to the policy enjoining the precaution by assuming*, unless there is evidence to the contrary, that the precaution would have averted the harm" (emphasis added).

representations."[186] It brings reassurance in the face of at least three unpredictable contexts: 1. some important real-world causal interrelationships are unobservable; 2. the environment has the potential to contain unknown background causes of an outcome; 3. it is always possible for background causes of an outcome to differ across contexts.[187]

When there is ambient uncertainty, it is all the more necessary to explain why, say, a specific harm occurred, when the evidence linking it to a cause is not available. Law may then step in and save the day by declaring 'what is what.' This is usually done through *generalizing causal invariance*. When an instrument like the AI Act states that biometric identification systems *typically* cause ethnic discrimination, it generalizes or exemplifies a causal link. This means that all biometric identification systems, past and present, have the potential of developing an ethnic bias which is, of course, an overstretch: they may be perfectly bias-neutral or develop biases on grounds like gender. Causal generalizations are logically 'thin': they take a plausible but narrow belief about reality and convert it into a general, supported-by-the-law example of how reality causally works.

Law has often been accused of being under-deterministic because it tends to introduce simplicity where simplicity is not warranted. An overview of the EU legislation on AI certainly reveals a tendency toward causal underdetermination. As we have argued in a recent study,[188] there is no evidence to overwhelmingly show that biometric identification systems are, *without a doubt* - what the AI Act calls - high-risk systems. On the contrary, we showed that, instead of being evidence-based regulation, the AI Act is primarily a market regulating one, barely relying on facts and mostly giving expression to a seductive value discourse according to which the four levels of risk mentioned[189] are justified by the aim to protect fundamental rights.[190]

In the field of epidemiological evidence, Haack also commented on the not so uncommon disconnection between law and reality: "there can be hard-and-fast rules for determining when epidemiological evidence indicates causation, *the legal penchant for convenient checklists* has led many to construe his list of (…) 'viewpoints' as criteria for the reliability of causation testimony."[191]

Law's causal invariance is convenient but sometimes insufficient because by *labelling causality* it limits the possibility of properly *discovering causality*: biometric identification systems do not develop ethnic biases simply because they perform biometric identification. It is because they - somehow - causally link ethnicity (or any other protected characteristic for that matter) with the purpose for which those systems

---

186 Jooyong Park, Shannon McGillivray, Jeffrey K. Bye, Patricia W. Cheng, "Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions" (2022) 132 *Cognitive psychology*, 1, at 3.

187 *Ibid*.

188 Ljupcho Grozdanovski, Jérôme de Cooman, "Forget the Facts, Aim for the Rights! On the Obsolescence of Empirical Knowledge in Defining the Risk/Rights-Based Approach to AI Regulation in the European Union" *cit. supra*.

189 The four levels of risk in the AI Act are presented *supra* in the Introduction of this paper.

190 See Ljupcho Grozdanovski, "The ontological congruency in the EU's data protection and data processing legislation: the (formally) risk-based and (actually) value/rights-oriented method of regulation in the AI Act" *in* Marton Varju (ed.) *Artificial Intelligence and Law: Values, Rights and Regulation in the European Legal Space* (Springer, 2025), 25 p. (forthcoming).

191 Susan Haack, "Correlation and causation. The 'Bradford Hill criteria' in epidemiological, legal and epistemological perspective," *in* Miguel Martin-Casals, Diego M. Papayannis (eds.), *Uncertain Causation in Tort Law* (CUP, 2015), 176, at 180 (emphasis added).

are used (say, selection of asylum applicants or prevention of crime). *That* causality needs to be uncovered through evidence, even if the evidence reveals causal links other than those that the law (like the AI Act) assigns to specific intelligent systems.

This being said, the discovery of actual, as opposed to preemptively exemplified causation is also tricky because it may show that a harm (say, an unfair bias) can be caused by a plethora of facts or events, each being a plausible candidate to qualify as cause.

ii. Causal Overdetermination

Contrasting law's underdetermination, empiricism faces the risk of *overdetermination.*[192] Pietsch illustrates this with the following example: "the current position of Jupiter might be used by a psychic to scare some poor person to an extent that she commits suicide confirming the very astrological prediction. It seems to follow that the position of Jupiter has to be held fix to fulfill homogeneity when examining causes of suicides."[193]

While courts seldom explain causation in reference to the movement of heavenly bodies, they are not immune to overdetermination. In trials, the risk of overdetermining can occur in essentially two series of cases. First, cases of so-called *concurrent causes i.e*. causes which occur simultaneously and present the equivalent potential of being 'necessary conditions' for a given harm.[194] Second, there is the so-called *pre-emptive kind of overdetermination* where the putative causes are chronologically ordered.[195] Suppose - Moore writes - a building caught fire, and by the time a second fire started, the building has already burnt down.[196] In such a case, we could intuitively assert that the 'necessary' condition for the harm (the burnt building) is the first fire. And yet, a strict counterfactual analysis may yield a "counterintuitive implication that *neither fire* caused the harm because neither fire was necessary (each being sufficient) for the harm."[197] Indeed, with preemptive determination, the problem is that of pinpointing the cause which appears to be the decisive one, in the presence of two or more chronologically ordered or concomitant causal candidates.

The business of linking an effect to its *actual* cause calls for caution in the criteria used to distinguish correlation from causation. This is an issue of both discovery (as an act of evidence-gathering) and explanation (as an act of interpreting the evidence gathered). It is an issue of discovery because the designation of a cause is - here again - largely dictated by the nature and probative value of the items of evidence available. It is an issue of explanation because the evidence is analyzed under specific criteria

---

192 Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, *cit. supra*, at 86.

193 Wolfgang Pietsch, *On the Epistemology of Data Science*, *cit. supra*, at 129.

194 As an illustration of concurrent causes, Moore gives the following example: "two fires, two shotgun blasts, two noisy motorcycles, are each sufficient to burn, kill or scare some victim. The defendant is responsible not for only one fire, shot or motorcycle. Yet his fire, shot or noise joins the other one, and both simultaneously cause their various harms. On the counterfactual analysis, the defendant's fire, shot or noise was not the cause of any harm because it was not necessary to the production of the harms – after all, the other fire, shot or noise was by itself sufficient." See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, *cit. supra*, at 86.

195 *Ibid*.

196 *Ibid*.

197 *Ibid* (emphasis added).

used to determine if one cause or a chain of causes had, indeed, a *decisive* influence on the harm materializing. In complex causal scenarios - where the cause-harm link is not straightforwardly discernable - the decisiveness aspect is usually uncovered through the search for the so-called *proximate cause*. Hart and Honoré tell us that under the heading of 'proximate cause' we find multiple methods of causal fact assessment. Almost always - they say - the relevant question is whether or not the harm would have happened without the defendant's act: "this factual component is variously termed 'cause in fact' 'material cause,' *conditio sine qua non*, and is the sole point of contact with what causation means apart from the law."[198] The authors further explain that the 'proximate' label can be used to explain (or not) cause-effect occurrences and is often given out of convenience, public policy, "a rough sense of justice"[199]… When a law decides that beyond a certain threshold of probability, the cause is no longer 'proximate' (*i.e*. can no longer be positively associated with a harm) causation becomes an issue of "practical politics."[200]

Proving and assessing the *degree of proximity* between a possible cause and a harm becomes even more complex when an alleged harm-doer appears to be causally far removed from the harm. This is usually the case in so-called *material contribution* cases and *vicarious liability* cases. In the former, the victim is typically required to show that the defendant's wrongful conduct had made "a 'material contribution' to the disease or injury. The doctrine of material contribution applies to conditions (…) which are known often to be caused by prolonged exposure to some agent (e.g. dust) but where the effect of any particular period of exposure is hard to argue."[201] Material contribution is a *faute de mieux* approach to causation, typically "when the actual cause of an occurrence is unknown in the sense that there is not sufficient evidence to show in detail what happened on the occasion in question."[202] In such a case, a court would look for evidence of "the characteristically different processes by which different causes produce their effects.[203]

In vicarious liability cases - relevant for commodities such as AI - the causal connection sought is that between "the servant's action or omission and the harm, and in no sense of causation is it necessary to establish any causal connection between the master's conduct and the harm."[204] In a scenario involving AI, the 'servant' would be the artificial system whose decision or prediction would act as the *apparent cause* of harm. However, the 'master' would - always - be a human agent exercising a legal right (ownership, use) and complying with a duty (e.g. control and oversight) over that system. The issue of AI liability will be discussed in more detail further.[205] At this juncture, may it suffice stressing that the world of causation (and the explanations thereof) is rich and complex, lending itself to a variety of explanatory possibilities. Let us, therefore, bring forward accuracy as *fil rouge* of this Section.

With accuracy in mind, the legally relevant issue becomes the following: *in a specific case* (*ergo* not generally), how can a harm be plausibly, if not accurately,

198 H.L.A. Hart, Tony Honoré, *Causation in the law*, *cit. supra*, at 90.
199 *Ibid*.
200 *Ibid*.
201 *Id*., at 410 (emphasis added).
202 *Ibid*.
203 *Ibid*.
204 *Id*., at 85.
205 See *infra*, Sub-Section 4.3.

viewed as the consequence of a conduct, phenomenon or event? This is not an issue of *deontic reasoning* (what the law orders us to view as cause). It is an issue of *practical reasoning* based on (and presupposing) a source of valid, trustworthy empirical information that supports the understanding of the relevant facts, offering an answer to a causal inquiry (e.g. who or what *actually* caused a harm?).[206]

In short, causation is a matter of getting the *right kind of evidence* and delivering the *right kind of understanding* based on that evidence because - as will be argued in the following sub-section - in the presence of multiple candidate-causes, *justice* requires that the actual cause of a harm be uncovered. In other words, what we're aiming at is distilling causation from a sea of correlations.

2. Causality Explained Ex Post (the 'Understanding That')

As argued previously, to explain causality is to give an 'accurate' (believable[207]) account of the various stages of a causal chain that connect a fact with an end-result (typically, a harm). We also alluded to the fact that the problem with AI is that the opacity of automated decisional processes makes it difficult to straightforwardly establish a cause-effect connection. Indeed, direct and probing evidence in support of causal explanations is often unavailable, pushing courts to call for expertise which - as the caselaw shows - may neither be available, nor clear on how a well-performing system should and is likely to operate (**A**).

If and when evidence on possible cause/effect correlations is given, courts typically seek to separate causal from correlative associations. To do so, liability doctrines and court practice offer a series of so-called causality tests: essentially, forms of counterfactual reasoning designed to determine if a fact, event or trope was both sufficient and necessary to yield a specific harmful result (**B**).

a. Lessons from North American Caselaw in the Field of AI Liability

The available examples of judicial instances in AI liability - mostly brought before North-American courts - give valuable insight into the evidence that both litigants and courts flag as necessary and probative for the purpose of explaining causation in connection to 'harmful' AI systems. To induce conclusions - as useful takeaways for the future application of the EU's regulation on AI liability - we will focus on the two, abovementioned set of factors that impact explanatory 'goodness.'

On the one hand, we argued that explanations are fact- and context-bound, their 'goodness' being largely dictated by the evidence of the facts that fall in the scope of the explanations. In the existing AI liability caselaw, *expertise* emerges as a privileged mode of evidence (***i***). On the other hand, we argued that a 'good' explanation is one that warrants believability: a situation where the explainee considers they have sufficient reason to accept an explanation as plausibly true. In the caselaw cited hereafter, two trends emerge regarding the conditions for believability litigants and

---

[206] Friedman rightly pointed out that "if epistemic rationality is a form of instrumental rationality, following one's evidence should be conducive to achieving one's epistemic goals." Jane Friedman, "Teleological epistemology" (2019) 176 *Phil. Studies*, 673, at 677.

[207] We allude to our comments on believability as standard for explanatory accuracy, see *supra* , Sub-Section 2.1.2.

courts appear to observe, when assessing if an explanation on a harmful AI system warrants acceptance (***ii***).

i. Lessons on the Fact-Correspondence of Causal Explanations: Expertise as a Preferred Type of Evidence

To distil causation proper from a multitude of correlations, causal explanations - factive as they are - require tangible, probing and verifiable evidence of the causal link between a defective product (like a biased AI) and a harm suffered (say, gender discrimination). When *direct evidence*[208] of that link is unavailable, courts may turn to expertise, the admissibility of which is usually framed by procedural requirements of 'scientificity,' reliability and trustworthiness.

In the US, Rule 702 of the Federal Rules of Evidence defines the essential features that expert evidence must present to be declared admissible. This provision states that "to be scientific knowledge (…) valid reasoning and methodology must be employed: (1) peer review and publication, (2) the known of potential rate of error, (3) general acceptance and (4) testing a theory by attempting to find evidence to disprove it ('falsification')."[209] The main purpose of these criteria is to support the monitoring of the reliability of expert testimony, allowing courts to 'weed out' so-called 'junk science.'''[210]

The criteria listed in Rule 702 are, in a sense, a codification of the 'original' expertise case *i.e. Frye.*[211] In this case, a person was being tried for murder. In their defense, they called an expert witness who testified on the results of a systolic blood pressure deception test, the argument of the defense being that blood pressure was influenced by the changes in the witness's emotions, being on the rise when the witness experienced nervousness. The obvious issue here was whether such a test could be admitted as legal evidence. The court's approach on this point was cautious: while it did not altogether dismiss scientific expertise as a mode of evidence, it defined a key admissibility requirement which referred to the *epistemic soundness* of the method used to yield a result the court might decide to consider as probing. Since judges are not scientists, the criterion used to determine if a method of discovery produced valid knowledge (as opposed to speculative information), it was stated in *Frye* that "while the courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to *have gained general acceptance* in the

---

208 In evidence scholarship, direct evidence (as opposed to indirect evidence) is usually understood to mean proof of fact which does not call for the reality of that fact to be inferred. According to Cansacchi, the 'directness' of evidence derives from the type of information an item of evidence reveals to an assessor (say a court). Direct evidence brings a reality directly to the knowledge of the assessor, without requiring any mediation (additional items of evidence) and without inviting the assessor to interpret what the item can mean. See Giorgio Cansacchi, *Le prezunzioni nel diritto internazionale : contributo allo studio della prova nel processo internazionale* (Eugenio Jovene, 1939), at 11.

209 Michael D. Green, Joseph Sanders, "Admissibility versus sufficiency. Controlling the quality of expert witness testimony in the United States," *in* Miguel Martin-Casales, Diego M. Papayanis (eds.), *Uncertain Causation in Tort Law* (CUP, 2015), 203, at 214.

210 *Id*., at 204.

211 Court of Appeals of the District of Columbia, 3 December 1923, *Frye v. US*, 293 F. 1013 (D.C. Cir. 1923).

particular field in which it belongs."[212] We find here a procedural translation of the principles of acceptance of knowledge discussed earlier:[213] a scientific community is likely to 'validate' knowledge based on the soundness and reliability of the methods used to produce it.

Since *Frye* (1923), the conditions under which the 'general acceptance of a scientific method' could be declared were further clarified in *Daubert*.[214] In this case, the parents of two minor children with birth defects alleged that those defects were due to the mothers' prenatal ingestion of a prescription drug marketed by the defendant. The probative issue was whether the available expertise revealed a risk that the drug might indeed be causally linked to those defects (which experts largely denied). The merit of *Daubert* is that it provides useful insight into the criteria applied to determine the probative quality of scientific expertise. Those criteria pertain to the *trustworthiness* and *admissibility* of expertise and to its *impact* on the outcome of a dispute.

On the point of trustworthiness and based on both *Frye* and Rule 702 of the Federal Rules of Evidence, the Supreme Court in *Daubert* first formulated the basic accuracy requirements, specifying that the adjective 'scientific' implies a "grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation."[215]

Within the framework of our discussion of objectivism/subjectivism in connection to scientific knowledge,[216] the US Supreme Court is - understandably - subjectivism-averse, since probative 'knowledge' cannot be reduced to mere 'subjective beliefs.' The Supreme court further distinguished between validity and reliability, although "the difference between accuracy, validity, and reliability may be such that each is distinct from the other by no more than a hen's kick."[217]

Translating this validity/reliability distinction in the context of dispute-resolution, the Supreme Court noted that "our reference here is to evidentiary reliability that is, trustworthiness."[218] In the interest of assessing the level of general acceptance of a discovery method, a "reliability assessment does not require, although it does permit, explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community."[219] *Widespread acceptance* "can be an important factor in ruling particular evidence admissible"[220] whereas "a known technique which has been able to attract only minimal support within the community (…) may properly be viewed with skepticism."[221] The focus - the Court stated - should be "*solely on principles and methodology, not on the conclusions that they generate*."[222] Based on these premises, the Court's conclusion was obvious: expert knowledge given as evidence in a trial

---

212 *Id*., at 1014 (emphasis added).
213 See *supra*, Sub-Section 2.1.2.
214 US Supreme Court, *Daubert v. Merrell Dow Pharmaceuticals*, 28 June 1993, 509 U.S. 579 (1993).
215 *Id*., at 590.
216 See *supra*, Sub-Section 2.1.
217 US Supreme Court, *Daubert v. Merrell Dow Pharmaceuticals*, *cit. supra*, at 590.
218 *Id*., at 592.
219 *Id*., at 594.
220 *Id*., at 595.
221 *Ibid*.
222 *Ibid* (emphasis added).

should meet at least basic validity requirements warranting acceptance in the relevant scientific field.

More interestingly on the second point - pertaining to the expertise/fairness interrelationship - the Supreme Court stressed that 'scientific' evidence, albeit relevant, can be excluded from a trial "*if its probative value is substantially outweighed by the danger of unfair prejudice*, confusion of the issues, or misleading the jury (…) Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it. Because of this risk, the judge in weighing possible prejudice against probative force under Rule 403 of the present rules exercises more control over experts than over lay witnesses."[223]

Regarding fairness, the ruling in *Daubert* is truly eye-opening because if confirms the specific status of *science-based judicial truth*: accuracy of the disputed facts is, indeed, a precondition for an informed, impartial and by that, fair adjudication. However, courts must remain mindful of the *finality of fairness* of adjudicatory procedures. This is especially true in cases - like those analyzed further in this paper - where consensus on a scientific method is not widespread, but the legal stakes of verifying the soundness of that method are high, especially in criminal proceedings where accurate and reliable information is paramount for the issuing of a verdict. Law asks for fairness and expediency while scientific discovery is ever evolving and seldom set in stone: "scientific conclusions are subject to perpetual revision. Law, on the other hand, must resolve disputes finally and quickly."[224]

In this context, general acceptance, as originally defined in *Frye* was to be viewed as "not a necessary precondition to the admissibility of scientific evidence under the Federal Rules of Evidence, but the Rules of Evidence - especially Rule 702 - do assign to the trial judge the task of ensuring that an expert's testimony both *rests on a reliable foundation* and is *relevant to the task at hand*. *Pertinent evidence based on scientifically valid principles will satisfy those demands*."[225]

The moral of the story in *Daubert* is that fact-accuracy is of course important, but it is function of *evidentiary relevance*: scientific expertise, when used in the courtroom, is not meant to answer a question of pure science; it *participates in answering a question of law*. In other words, the admissibility of (scientific) expert evidence should not rely solely on scientists' opinions *en général*; it should meaningfully guide a court in the latter's application of the law to a specific factual situation.

Following *Frye* and *Daubert*, Anglo-American scholarship explored subsequent applications of this caselaw, in an attempt to induce general criteria (or court trends) used to assess the reliability of scientific expertise. Braford-Hill[226] famously suggested a list, arguing that reliability of scientific evidence - especially in causal scenarios - is most frequently function of the *strength* of the causal association,[227]

---

223 *Ibid* (emphasis added).
224 *Id*., at 597.
225 *Ibid*.
226 See Austin Bradford Hill, "The Environment and Disease: Association or Causation?" (1965) 58 *Proceedings of the Royal Society of Medicine*, 295.
227 Susan Haack, "Correlation and causation. The 'Bradford Hill criteria' in epidemiological, legal and epistemological perspective," *cit. supra*, at 182.

consistency (stemming from the converging results from different investigations performed in different places),[228] *specificity* (the association should be restricted to a specific cause-effect interrelationship), [229] *temporal precedence* (the cause must consistently precede the harm)[230] a *gradient* (essentially a threshold of gravity)[231] and *plausibility* (the cause-effect connection should be plausibly considered as causation),[232] *coherence* (the causal interpretation should not seriously conflict with known facts about the cause-effect interrelationship).[233]

The trouble in AI litigation is that expertise, fitting all of the Bradford-Hill criteria, is often not available. More often than not, direct evidence of a system's 'inner workings' - at the time a harm occurred - will not be available. In a world where transparency and explainability would reign, whenever harm would be causally linked to an AI system, an independent expert would be called to reverse-engineer that system's decisional process, zooming in on the point where the harm-causing 'glitch' appeared. However, save in cases of fully transparent and explainable systems, the scenario of experts stepping in to crack open the black box and save the day is not, and will not be as frequent. If independent expertise is not likely to be feasible, which evidence can courts rely on to discern causation? The *Pickett*[234] and *Loomis*[235] cases can shed some light in this regard.

ii. Lessons on the Believability Dimension of Causal Explanations: the Types of Understanding Sought

(1) The Understanding Sought by Courts: the Shift from 'What Experts Prove' to 'What Experts Say' in Pickett

In 2017, two police officers travelled in an unmarked vehicle in New Jersey. A group of men wearing ski masks and armed with handguns fired in a crowd causing the death of one person. Shortly thereafter, they were arrested. A ski mask, recovered by the police, was analyzed for DNA. The analysis showed two specimens of saliva. A buccal swab from the suspects showed that one of them was the main source contributor. The remaining specimen could not be analysed using traditional DNA testing. The samples were then sent to Cybergenetics (a private laboratory), owner of the TrueAllele software program, assumed to be far superior in terms of accuracy to traditional forensic DNA tests, especially when dealing with complex DNA mixtures. The results correlated the DNA specimen to the defendant (Pickett). He challenged the accuracy and reliability of the probabilistic genotyping, calling for *independent studies* to investigate whether TrueAllele correctly applied the probabilistic genotyping methods.

---

228 *Id*., at 183.
229 *Ibid*.
230 *Id*., at 184.
231 *Ibid*.
232 *Ibid*.
233 *Ibid*.
234 Superior Court of New Jersey (Appellate Division), 2 February 2021, *State of New Jersey v. Corey Pickett,* Docket N° A-4207-19T4.
235 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, 881 N.W. 2d 749 (2016) 2016 WI 68.

The experts stressed that the software program contained approximately 170'000 lines of code written in MATLAB (a programming language designed specifically for visualizing and programming numerical algorithms).[236] They claimed it would take hours to decipher a few dozen lines of the 'dense mathematical text' comprising the code,"[237] leading up to "about *eight and a half years* to review the code in its entirety."[238] In other words, reverse-engineering was not an option, namely because of the excessive duration which would adversely impact the reasonable duration of the trial. In the absence of expert evidence, the courts reverted to *alternative evidentiary strategies*. The question having guided their reasoning is the following: if an *expert cannot prove* the accuracy of TrueAllele's decision in the specific case of Pickett, *what do experts say* on the system's aptitude for accuracy *in general*?

This shift from '*what can experts prove*' to '*what do experts say*' has an important procedural repercussion because it shifts the debate from evidence that is case-specific, highly probative but unavailable (reverse-engineering) to information that is available, but not case-specific and not particularly probative (general expert opinions). Following this approach, the Attorney General in *Pickett* considered three types of evidence: the testimony given by Cybergenetics' expert, validation studies and publications on TrueAllele and opinions from other jurisdictions on the system's performance. All three types of evidence converged on the point that TrueAllele was, *in principle*, reliably accurate[239] which, of course, triggered some discontent.

It was argued that *general expert acceptance* of a model's accuracy (providing, at best, presumptive evidence of debatable probative force) is no substitute for independent, unsupervised review of the source code (providing direct evidence, with strong probative force).[240] It was also argued that even simple software programs are "prone to failure, and that an error in any one of the three domains of software engineering - problem identification, algorithm development and software implementation - undermines the trustworthiness of the science underlying the relevant expert testimony."[241] These opinions are, of course, legitimate. But if *Pickett* confirms anything about the quality of evidence used for the purpose of arriving at a (plausible) causal explanations, it is that in future AI liability cases, the most conclusive evidence (reverse-engineering) may not always be within reach. In cases where the *probatio* (expertise on AI concrete performance) is unavailable, courts are likely to turn to *fama* (an AI's reputed performance in a majority of cases). This redirection from *in concreto* evidence assessment to general opinions is what Duede called *brute inductive consideration i.e.* a belief that an AI system is reliable based on past reliability evaluations.[242] And such a reasoning is 'all too human': given that AI systems can be opaque (therefore, inscrutable and unpredictable), courts 'naturally' search for expert

---

236 *Id*., at 17.
237 *Ibid*.
238 *Ibid* (emphasis added).
239 *Id*., at 25.
240 Some experts stated that the reliability of the TrueAllele software "cannot be evaluated without full access to 'executable source code and related documentation,' something that no one to date has seen." See *State of New Jersey v. Corey Pickett*, Docket N° A-4207-19T4 *cit. supra*, at 34.
241 *Id*., at 35.
242 Eamnon Duede, "Instruments, agents, and artificial intelligence: novel epistemic categories of reliability" (2022) 6 *Synthese*, 1, at 3. Audi called this derivative reliability which, in essence, warrants trust in an information based on the reliability of the source of that information. See Robert Audi, "Reliability as a virtue" (2009) 142 *Phil. Studies*, 43, at 46.

opinions that can confirm a system's *behavioral consistency*. But is this good enough from the litigants' perspective? The answer is 'no;' the *Loomis* case gives hints as to why.

(2) The Understanding Sought by Litigants: the Reasons for (Human) Reliance on AI Output in Loomis

Loomis[243] deals with the use of COMPAS, a risk-need assessment tool designed to predict recidivism and to identify program needs in areas such as employment, housing and substance abuse. The claimant was accused of being involved in a drive-by shooting which he denied. He was charged with five counts and pleaded guilty to only two of the less severe charges. After accepting Loomis's plea, the circuit court ordered a presentence investigation which included a COMPAS risk assessment. The risk scores in this assessment were intended to predict the general likelihood that those with a history of offending are either less likely or more likely to commit another crime following their release from custody. The prediction was based on a comparison between information pertaining to an individual and information pertaining to members of a similar data group. It should be stressed that the risk scores produced by COMPAS were not intended to determine the severity of the sentence or whether the offender should have been incarcerated.

In *Loomis*, the defendant contested the court's *reliance* on COMPAS's allegedly biased prediction which resulted in predicting a higher risk of recidivism, naturally leading to a more severe sentence. In essence, the defendant contended that *by slavishly relying on COMPAS*, the sentencing court erroneously exercised its discretion by not basing its decision on other facts in the record. The consequence of this - it was argued - was the violation of the defendant's due process rights namely, the right to be sentenced "based on accurate information;" the right to an individualized sentence and the improper use of gendered assessments in sentencing.[244]

Loomis is foretelling of a caselaw we will likely see develop in the future because it points to the *reasons underlying the human reliance on a given AI output*. Indeed, the evidentiary (and explanatory) issues we will see down the line will likely not only focus on whether the author of harm was an AI or a human, but if it was a human agent's slavish (non-reasoned) reliance on an automated decision/prediction that caused the harm. To make their argument in this sense, a litigant would need to demonstrate that: 1. an AI output was inaccurate (e.g. biased), which would require proof and explanation on the system's functioning and performance; 2. that the reliance on that output was harmful, which would require evidence on *pre*-, *prae*- and *post*-use accuracy checks.

Like in *Pickett*, in *Loomis*, reverse-engineering of COMPAS was not performed. Rather, the Wisconsin Supreme Court turned to sources, external to the dispute, to arrive at a conclusion on the system's *general accuracy* (thus confirming the above-mentioned shift from *probatio* to *fama*). The Court found e.g. that some States - like New York - have conducted validation studies of COMPAS concluding that its risk

---

243 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra*.

244 *Id*., § 34.

assessments were generally accurate.[245] The defendant, however, cited a 2007 California Department of Corrections and Rehabilitation (CDCR) study which concluded that there was "no sound evidence that COMPAS can be rated consistently by different evaluators (…) that it predicts inmates' recidivism."[246] This study notwithstanding, the Wisconsin Supreme Court considered that the sentencing court used COMPAS merely as an 'aid,' not as a basis for its decision. But the 'battle of experts' in *Loomis* is not reassuring because - again - general expert opinion is hardly strong proof of a system's accuracy *in a specific case*. What would have happened if the California study, critical of COMPAS, was seen as more probative than other (contradicting) studies?... The relevant caselaw is too embryonic to infer the criteria used by courts in their selection of reliable and trustworthy expertise, as regards the aptitude for accuracy of AI systems. For the purpose of this paper, we will view *Pickett* and *Loomis* as examples showcasing an emerging (but not consolidated) trend of *casting a wide net on evidentiary relevance*: when concrete, case- and AI-specific expertise is desirable but unfeasible, general (and reliable) expert opinions on the AI concerned will have to do.

The cited cases also illustrate that evidence is but the first step of the causal explanatory enterprise in cases of AI liability. Save in rare instances where evidence is self-explanatory (e.g. the training data reveals the presence of a bias) the items of evidence discussed before a court will usually be integrated into explanatory narratives which - as mentioned earlier - will aim at delivering causal understanding that explainees (*i.e*. courts) can 'buy into.'[247] To assess the level of understandability and believability, courts use a number of so-called causality tests. These usually play an *exclusionary role*: they are meant to allow the assessment of the 'goodness' of the understanding that explanations deliver, in view of eliminating those which (plausibly) show *correlation* from those that (plausibly) show *causation*.

b. The 'Tests' Used to Explain Causation: But-For and its Variants

Causality 'tests' are used in many legal systems but have especially been developed in Anglo-American court practice and statutory evolution. There are usually notable differences in the ways in which they apply, depending on whether causation is proven in the context of tort or criminal law.[248] As a general *summa divisio* - and based on Moore's work - these tests can be perceived as variations of one test, seen as fundamental across Common law systems: the *sine qua non* or *but-for* test.

This test supports the following *counterfactual reasoning*: but for the defendant's action, would the victim have been harmed in the way that law prohibits?[249] In both criminal law and tort law - as well as in direct and proximate cause scenarios - the but-for test allows courts and juries to zoom in on two points which, if supported by evidence, are likely to uncover the causal or correlational nature of a fact/harm link.

---

245 *Id*., § 59.
246 *Id*., § 60.
247 See *supra*, Sub-Section 2.1.2.
248 Moore argues that criminal law has been a 'borrower' from torts regarding the 'tests' aimed at proving and assessing causation. However, this "borrowing has not been uniform and without reservation (…) the criminal sanction of punishment is sometimes said to demand greater stringency of causation than is demanded by the less severe tort sanction of compensation." See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, *cit. supra*, at 83.
249 *Ibid*.

These points are the *necessity* of the cause for the harm to occur (meaning that without a specific event acting as cause, a harm would have not materialized) and the *sufficiency* of that cause (meaning that the cause was a determining factor for the harm to materialize). By applying the counterfactual reasoning based on the but-for test, court practice has developed a series of variants - specific tests to assess causal necessity and sufficiency. Moore cites, as examples, the *necessary element test*; the *necessary to the time, place and manner* of an effect's occurrence; *asymmetrically temporal* test, the *necessary to accelerations* test; the *necessity of negligent aspects* of acts; necessity as a usually present and always sufficient criterion of *'substantial factor' causation* and causation as *necessity to chance*.[250]

In the field of AI, a peculiar application of the but-for test can be detected in *Loomis*.[251] In assessing the sentencing court's reliance on COMPAS when reaching a verdict (as the causal issue in this case), the Wisconsin Supreme Court found that, the COMPAS assessment was not "determinative in deciding whether Loomis should be incarcerated, the severity of the sentence or whether he could be supervised safely and effectively in the community."[252] To support this argument, the Wisconsin Supreme Court applied a peculiar 'but-forian' reasoning, arguing that the circuit court *would have imposed the exact same sentence* even without having used the COMPAS system.[253]

Loomis gives a glimpse into the reasoning courts are likely to apply in future AI liability cases which will depart from the following question: *would the user of the AI system arrive at the same (harmful) decision, had they not used the system in the first place*? Asking this question is tricky because it opens the door to speculation. To avoid this, we will perhaps see the emergence of additional tests down the line. For example, a '*reasonable user*' test might emerge, which would translate to examining an agent's conduct in a specific occurrence and seek to determine if the alleged harm would have nevertheless occurred, without that agent's conduct.[254] It is - again - too early to speculate on the ways in which the but-for test might be applied in future AI liability cases.

The *second type* of tests include a variety of policy-based tests such as the *reasonable foreseeability* and *harm-within-the-risk* tests. According to Moore, the goal of those is to "describe a factual state of affairs that plausibly determines both moral blameworthiness and duties to compensate, and that plausibility connects a defendant's

---

250 *Ibid*.

251 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra*.

252 *Id*., § 109.

253 *Id*., § 110.

254 Bathaee argues that the principal-supervision rule derived from standard principles on agency can applies in assessments of causation in AI liability cases. The test would basically seek to establish if the programmer or user of the AI system exercised reasonable care in processes like monitoring, designing, testing or deploying. Of course, the principal-supervision rule is applicable in instances where supervision is possible. In cases of unsupervised ML, the relevant issue - Bathaee stresses - is whether it is at all reasonable to have used or deployed such a system. The answer, the scholar says, may be no, which would mean that the creator or user of that system would be liable for any harm that it might cause. See Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation" (2018) 2 *Harv. J. L. & Tech'y*, 890, at 936.

culpability to particular harms."[255] These are the tests we alluded to when we discussed causal invariance (where the law connects specific causes to specific harms).[256]

The *harm-within-the-risk* test essentially serves to discern causation when a cause is associated with - so to speak - a *family of harms*.[257] Think of a recruitment AI: though they are commonly associated with gender biases, we would hardly be surprised if they, at some point, expressed an ethnic bias. Before we witnessed the outcome of the EU's rights-matter-to-us regulatory framework on AI (AI Act, AILD), part of scholarship - including the author of this paper - pleaded in favor of an acceptance-of-risk criterion, serving as referent for identifying the agent having accepted that an AI system may cause harm and can, because of that acceptance, be held responsible.[258] The AI Act essentially integrates the *harm-within-the-risk test* by introducing a form of causal invariance for so-called high-risk AI systems. The invariance aspect is visible in the list of sectors and uses that the AI Act flags as falling in the 'high risk' category. For example, in the field of migration, asylum and border control management, it mentions systems used as polygraphs (and similar tools) aimed at detecting emotional state(s) of a natural person. Intuitively, we could agree that this is, indeed, a high-risk use: errors in detecting emotional states can produce unwanted consequences, especially when such detecting is performed in the processing of asylum applications. The procedural question is whether this causal invariance in the AI Act would somehow *lighten* the burden for victims to prove causality. Imagine an asylum seeker who underwent an emotion recognition test which concluded that the applicant was lying when they explained the reasons why they were forced to flee their country of origin. Based on that decision, their asylum application would presumably be rejected. Suppose the applicant wished to contest that rejection. Would they be required to prove the cause (the system's error) and its harmful consequence (the rejection of the asylum application), given the AI Act states that emotion recognition systems are 'high risk' *anyway*? Now that we have the EU's AI Liability framework, the answer is 'no': though the list of 'high risk' systems in Annex III of the AI Act integrates a causal invariance rationale, it does not create a *general presumption of harm and causation* when those systems are used in practice. The AI Act merely circumscribes the scope of the harms associated with 'high risk' AI, but does not include a *general liability test*, nor does it attach any procedural consequence (e.g. discharge of the burden to prove harm) for high-risk systems. The evidentiary issues associated with those systems are addressed in the EU's forthcoming legislation on AI liability, which will be analyzed further in this paper.

Under the *foreseeability test*, the relevant question to ask is whether a harm was intended, foreseen and foreseeable enough "to render any actor unreasonable for not

---

255 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra,* § 110.

256 See *supra*, Sub-Section 2.2.1. (B).

257 Moore writes that "the harm-within-the-risk test is in the service of justice-oriented policy in its seeking of a true desert-determiner and the test does not ask a redundant question (…) The real question for the harm-within-the-risk test is whether the grading by culpable mental states is all that is or should be going on under the rubric 'legal cause.'" See Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics, cit. supra*, at 100.

258 Ljupcho Grozdanovski, « L'agentivité algorithmique, fiction futuriste ou impératif de justice procédurale ? Réflexions sur l'avenir du régime de responsabilité du fait de produits défectueux dans l'Union européenne », *cit. supra*.

foreseeing it."[259] Of course, the specificity of this test is that it necessarily includes a *subjective element*. Moore calls it the *fit problem*: "fact-finders have to fit to the mental state of the defendant had to the actual result he achieved and ask whether it is close enough for him to be punished for a crime of intent."[260] In criminal law, intent is paramount considering that, for many criminal offences, evidence of the intent-to-harm is required. An interesting example of foreseeability and the (impossible) proof of *mens rea* in the field of AI is given by the *Coscia* case.[261] Here, high-frequency trading algorithm performed spoofing *i.e.* placed phantom orders in the market, then withdrew them when the markets began to 'move' in a desired direction. Since spoofing is a criminal offence in US law, the proof of spoofing requires evidence of *intent to harm* (*mens rea*), which the algorithm in *Coscia* of course did not have. The US courts' found themselves in an unenviable position: on the one hand, they were held to ask for and assess intent-to-harm evidence but were, on the other hand, faced with an objective, practical difficulty to access such evidence, since AI autonomy does not include intentionality *per se*. In this procedural setting, the courts' reflex was to, essentially, *broaden the scope of admissible evidence* and require that the parties '*prove until a responsible human is found*.' Testimonial evidence was ultimately key in adjudicating this case: it was the system's programmers who, in their testimony, revealed that it was the user who 'commissioned' a system capable of spoofing.

In *Coscia*, the intent-to-harm test, when applied, did ultimately direct the court to a human agent. We may however imagine and even expect instances where this might not be the case, leaving open the question of the human who ought to be criminally responsible when *no evidence* shows any trace of criminal (human) intent. This issue will likely not be raised in the EU, since the AILD regulates civil liability. But national courts (including those of the EU Member States) may, at some point in the future, be confronted with scenarios like the one in *Coscia*, only without testimonial evidence to guide them to a responsible human.

II. ACCURACY IN CONNECTION TO EXPLAINABLE AI (XAI)

In connection to AI, accuracy is a tricky concept for two reasons. First, on a theoretical level, AI technologies are slowly pushing changes on some of the bedrock-principles of epistemology: we are now in the era of data-driven science which "seeks to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon."[262] This new field of data science seeks to "incorporate a mode of induction into the research design, though explanation though induction is not the intended end-point (as with empiricist approaches)."[263] Instead, "it forms a new mode of hypothesis generation before a deductive approach is employed. Nor does the process of induction arise from nowhere, but is situated and contextualized within a highly evolved theoretical domain."[264]

---

[259] Michael S. Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics, cit. supra*, at 100.
[260] *Ibid*.
[261] US Court of Appeals for the 7th Circuit, *US* v. *Coscia*, 866 F.3d 782 (2017).
[262] Rob Kitchin, "Big Data, New Epistemologies and Paradigm Shifts" (2014) 1-1 *Big Data & Society*, 1, at 5.
[263] *Id*., at 6
[264] *Ibid*.

Second, and more importantly, there is the question of *law's response* to these new 'epistemic actors.' The fundamental issue here is whether evidentiary causal explanations in AI liability cases can, or even should integrate explanations of specific AI output (*i.e*. if causal explanations about AI-related harm require explainable AI).

Before we tackle this issue in the context of the EU's procedural framework on AI liability, we should pay closer attention to the criteria according to which AI output can be viewed as accurate (**Sub-Section 3.1.**). In light of those, we will then explore the conditions that *explanations on AI* should meet in order to, themselves, be qualified as accurate or, at the very least plausible (**Sub-Section 3.2.**).

A. Accuracy Standards for AI Output

When Badea and Artus defined 'intelligence' in connection to *artificial* intelligence, they gave the impression of weighing their words and rightfully so: the only referent we have for intelligence is that of *human* intelligence which smart technologies are capable of simulating, without - yet - fully reaching the intelligent-as-a-human standard: "by intelligence, we of course do not necessarily mean anything as grand as consciousness or Artificial General Intelligence (AGI), but, rather, the *ability to be an effective and creative utility* (or function) maximiser, *i.e*., a machine that is 'clever' at finding ways to achieve the goals we set for it."[265]

But machines can be 'clever' in achieving preassigned goals in ways that humans (clever as they themselves are) are not always capable of discerning or foreseeing. In this context, the question 'what is *accurate* AI output?' depends on first addressing the issue of 'how does AI produce knowledge of the world in the first place?' To address these questions, it is necessary to first explore the peculiar epistemic status of intelligent technologies which albeit created by humans, gradually become their (mighty) fellow-knowers (**Sub-Section 3.1.1.**). Against this backdrop, we can then explore the challenges that humans experience when explaining how AI systems actually 'understand' information about reality (data), when they have nothing else to go by but the output those systems produce (**Sub-Section 3.1.2.**).

1. The Epistemic Specificity of Non-Human 'Knowers'

From the perspective of 'standard' knowledge-construction theory[266] whereby human agents are the sole 'knowers' of the world, AI technologies are certainly avantgarde: for the first time in history, non-human entities are capable of employing the reasoning models historically associated with humans. Because of this, we would be inclined to assume an *epistemic parallelism* between human and non-human 'knowing': since both deploy the same reasoning models, they must also share the same standards by which the knowledge they acquire can qualify as accurate. A nuance should however be highlighted. It is one thing to draw parallels between humans and AI on how they go about acquiring knowledge. It is another thing to inquire on how humans arrive at such knowledge when the object they seek to 'know' (or understand) is an AI system and its output. Epistemically speaking, we are in the presence of *two*

---

[265] Cosmin Badea, Gregory Artus, "Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents," *cit. supra*, at 125.

[266] See *supra*, Sub-Section 2.1.

sets of accuracy standards: those that apply to AI output and those that apply to the explanations pertaining to that output.

In AI scholarship, accuracy has been closely associated with *performance*. According to Liang *et al.*, it "highlights any performance benefits of relying on the recommendation and offers a benchmark against which individuals can judge their own performance."[267] Alternatively, explanations pertaining to AI output "are able to measure the importance of parts of the input or intermediate features towards a model's decision - and can therefore be viewed as an additional and high-dimensional measurement for the discussed properties, depending on the application."[268] In AI jargon, explanations are meant to allow "for a better (compared to, e.g., just relying on the prediction error) control of the model behavior."[269]

The oft-recalled trouble with advanced ML systems is *opacity*. As Edwards and Veale put it, AI technologies may exhibit implicit rather than explicit logics since the ways in which they learn about, and shape reality do not often offer the opportunity to backtrack the stages of their inferential process.[270] Inscrutability of ML and DL models is an epistemic concern, where explanations and understanding are considered as central epistemic virtues.[271] This inscrutability is - Duede points out - that the relationship between an ML or DL model and the real world is *mediated by the logic of what the system learnt:* "no direct causal connection between the world and the DLMs mediates the model's output of a given value."[272]

To illustrate this: say a recruitment algorithm was programmed based on a simple 'if-then' rule.[273] The application of this rule would allow the system to view factors (education, work experience, career advancement, languages spoken etc) as indicators of work performance and, based on those, it would be able to infer a person's level of skill. Suppose that, when processing data not seen during training, the system - somehow - associated gender with work performance concluding that, because men's professional advancement is historically more common, they must be more skilled than women.[274] The consequent inference would be that gender is a sign of high work

---

267 Garston Liang, Jennifer F. Sloane, Christopher Donkin, Ben R. Newell, "Adapting to the algorithm: how accuracy comparisons promote the use of a decision aid" (2022) 14 *Cognitive Research: Principles & Implications,* 1, at 2.

268 Leander Weber, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement" (2023) 92 *Information Fusion*, 154, at 165.

269 *Ibid*.

270 Lilian Edwards, Michael Veale, "Slave to the Algorithm: Why a Right to an Explanation is Probably Not the Remedy You Are Looking for" (2017-2018) 18 *Duke L. & Tech'y Rev.*, 8, at 25.

271 Eamon Duede, "Instruments, agents, and artificial intelligence: novel epistemic categories of reliability" *cit. supra*, at 491.

272 *Id*., at 500.

273 If-then models are typical of so-called conditional reasoning consisting in matching a set of conditions (if a person has university education) with consequences that follow from those conditions (then the person is a highly qualified worker). Our explanation here seeks simplicity, a critical analysis of the theory of conditional reasoning being beyond the scope of this paper. For such analysis, see e.g. Ruth M.J. Byrne, Philip N. Johnson-Laird, "'If' and the problems of conditional reasoning" (2009) 7 *Trends in cognitive sciences*, 282.

274 On the effects of gendered AI systems, see Lena Wang, "The Three Harms of Gendered Technology" (2020), 24 *Australasian J. Inf. Systems*, 1.

performance *i.e*. one that the labour market favors. Amazon's gender-discriminating recruitment system provides the topical example of this.[275]

The problem with this scenario is that the gender-skill association, made by a system in its 'discovery' of the real world, may not always be foreseen by the users and even programmers.[276] This has its importance in the context of harm (unfair biases, physical injuries, illegal investments, medical misdiagnosis etc). As a matter of principle, AI-related harm is usually thought to be the result of miscalculation, error, deviation from that for which the system was trained to do.[277] The question is: how to *causally explain* the occurrence of such harm? Realist epistemic currents do not help much in answering this question. Their postulate is, essentially, that the objects of cognition are tangible occurrences with relatively discernable causes: if snow falls, we may - as some philosophers have - engage in extensive debates on the conditions under which we may assert that 'snow is white.'

In our recruitment hypothetical, the real or tangible occurrence (the AI output) does not seem to reveal a lot on the causal interrelationship (in the form of variable-association) underlying it. This leads to an important epistemic consequence. Kitchin[278] commented that, in pre-AI times, the operative assumption was that any scientific hypothesis could be tested and verified.[279] This paradigm - he argued - consisted of "overly sanitized and linear stories of how disciplines evolve, smoothing over the messy, contested and plural ways in which science unfolds in practice."[280] AI disrupted this 'sanitized' view, upsetting epistemologists' *penchant* for methodological reliability, expressed in the belief that procedures designed to produce knowledge *reliably produce* the knowledge they are designed for. In this context, is AI's capacity for knowledge-construction different from (or more sophisticated than?) that of human 'knowers'? The answer is no… and yes.

2. The Specificity (and Interpretability) of AI 'Knowledge'

The answer to the above-mentioned question ('is AI's capacity for knowledge-construction different from, or more sophisticated than, that of human 'knowers?') is

---

275 Roberto Iriondo, "Amazon scraps secret AI recruiting tool that showed bias against women," Carnegie Mellon University (available on: https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html, last visited, 20 Jan. 2024).

276 See Weston Kowert, "The Foreseeability of Human-Artificial Intelligence Interactions" (2017) 1 *Texas L. Rev.*, 181, at 204: "once the artificial intelligence is sent off to the buyer, the programmer no longer has control and the artificial intelligence could be shaped by its new owner in uncountable ways."

277 Schiyong and Kaizhong essentially view error as a judgment made in application of rules that in a given 'universe of discourse' allow to identify erroneous definitions or assertions. See Liu Shiyong, Guo Kaizhong, *Error logic: paving pathways for intelligent error identification and management* (Springer:2023), at 2-3. Chanda's and Banerjee's definition of error is more functional in the sense that they define errors in reference to the objectives (and expected outputs) of AI systems. For them, errors are 'inadequacies' which can be of two kinds: errors of comission (doing something that should not have been done) and errors of omission (not doing something that should have been done). See Sasanka Sekhar Chanda, Debarag Narayan Banerjee, "Omission and commission errors underlying AI failures" (2022) *AI & Society*, 1, at 1. In short, errors (like unfair biases) are deviations from a model's basic programming.

278 Rob Kitchin, "Big Data, new epistemologies and paradigm shifts" (2014) 1 *Big Data & Society*, 1.

279 *Id*., at 3.

280 *Ibid*.

'no' because, as already mentioned, AI is programmed based on human reasoning models, only - or so the story goes - they seem to apply those models in ways that average human agents do not.

If one lends an ear to some mainstream narratives, the attractiveness of AI stems precisely from its ability to outperform humans.[281] To a certain degree, this holds. In the field of medicine e.g., Arno *et al*.[282] sought to determine if the accuracy of AI-assisted risk-of-bias detection was comparable (noninferior) to human-only assessments. They found that in terms of *efficacy* - essentially the margin of statistical error between automated and human-only assessments - AI reached an accuracy threshold of 0.89/1 whereas for humans, the threshold was of 0.90/1.[283] AI-assisted decisions were therefore not inferior to human decisions in terms of efficacy but - the authors point out - efficacy is not an indicator of *effectiveness*, understood as the possibility for AI to produce the output that is not only accurate, but *desired* in real-life contexts. Think of the recruitment AI: if the system found that, historically, part-time workers are mostly female - which may be statistically correct - it should not be programmed to make the generalization that all women underperform in comparison to men. In this scenario, an efficacious output (though backed by statistical data) will not necessarily be viewed as effective, as it would possibly lead to restricting access to work for women, causing a text-book example of gender discrimination.

These observations allow us to fine-tune the concept of AI accuracy flagged at the beginning of this Sub-Section: although this concept is linked to the quality of AI's probabilistic reasoning, *it does matter* how this reasoning will impact the reality of humans. A well performing (accuracy-apt) system is one that would achieve a difficult double task: be statistically correct (efficacious)[284] and value-conform (effective). In this regard, regulators and scholars seem to have reasoned in terms of another *procedural parallelism*: the *design* of the inception procedures of AI systems directly shapes those system's *aptitude for accuracy*. In terms of cognition, the way knowledge about the world is *represented* in the coding phase of AI will shape the way in which AI will subsequently 'know' and 'act' in the world. In this context, it is not very surprising that regulatory and savant attention turned to the criteria used for the establishment of *ground-truths*, as a form of proto-knowledge comprised of data that an AI system can refer to when confronted with new data that is, data not seen during training.[285]

---

[281] See Katja Grace, Allan Dafoe, Baobao Zhang, Owaian Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts" (2018) 62 *J. of AI Res.*, 729.

[282] Anneliese Arno, James Thomas, Byron Wallace, Iain Marshall, Joanne E. McKenzie, Julian H. Elliot, "Accuracy and Efficiency of Machine Learning-Assisted Risk-of-Bias Assessments in 'Real World' Systemic Reviews: A Noninferiority Randomized Controlled Trial" (2022) 7 *Annals of Internal Medicine*, 1001.

[283] *Id*., at 1004.

[284] Efficacy is essentially a matter of accurate representation, not only of concrete outputs, but also of how accurately AI systems represent their targets. See Eamon Duede, "Instruments, agents, and artificial intelligence : novel epistemic categories of reliability" *cit. supra*, at 496.

[285] Lebovitz *et al.* define the term 'ground truth' as referring to the labels assigned to the data sets used to train a ML model to link new inputs to outputs and to validate its performance. See Sarah Lebovitz, Natalia Levina, Hila Lifshitz-Assa, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What" (2021) 3 *MIS Quart'y*, 1501, at 1509.

It goes without saying that the selection of the data used to constitute ground truths should be performed with great caution in the protocolized process called *labelling*: the assembling and 'cleaning' of data used during a model's programming.[286] Ground data constitutes the *cognitive referent* the system will use when performing in practice. To assess the quality of this performance, the model undergoes *training* that is, the phase where it is confronted to sub-sets of preselected data. If the model performs well (*i.e*. the risk of error is minimal or 'tolerable'), the model would go on to the so-called *validation stage*.

It should be mentioned that a well performing AI is never bias-free, but one that arrives at statistically accurate outcomes *in spite of the biases* that may be either embedded in the ground data or learnt during the model's lifetime. We have discussed elsewhere that accuracy, in AI jargon, is really a balance between *bias* (preferences embedded in the ground data) and *variance* (a model's ability to make relevant decisions and predictions when confronted to data not seen during training).[287] This balance is struck through much testing and controlling of the sample size used in the training stage. With accuracy as *fil rouge* of this paper, we will rather focus on the epistemic conditions that usually warrant 'accurate' AI output. In this vein, ground truths play the role of *premises* the accuracy of which should, logically, dictate the accuracy of the conclusions.

This is the underlying *leitmotiv* of labelling: once ground truths are selected, the systems are trained to create associations between variables, generating a series of relative weights that can be applied to future data inputs.[288] Lebovitz *et al.* refer to - what they view as - a standard method of measuring the quality of an AI model which involves the calculation of how often the model's predicted outputs match the label *a priori* defined as accurate in the data set reserved for model validation.[289] This assessment of course requires expertise, but not only. The authors cite radiology as an example: professionals in this field are trained to refer to the 'Area Under Curve' (AUC) when determining if any technological tool (ranging from imaging equipment to analytical tools) improves diagnostic accuracy.[290] AUC is therefore "primary evidence of performance"[291] supported by larger scientific acceptance (expertise published in specialized journals e.g.) and combined with other methods available for the accuracy

---

[286] Carbonara and Sleeman sfocus on the process of knowledge construction for the purpose of AI programming. For any knowledge-based system - they argue - the process of *accurate representation* of domain knowledge includes three main stags: *knowledge elicitation*, *knowledge representation* and *testing/refining* the initial knowledge base (KB_0). In the first two stages consist in using various automated tools for knowledge elicitation and representation. Knowledge refinement is a process through which the initial knowledge base KB_0 is tested and fine-tuned. To do so, two sets of cases are used: training cases used for knowledge refinements and training cases used to measure the effectiveness of those refinements, thus allowing to measure a system's effectiveness and performance in practice. See Leonardo Carbonara, Derek Sleeman, "Effective and Efficient Knowledge Base Refinement" (1999) 37 *ML*, 143, at 144.

[287] Ljupcho Grozdanovski, "In Search for Effectiveness and Fairness in Proving Algorithmic Discrimination in EU law" (2021) 58 *CMLREv.*, 99, at 107.

[288] Sarah Lebovitz, Natalia Levina, Hila Lifshitz-Assa, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *cit. supra*, at 1503.

[289] *Ibid*. The calculation is represented by a metric called the 'Area Under the Receiver Operating Curve' (AUC) and plotted on two-dimensional graphs. The AUC is a summary of a model's success and error rates, with predictions of possible false negatives and false positives. See *ibid*.

[290] *Id*. at 1508.

[291] *Ibid*.

assessment of a given system. This suggests that AI programming is integrated in broader scientific and social contexts with already existing methods of seeking and verifying information: "in knowledge-intensive contexts, experts developed over the years rich know-how practices to form high-quality knowledge outputs."[292]

Because expert fields and new technologies evolve side by side, coding should be an extremely cautious process when it comes to 1. deciding which data is 'true enough' (at a given point in time) to be used as ground data; 2. embedding models of reasoning that can allow a system to rely on that data and produce an accurate (*i.e.* efficacious *and* effective) outcome.[293] Of course, high-quality, bias-free ground data gives some assurance that a system will perform well when 'released into the wild,' but this assurance is not absolute certainty. There is always a margin of doubt that an AI system may not produce the type of output it was programmed to produce.

This unpredictability is, arguably, why AI technologies upset standard epistemology (the 'yes' answer to the question mentioned earlier): the absence of unfair biases in the labelled data does not automatically imply that a system's output will *systematically* be bias free.

The fact that we can no longer reliably assume the input/output parallelism (in terms of accuracy) is a sign of a much deeper epistemic shift triggered by Big Data. Indeed, the possibilities for various scientific and non-scientific communities to interact within - to borrow Floridi's jargon - the *infosphere*[294] hold the remarkable potential of increasing the speed with which (valid) knowledge is produced and disseminated. In addition, the sheer volume of Big Data presents several epistemic advantages: it can capture a whole domain and provide full resolution; there is no need for *a priori* theories, models or hypothesis for knowledge to be - as it were - distilled from the vast volumes of data; through the application of *agnostic* data analysis, the data can speak for themselves free of human bias; any patterns and relationships within Big Data are (presumed to be) meaningful and truthful; learning transcends context or domain-specific knowledge, thus can be interpreted by anyone who can code a statistic or data visualization…[295]

In this context, scholars have detected the "*troubling disconnection* between ML-based AI quality measures that were based solely on know-what aspects of knowledge and the rich know-how practices experts rely in their daily work."[296] This of course had a profound implication on the ability to assesses a system's potential risks and benefits.[297] If the process (the 'how') preceding an output could not be sufficiently explained based on output alone, quality measures needed to be put into place for in-depth assessments to be made possible. In the trials conducted by Lebovitz *et al*., the

---

292 *Id*., at 1512.

293 *Id*., 1513-1514: "to evaluate AI outputs, managers began reflecting on the know-how practices that enable internal experts to grapple with uncertainty in their daily work and produce high-quality judgments."

294 Luciano Floridi, "Ethics after the Information Revolution" *in* Luciano Floridi (ed.), *The Cambridge Handbook of Information and Computer Ethics* (CUP, 2012), 3, at 6.

295 Rob Kitchin, "Big Data, new epistemologies and paradigm shifts" 1 Big Data & Society (2014), 1, at 4.

296 Sarah Lebovitz, Natalia Levina, Hila Lifshitz-Assa, "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *cit. supra*, at 1514 (emphasis added).

297 *Ibid*.

qualifications of the labelers were under high scrutiny, as was the "taken-for-granted representations of knowledge."[298] This eventually led to admitting that "even labels generated by experts limited [the] evaluations since experts' knowledge outputs were subject to deep underlying uncertainty and ignored know-how aspects of knowledge that were essential to producing knowledge in practice."[299]

In light of the above, it was a set of *professional standards* established, not so much as guaranteeing AI accuracy, but as supporting the belief - namely of users - that *accuracy was likely*.[300] In the AI Act, the accuracy-enhancing (and, trust-engineering) standards target, in particular, the so-called high-risk systems. Interestingly - but understandably - accuracy is seen as a *byproduct of resilience*. For example, Article 15 AI Act states said systems should be resilient as regards "errors, faults or inconsistencies that may occur within the system or the environment in which it operates, in particular due to their interaction with natural persons or other systems."[301] They will also be resilient with regard to attempts by unauthorized third parties to alter their use or performance by exploiting the system vulnerabilities.[302]

The technical solutions to address AI specific vulnerabilities shall include - the AI Act states - measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws.[303] In essence, high-risk AI systems should be resilient to anything that might cause them to deviate from their purpose. Whether this level of resilience can be achieved through technical standardization is an issue we have explored elsewhere.[304] At this stage, the takeaway from our observations on accuracy is that as a *concept*, as an *aptitude* (of a model) and as a *property* (of both ground data and AI output) *perfect accuracy* is technically difficult to instill and comes with no guarantees: try as they might, AI programmers are seldom in a position where they can predict that a well-performing AI system will invariably hit the mark in producing perfectly efficacious and effective output. This is a constant not only in discourse on expert systems (by now associated with the 'stone age' of AI development)

---

298 *Ibid*.

299 *Ibid*.

300 Commenting on the regulatory discourse on trustworthy AI and the use of technical standardization as the means to make AI 'trustworthy', Laux *et al*. stress the possibility that standardization is meant to 'engineer' trust. See Johann Laux, Sandra Wachter, Brent Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk, (2023) *Regulation & Governance*, 1-30, at 2.

301 AI Act*, cit supra,* Art. 15-3.

302 *Id*., Art. 15-4.

303 *Id*., Art. 15-4.

304 Ljupcho Grozdanovski, "The ontological congruency in the EU's data protection and data processing legislation: the (formally) risk-based and (actually) value/rights-oriented method of regulation in the AI Act" *cit. supra.*

but also with generative AI. Much like its more primitive predecessors, ChatGPT was also found to produce output 'tainted' by un unfair bias.[305]

In causal explanatory contexts, the million-dollar question is, of course, *why*? To give a plausible answer to this question there seem to be two sets of conditions: 1. that a given output *lends itself* to an explanation (explainability-as-interpretability); 2. that the explanation provides *adequate understanding* of the process through which that output was produced (explainability proper).

B. Accuracy Standards for Explanations of AI Output

A key doctrinal referent in this sub-section is the remarkable study produced by Barredo Arrieta *et al*.[306] on XAI where the authors highlight five operative concepts. First, *understandability* or *intelligibility*, which denotes "the characteristic of a model to make a human understand its function - how the model works - without any need for explaining its internal structure or the algorithmic means by which the model processes data internally."[307] Second, *comprehensibility* which "refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion."[308] Third, *interpretability* defined as "the ability to explain or to provide the meaning in understandable terms to a human."[309] Fourth, *explainability*, "association with the notion of explanation as an interface between humans and a decision maker (and is) at the same time, both an accurate proxy of the decision maker and comprehensible to humans."[310] Finally, *transparency*: "a model is considered transparent if by itself it is understandable."[311]

We can derive from the relevant scholarship that, in the field of AI, explainability can be *a priori* or *ex post*. *A priori* (*ad hoc*) explainability pertains to the criteria or standards which, if followed, are assumed to, if not guarantee, at least contribute to a system's explain-*ability* down the line (**Sub-Section 3.2.1.**) *Ex post* (*post hoc*) explainability pertains to the interpretation (retro-rationalization) of AI output, once such output is produced (**Sub-Section 3.2.2.**).

---

305 A recent study analyzing the output of two large language models (LLMs) namely ChatGPT and Alpaca, charged with drafting recommendation letters for hypothetical workers. It was observed that the language used by both systems to describe the workers was heavily gendered (using 'expert' and 'integrity' for men and 'beauty' or 'delight' for women). See Christ Stokel-Walker, "ChatGPR Replicates Gender Bias in Recommendation Letters" available on: https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/#:~:text=But%20a%20new%20study%20advises,recommendation%20letters%20for%20hypothetical%20employees (last accessed on 20 Jan. 2024).

306 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (2020) 58 *Information Fusion*, 82.

307 *Id*., at 84.

308 *Ibid*.

309 *Id*., at 85.

310 *Ibid*.

311 *Ibid*. In their study, Barredo Arrieta et al. divide transparent models into three categories: simulatable, decomposable and algorithmically transparent.

1. Ad Hoc Explainability: Embedding Transparency, Hoping for Explicability

The object of *ad hoc* explainability is a matter of standardization, essentially translating to the observance of pre-established functional and operational requirements meant to enhance a model's comprehensibility.[312] This is, no doubt, the reason why technical standardization was ultimately favored by the EU's legislature in regulating AI systems. The 'standardization narrative 'can be traced back to the HLEG's Ethics Guidelines[313] where explicability appears as one of the four cardinal principles for ethical AI, alongside the respect for human autonomy, prevention of harm and fairness. This principle - the experts argued - is crucial for building and maintaining trust in AI.[314] Curiously, the HLEG distinguished between *explicability* and *explainability*.

According to the Guidelines, *explicability* refers to the factors that support and reinforce it. Those factors are unsurprising: transparency and clarity of communication.[315] Where explicability is obstructed, the HLEG stressed that other measures (e.g. traceability, auditability and transparent communication on system capabilities) can be required, "provided that the system as a whole respects fundamental rights."[316] Alternatively, *explainability* is a *component of transparency*, pertaining to the "ability to explain both technical processes of an AI system and the related human decisions." [317] In connection to explainability, the HLEG emphasized human understandability[318] derived from explanations of the degree to which an AI system influences and shapes the decision-making process, design choices of the system and the rationale for deploying it.[319]

The distinction between explicability and explainability in the HLEG's Guidelines is interesting. Explicability seems to refer to the factors (transparency and clarity) that support a model's *interpretability*. From the vantage point of explanatory epistemology examined previously, it is possible to argue that those factors are meant to support an explanation's *objectivist* dimension or facticity.[320] In other words, transparency and clarity should make - what in a legal setting would be considered as - *elements of fact* (ground data, programming, training and validation etc) discernable, so that a model's functioning and output can *in fine* be interpreted. Alternatively, explainability - as the HLEG seems to understand it - is more *subjectivist*, explainee-oriented, focused on the *format* and *features* that explanations must have to be *understandable*.

---

312 According to Guidotti *et al*., the functional requirements of XAI are those that identify the algorithmic adequacy of a particular approach for a specific application, while operational requirements take into consideration how users interact with an explainable system and what is the expectation. See Ricardo Guidotti, Anna Monreale, Dino Pedreschi, Fosca Giannotti, "Principles of Explainable Artificial Intelligence" in Moamar Sayed-Mouchaweh (ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications* (Springer, 2021), 9, at 12.

313 High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI* (2019), *cit. supra*.

314 *Id*., at 13.

315 *Ibid*.

316 *Ibid*.

317 *Id*., 18.

318 *Ibid*.

319 *Ibid*.

320 See *supra*, Sub-Section 2.1.1.

We consider that the explicability/explainability distinction in the HLEG's Guidelines is an issue of semantics. As will be argued further, XAI is multilayered. However, concepts such as interpretability, comprehensibility and transparency are *instrumental* to explainability as the generic, operative concept in the field of XAI. In light of this, in the remainder of this paper, we will not use the HLEG's explicability/explainability distinction but will instead generically use explainability in our analysis of both the factive and subjective aspects of explanations pertaining to AI performance and outpout. Semantic parenthesis closed.

Following the HLEG's Guidelines, the AI Act translated the requirements on explainability in technical standards targeting, in particular, the so-called high-risk systems. These can be clustered in roughly *three families*.

The *first* includes standards that generate *requirements for accuracy* (of the ground data) and *transparency*. These requirements pertain to data governance and management practices such as relevant design choices,[321] data collection;[322] relevant data reparation processing operations, such as annotations, labeling, cleaning, enrichment and aggregation,[323] the formulation of relevant assumptions, namely with respect to information that the data are supposed to measure and represent,[324] prior assessment of the availability, quantity and suitability of the data sets that are needed,[325] examination in view of possible biases [326] and identification of data gaps or shortcomings, and how those can be addressed. [327] Unsurprisingly, the AI Act expresses a basic requirement that training, validation and testing data sets be *relevant*, *representative*, *free of errors* and *complete* [328] taking into account, "to the extent required by the intended purpose" the characteristics pertaining to specific geographical, behavioral and functional setting within which the high-risk system is intended to be used. [329] The *data governance requirement* is, of course, meant to increase the transparency and provision of information to users. Article 13(1) states that high-risk AI systems shall be "designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately." Perhaps naively, this Article states that the "appropriate type and degree of transparency" - whatever 'appropriate' is - will be reached through compliance with the obligations set out in the AI Act.[330] High-risk systems shall, in addition, be accompanied by instructions for use, in a digital format, that include concise, complete, correct and clear information that is "relevant, accessible and comprehensible for users." [331] The information required includes *inter alia* the characteristics, capabilities and limitations of performance of the high-risk system including its intended purpose,[332] the levels of accuracy robustness and cybersecurity against which the system had been tested and validated and which "can be expected"

321 AI Act, *cit. supra*, Art. 10-2 (a).
322 *Id*., Art. 10-2 (b).
323 *Id*., Art. 10-2 (c).
324 *Id*., Art. 10-2 (d).
325 *Id*., Art. 10-2 (e).
326 *Id*., Art. 10-2 (f).
327 *Id*., Art. 10-2 (g).
328 *Id*., Art. 10-3.
329 *Id*., Art. 10-4.
330 *Id*., Art. 13-1.
331 *Id*., Art. 13-2.
332 *Id*., Art. 13-3(b)(i).

as well as any known and foreseeable circumstances that may have an impact on the expected level of accuracy, robustness and cybersecurity,[333] any known or foreseeable circumstance related to the use of a high-risk system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety of fundamental rights,[334] its performance as regards the persons or groups on which the system is intended to be used,[335] when appropriate, specifications for the input data, or any relevant information in terms of training, validation and testing data sets used, taking into account the intended purpose of the AI system.[336] The information should further include the *changes* of the high-risk AI system determined by the provider during the initial conformity assessment,[337] the human oversight, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users,[338] the expected lifetime of the high-risk system and any necessary maintenance and care measures to ensure the proper functioning of that system, including as regards software updates.[339]

The *second family* of standards create requirements to *produce proof of compliance* and *traceability*. Under these requirements, the programmer is held to keep technical documentation,[340] drawn up "in such a way to demonstrate" the compliance of a high-risk AI system with the AI Act. They should also perform record-keeping able to show that high-risk systems are designed with capabilities enabling the automatic recording of events ('logs') while those systems are operating.[341] The logging capabilities should increase the level of traceability[342] and facilitate monitoring of a system's operation in situations where it may present a risk of harm.[343] In a similar vein, Article 11(4) of the AI Act states that the logging capabilities should provide, "at a minimum" recording of the period of each use of a given system,[344] the reference database against which input data has been checked by the system,[345] the input data for which the search has led to a match[346] and the identification of natural persons involved in the verification of the output.[347]

The *third family* of standards pertain to *human oversight*. Article 14 of the AI Act creates the obligation to provide appropriate human-machine interface tools so that high-risk AI systems can be effectively overseen by natural persons during those systems' use.[348] It should prevent and minimize the risks to health, safety or fundamental rights that may emerge during the intended use of the AI system or in conditions of reasonably foreseeable misuse.[349] In a positive sense, human oversight

333 *Id*., Art. 13-3(b)(ii).
334 *Id*., Art. 13-3(b)(iii).
335 *Id*., Art. 13-3(b)(iv).
336 *Id*., Art. 13-3(b)(v).
337 *Id*., Art. 13-3(c).
338 *Id*., Art. 13-3(d).
339 *Id*., Art. 13-3(e).
340 *Id*., Art. 11.
341 *Id*., Art. 12-1.
342 *Id*., Art. 12-2.
343 *Id*., Art. 12-3.
344 *Id*., Art. 11-4(a).
345 *Id*., Art. 11-4(b).
346 *Id*., Art. 11-4(c).
347 *Id*., Art. 11-4(d).
348 *Id*., Art. 14(1).
349 *Id*., Art. 14(2).

should be ensured through measures such as identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service,[350] identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the user.[351] These measures are meant to enable individuals to whom human oversight is assigned to fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;[352] remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;[353] be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;[354] be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;[355] be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure.[356]

No doubt for convenience, the rationale which transpires from these 'families of standards' is one of *epistemic parallelism* by virtue of which *procedures designed to increase AI accuracy should yield accurate and explainable outcomes*. But is this parallelism tenable? Though several factors can explain the EU's *penchant* for standardization, it is open to criticism on namely *three points*: 1. the technical standards are descriptive and vaguely worded. Presumably, even if the AI Act did not set out a duty of transparency, software engineers would still abide by it as a *deontic requirement* in their sector of activity; 2. the procedures/outcomes parallelism as underlying rationale of the AI Act is somewhat naïve. Bearing in mind our observations on the epistemology of AI knowledge construction,[357] there are no absolute guarantees that systems' conformity to technical standards will prevent them from 'deviating' from their original programming; 3. the parallelism assumption seems to have shaped regulators' view of how to achieve explaninability. The propositional (if/then) logic that characterizes this view can be summarized as follows: *if* there is compliance with the standards in the AI Act *then* AI output is accurate and explainable (statement labelled as true); a natural or legal person has complied with the AI Act (premise), a system's output is surely accurate and explainable (conclusion).

The peculiarity of this reasoning is that explainability becomes a *byproduct of lawfulness*. On the one hand, this is not surprising. When legislation includes series of technical standards, those are presumably drawn from existing business practices of, say, manufacturing a specific type of products. Through their translation into law, those standards acquire the authority of the law and generate mandatory requirements which serve as referents for the assessment of the legality of market actors' conduct.

---

350 *Id*., Art. 14-3(a).
351 *Id*., Art. 14-3(b).
352 *Id*., Art. 14-4(a).
353 *Id*., Art. 14-4(b).
354 *Id*., Art. 14-4(c).
355 *Id*., Art. 14-4(d).
356 *Id*., Art. 14-4(e).
357 See *supra*, Section 2.

On the other hand however, the argument of lawfulness is not fully satisfying for the purpose of giving fact-of-the matter causal explanations. What victims *need*, in terms of understanding, is an explanation of how a system operating in a specific context developed, say, a bias. This bias may, of course, be the consequence of non-compliance with the AI Act, but it may occur even when the standards in this instrument were religiously observed. Selecting lawfulness as the be-all-end-all factor for accurate AI output is too limiting in cases where the cause of AI-related harm may reside with a system having acted alone. *Ad hoc* explainability provides understanding on what *ought to be done* for AI output to be explainable; it does not necessarily deliver understanding on the decisional process that led to an output which failed to be explainable. For that type of understanding to be given, *post hoc* explainability is paramount, translating to several (some sophisticated and complex) explanatory methods and techniques experts apply once - possibly harmful - AI output has been produced.

2. Post-Hoc Explainability: Experiencing Opacity, Attempting Explanation

The impression one has when reading the AI Act is that of a binary view of explainability: a system is either created transparent and is therefore explainable, or it is not. In software engineering, explainability, especially *post hoc* explainability is a *spectrum*. The nature and feasibility of *post hoc* explanations are largely dictated by the complexity of the models used in the programing of AI systems. The general rule of thumb is not difficult to understand: the more 'linear' the model (*i.e.* where the association between variables is continuous), the more transparent and explainable the system. From the perspective of AI programming, there are several techniques available:

text explanations,[358] visualizations,[359] local explanations,[360] explanations by example,[361] explanations by simplification[362] and feature relevance.[363]

Barredo Arrieta *et al*.[364] produced a well-documented study showcasing the various reasoning models and corresponding levels of explainability. There are, indeed, models that can reliably be qualified as transparent are explainable. They generally apply linear/logistic regression[365] meaning that they are rule-based and operate on the assumption of a linear dependence between predictors and predicted variables. They are 'stiff' as they do not tend to deviate from the rules which makes them predictable and transparent and their output *prima facie* explainable. This family of explainable models includes *inter alia decision trees* which are hierarchical structures used to support regression and classification. Guidotti *et al*.[366] explain that decision trees exploit a graph-structure with so-called internal nodes representing tests on features or attributes (e.g., whether a variable has a value lower than, equal to, or greater than a threshold) and so-called leaf nodes representing a decision. Each 'branch' is a possible outcome. The connections from the 'root' to the 'leaves' represent the so-called

---

358 Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Expainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 88: "text explanations deal with the problem of bringing explainability for a model by means of learning to generate *text explanations* that help explaining the results from the model. *Text explanations* also include every method generating symbols that represent the functioning of the model. These symbols may portrait the rationale of the algorithm by means of a semantic mapping from model to symbols."

359 *Ibid*: "Visual explanation techniques for post-hoc explainability aim at visualizing the model's behavior. Many of the visualization methods existing in the literature come along with dimensionality reduction techniques that allow for a human interpretable simple visualization. Visualizations may be coupled with other techniques to im- prove their understanding, and are considered as the most suitable way to introduce complex interactions within the variables involved in the model to users not acquainted to ML modeling."

360 *Ibid*: "local explanations tackle explainability by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model. These explanations can be formed by means of techniques with the differentiating property that these only explain part of the whole system's functioning."

361 *Ibid*: "Explanations by example consider the extraction of data examples that relate to the result generated by a certain model, enabling to get a better understanding of the model itself. Similarly to how humans behave when attempting to explain a given process, *explanations by example* are mainly centered in extracting representative examples that grasp the inner relationships and correlations found by the model being analyzed."

362 *Ibid*: "*Explanations by simplification* collectively denote those techniques in which a whole new system is rebuilt based on the trained model to be explained. This new, simplified model usually attempts at optimizing its resemblance to its antecedent functioning, while reducing its complexity, and keeping a similar performance score. An interesting byproduct of this family of post-hoc techniques is that the simplified model is, in general, easier to be implemented due to its reduced complexity with respect to the model it represents."

363 *Ibid*: "feature relevance explanation methods for post-hoc explainability clarify the inner functioning of a model by computing a relevance score for its managed variables. These scores quantify the affection (sensitivity) a feature has upon the output of the model. A comparison of the scores among different variables unveils the importance granted by the model to each of such variables when producing its output. *Feature relevance* methods can be thought to be an indirect method to explain a model."

364 *Id*., at 82.

365 *Id*., at 88-90.

366 Ricardo Guidotti, Anna Monreale, Dino Pedreschi, Fosca Giannotti, "Principles of Explainable Artificial Intelligence" in Moamar Sayed-Mouchaweh (ed.), *Explainable AI Withiin the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications* (Springer, 2021) 9.

classification rules. The most common rules are the conditional if-then rules, where the 'if' clause provides a set of conditions on the input variables. If the conditions are met, the system proceeds to drawing a corresponding conclusion (the 'then' portion of the reasoning). For a list of rules, the AI "returns as the decision the consequent of the first rule that is verified. Linear models allow visualizing the *feature importance*: both the sign and the magnitude of the contribution of the attributes for a given prediction."[367] In the simplest of their flavors - Barredo Arrieta *et al.* write - trees are simulatable models, manageable by human agents: "many applications of these models fall out of the fields of computation and AI (…) meaning that experts from other fields usually feel comfortable interpreting the outputs of these models."[368] However, the authors stress that decision tress have poor generalization properties which make them less interesting for businesses. Instead, so-called *K-Nearest Neighbors (KNN)* are more attractive.

KNN learning "combines the target values of K selected neighbors to predict the target value of a given test pattern."[369] When predicting a class of a test sample, they refer to classes of its K nearest neighbors (the 'neighborhood' relation being function of distance between samples).[370] KNN models work by association, much like humans who 'learn' from new experiences by associating them to similar past experiences.[371] When confronted to new sets of data, KNN models classify them in categories of the basic dataset that are similar to the data unseen during training. The simplest use of these models is e.g. that of pattern/image recognition.[372] In principle, they are predictable and explainable, which means that, to determine why a new sample has been classified inside a group, an explainer would need to refer to that sample's neighbors to infer how a 'new' sample interacted with those.[373]

In the class of linear models, Barredo Arrieta *et al.* further mention *rule-based learning*. The systems programmed with this method generate rules to characterize the data they learn from. Those rules can be linear (e.g. if-then) or combinations of such rules. So-called *fuzzy rule-based systems* enable the definition of verbally formulated rules over imprecise domains.[374] The specificity of fuzzy reasoning models is that they depart from the standard true/false dichotomy. Propositional logic typically offers a binary view: if a premise 'A' is true, the consequent 'B' is also true. Fuzzy logic deals

---

367 *Id*., at 15.

368 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *cit. supra*, 91.

369 Mahmood Akbari, Peter Jules van Overloop, Abbas Afshar, "Clustered K Nearet Neighbor Algorithm for Daily Inflow Forecasting" (2011) 5 *Water resources management*, 1341, at 1343.

370 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 91.

371 *Ibid*.

372 See e.g. Si-Bao Chen, YU-Lan Xu, Chris H.Q. Ding, Bin Luo, "A Nonnegative Locally Linear KNN model for image recognition" (2018) 83 *Pattern Recognition*, 78.

373 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 91.

374 *Ibid*.

with degrees, rather than fixed values of truth and falsity. Fuzzy systems - Barreda Arrieta *et al.* argue - empower more understandable models, since they operate in linguistic terms and perform better than classic rule systems in context with degrees of uncertainty.[375] Those systems are used e.g. in trading, in cases where traders seek to optimize portfolios while taking into consideration several factors.[376] In principle, fuzzy models are interpretable, though problems may arise when the rules they generate are too long.[377] A design goal usually sought by a user is to be able to analyze and understand the model; the number of rules in a model clearly improves its performance but also compromises its interpretability. In addition to the number of rules, their specificity may also adversely affect interpretability: a high number of antecedents and/or consequences might become difficult to interpret.[378]

In a similar vein, *Generalized additive models (GAM)* should be mentioned. They include two variables: a *response variable* (the consequent) and *predictor variables* (antecedents). They are 'linear' because their responses depend on so-called *unknown smooth functions of predictor variables*. 'Smoothness' is function of continuous derivatives in a given set called the *differentiability class*. In essence, continuous derivates are sign of stability of the variables and tend to 'stabilize' the response variable. GAMs are thus able to infer the smooth functions whose aggregate composition approximates the predicted variable.[379] In principle, GAMs too are interpretable, allowing users to verify the importance of each variable and how it affects the predicted output. The last model Barredo Arrieta *et al.* cite as interpretable are *Bayesian networks*. They make links that represent the conditional dependencies between a set of variables and "fall below the ceiling of transparent models"[380] because they are simulatable, decomposable and algorithmically transparent.

Regarding the less or non-interpretable (because non-linear) models, Barreda *et al*. cite essentially *three families of models*. First, the so-called *tree ensembles*, *forests* and *multiple classifier systems*. These are - arguably - among the most accurate (in terms of efficacy) because they are assumed to improve *generalization* capability of single-decision trees which are usually prone to so-called overfitting.[381] To avoid overfitting, tree ensembles combine different trees to obtain an aggregated

---

375 *Ibid*

376 See e.g. Yong Zhang, Weilong Liu, Xingyu Yang, "An automatic trading system for fuzzy portfolio optimization problem with sell orders" (2022) 187 *Expert Systems with applications*, 115822.

377 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 91.

378 *Ibid*.

379 *Ibid*.

380 *Id.*, at 92.

381 *Id*., at 94 Overfitting refers to a case where a system's variance (essentially the ability to 'learn' from new data) is high, running the risk of the system taking into account elements that are irrelevant for the performance of a given task. Overfitting is usually thought to be the consequence of a system's exposure to noise *i.e.* irrelevant data. We have previously discussed overfitting in recruitment scenarios. A recruitment AI might 'overfit' if e.g. it considered that job applicants who do not update their LinkedIn status regularly are introverts, not fit to work in teams. This might cause the system to exclude such applicants from the recruitment process. The 'overfitting' would essentially stem from the fact the system would not prioritize hard skills to shortlist job applicants, but due to its exposure to 'noise' would consider as determining factors that might have little or nothing to do with a set of job requirements. See Ljupcho Grozdanovski, "In search for effectiveness and fairness in proving algorithmic discrimination in EU law," *cit. supra*, at 108.

prediction/regression.[382] Though overfitting can be avoided, the combination of models makes the interpretation of an overall ensemble more complex than that of each of its compounding elements, forcing the user to employ *post hoc* interpretation techniques such as simplification, feature relevance estimators, text explanations, local explanations and model visualizations. Simplification consists in the creation of a less complex model from a set of random samples from the labeled data. It can also include a so-called *Simplified Tree Ensemble Learner (STEL)* which - again - consists in using two models, one simple and one complex, the former being used to interpret the latter through so-called Expectation-Maximization and Kullback-Leibler divergence.[383]

Another technique is *feature relevance*, especially used in tree ensembles. Feature relevance consists in measuring the so-called *Mean Decrease Accuracy (MDA)* of a forest, when a certain_variable is randomly permuted in the out-of-bag samples. This method allows experts to determine how the usage of variable importance reflects the underlying relationships in a Random Forest. Finally, a so-called *crosswise technique* proposes a framework that poses recommendations which convert an example from one class to another. The idea here is to disentangle the variables' importance in a way that is further descriptive.[384]

The second type of less/non-interpretable models cited by Barreda *et al.* are the so-called *Support Vector Machines (SVM)* which are more complex and opaque than tree ensembles.[385] SVMs construct so-called hyper-planes (or a set of hyper-planes) in a high (or infinite) dimensional space, which can be used for classification, regression or other tasks.[386] The accuracy of SVM is a function of the distance (functional margin) between the hyperplane and the nearest training-data point of any class. The larger the margin, the lower the generalization error of the classifier[387] (namely because distance reduces noise and allows the classifier to 'zoom in' on relevant training data points). The techniques used to explain SVMs are simplification, local explanations, visualizations and explanations by example. Simplifications here include four classes. First, building of rule-based models from the support vectors of a training model. This approach consists in extracting rules from the support vectors of a trained SVM using a modified sequential covering algorithm.[388] This may yield fuzzy rules in lieu of standard, propositional rules.[389] The argument voiced by experts is that long antecedents reduce comprehensibility, and a fuzzy approach allows for a more linguistically understandable result.[390]

The second approach consists in adding an SVM's hyperplane, along with support vectors, to the components in charge with creating the rules. This translates to

---

382 Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvrod Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" *cit. supra*, at 94.

383 *Id*., at 94. The Kullback-Leibler divergence allows to measure the degree of dissimilarity between two probability distributions.

384 *Ibid*.

385 *Id*., at 95.

386 *Id*., at 95.

387 *Ibid*.

388 *Ibid*.

389 *Ibid*.

390 *Ibid*.

creating hyper-rectangles from the intersections between the support vectors and the hyper-plane.[391]

The third approach consists in adding the actual training data as a component for building the rules - this would translate to creating a clustering method to group prototype vectors for each class. This combination allows for the defining of ellipsoids and hyper-rectangles in the input space.[392]

The fourth method is using SVC to give an interpretation to SVM decisions in terms of linear rules that define the space in Voronoi sections from extracted prototypes.[393]

Finally, there are the *Deep Learning models* - multi-layer networks capable of inferring complex relations among variables.[394] Because of this, they are assumed to be highly performing, but also raise serious interpretability/explainability issues. The techniques used to increase explainability are model simplification, feature relevance estimators, text explanations, local explanations and model visualizations. Barredo Arrieta *et al.* cite, as an example, the Deep RED algorithm, which extends the decompositional approach to rule extraction (essentially splitting the neuron level) for multi-layer neural network by adding more decision trees and rules.

Among generally used simplification techniques, a method called *Interpretable Mimic Learning* is used to extract an interpretable model by means of gradient boosting trees. Experts propose a hierarchical partitioning of the feature space that reveals the rejection of unlikely class labels, until association is predicted.[395] Since simplification of multi-layer neural networks is increasingly complex as the number of layers increases, feature relevance methods have become more commonly used for increasing explainability. One approach here would be to decompose the network classification decision into contributions of its input elements. This would translate to considering each neuron as an object that can be decomposed and expanded then aggregate and back-propagate these decompositions through the network, resulting in a deep Taylor decomposition.[396]

The main takeaway from our brief - though technical - overview of *post-hoc* explainability is its *complexity*. Engineers seem to have quite the 'toolbox' of techniques and methods that can easily adapt to the type of model that requires explanation. However, none of the *post hoc* explainability techniques and methods magically delivers *accurate* explanations. Explanation methods as a *post-hoc* on black-box models are not 100% faithful to the original and often do not provide enough detail to understand how the black-box models are predicting.[397] Yet, *post hoc* explanations are perhaps those capable of providing the most convincing (plausible and probative) understanding of causation in AI liability cases. In other words, XAI is - or should be - a prerequisite to the litigants' ability to give to causal explanations when debating the

391 *Ibid*.
392 *Ibid*.
393 *Ibid.*
394 *Ibid*.
395 *Id*., at 96.
396 *Ibid*.
397 Uday Kamath, John Liu, E*xplainable Artificial Intelligence: An Introduction to Interpretable Machine Learning* (Springer, 2021) at 122.

origin of AI-related harm. As intuitively obvious as this might seem, legal views are diverging. The following Sub-Section will showcase that divergence by outlining three legal perspectives.

III. XAI, INTEGRAL TO CAUSAL EXPLANATIONS? THREE PERSPECTIVES

With our discussion of explanatory accuracy and accuracy in connection to XAI in the backdrop, the *relevant procedural question* is whether plausibly accurate (or believable) causal explanations require understanding provided through the explainability methods (*ad hoc* and *post hoc*) mentioned above. Intuitively, the answer would be 'yes.' After all, when harm is *occasioned* by the use of an AI system, it is only natural to seek to uncover the role the system played in that harm materializing. This suggests that the law - including EU law - should include a set of *procedural abilities* that would allow litigants to engage in a discovery of facts that would reveal: 1. the actual (as opposed to the presumed) causal power of the AI system to be established and explained; 2. the nature and the extent of the human involvement in the system's harmful output; 3. the agent who should be held to compensate the harm occasioned by the system. In sum, the law should give an appropriate response to the *epistemic needs* of litigants in AI liability cases, in order to support their meaningful (and effective) participation in the resolution of AI liability cases. But what exactly are those needs? To use explanatory jargon, *what type(s) of understanding* do litigants flag as necessary to play an active role in the adjudication process? The emerging caselaw, as well as the EU's regulation on data processing and AI liability reveals three perspectives.

In several studies of the General Data Protection Regulation (GDPR)[398] scholars have interpreted the so-called *right to a human explanation* as needing to yield understanding of the functionalities of an AI system, therefore include *post-hoc* explainability (**Sub-Section 4.1.**). Emerging North-American caselaw in AI liability gives an additional hint: the litigants in many judicial instances do indeed seek to understand how a given system worked, but they also flagged as necessary the understanding of the *reasons why reliance on a given AI output was justified* (**Sub-Section 4.2.**). Finally, there is the EU perspective which is peculiar: the understanding the forthcoming AI liability regulation will support is neither on a system's functionalities, nor on the reasons underlying the decision to rely on that system's output. The understanding said regulation will enable pertains to the level of compliance of defendants (programmers or users) with applicable technical standards such as those enshrined in the AI Act (**Sub-Section 4.3.**).

A. 'It's about Understanding How (A System Works)' - Experts Said

The GDPR does not explicitly mention a *right* to (human) explanation. It does, however, include a *provision on transparency*, as a necessary legal (and epistemic) precondition for explainability. The normative blueprint for the principle of transparency comes from Article 12 GDPR which states that "any communication"

---

398 Regulation n° 2016/679 of the European Parliament and of the Council, of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46 (General Data Protection Regulation - GDPR), *OJ n° L 119, 4.5.2016, p. 1.*

relating to the data subject should be given by the data controller in a "concise, transparent, intelligible and easily accessible form, using clear and plain language."[399] The meaning of transparency we can derive from this Article is not difficult to grasp: for data processing to be transparent, the data subject should have access to *relevant* information - whatever those are - which should be conveyed to them clearly. The Article 29 Working Party (A29WP) - the predecessor to the European Data Protection Board (EDPB) - made the additional connection between transparency, fairness and accountability. It stressed that "the controller *must always be able to demonstrate* that personal data are processed in a transparent manner in relation to the data subject."[400]

If we read the A29WP guidelines through the *lens of evidence*, the Working Party seems to place, on the controller, the *onus* of proving transparency. They should be able to meet this 'burden' in three key stages of a data processing cycle: *before* this process is launched (when the personal data is collected either from the data subject or otherwise obtained), *throughout* the data processing (when communicating with data subjects about their rights) and *at specific points* while processing is ongoing (say, when data breaches occur or in the case of material changes to the processing).[401] To 'demonstrate' transparency, data controllers are required to present information/communication "efficiently and succinctly"[402] and the information "should be clearly differentiated from other non-privacy related information such as contractual provisions or general terms of use."[403]

It should of course be mentioned that transparency in the context of the GDPR applies in the processing of *personal data* only. There is room for debate on whether 'transparency' as enshrined in said instrument is equivalent to transparency as interpreted in connection to AI (which could process both personal and non-personal data). This is a debate deserving of a separate study. For the purpose of this paper, we shall assume that Article 12 GDPR (as interpreted by the A29WP) gives the canon on how a *generic* duty of transparency should support explainability in any data processing context. Based on this assumption, let us zoom in on the application of this 'generic understanding' of transparency in the context of *automated* data processing. Article 22 GDPR is relevant here.

By virtue of said article, the data subject has the right *not to be subject* to a decision "based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[404] Exceptionally, automated data processing can be allowed in three cases: 1. for the entering into, or performance of, a contract between the data subject and the data controller;[405] 2. when such processing is authorized by Union or Member State law to which the controller is subject;[406] 3. when the decision is based on the data subject's

399 *Id*., Art. 12(1).
400 Article 29 Working Party, *Guidelines on transparency under Regulation 2016/679* (29 November 2017, last revised on 11 April 2018), available on: https://ec.europa.eu/newsroom/article29/items (last accessed on 20 Jan. 2023), at 5.
401 *Id*., at 6.
402 *Id*., at 7.
403 *Id*., at 7.
404 GDPR, *cit. supra*, Art. 22(1).
405 *Id*., Art. 22(2)(a).
406 *Id*., Art. 22(2)(b).

explicit consent.[407] In these 'exceptional' cases, the data controller is required to implement "suitable measures" to safeguard the data subject's rights, freedoms and legitimate interests, "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."[408]

Article 22 GDPR has been interpreted as integrating human explanation in an *entitlement* (right), though this provision does not at all address the content and scope of that explanation. It does however highlight its finality which is *procedural*: the explanation given should enable the data subject to 'contest the decision,' presumably in dispute-resolution procedures launched before a national data protection authority or a court. The A29WP's Guidelines on automated individual decision-making and profiling[409] shed more light on that which ought to be explained on the grounds of said Article. First, the Working Party stressed that the term 'right' (to an explanation) entails a "general prohibition for decision-making based solely on automated data processing," [410] the implication being that such processing is "not allowed generally."[411]

Second - and more interestingly - 'automated decision' according to A29WP is one that implies *no human involvement*: "to qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than a token gesture."[412] The Guidelines further state that this "should be carried out by someone who has the authority and competence to change the decision."[413] This type of decision should, moreover produce effects that "must be sufficiently great or important to be worthy of attention."[414] Typically, 'significant effects' are produced from, say, automatic refusal of an online credit application or e-recruiting practices without any human intervention. In essence the automated decision should have the potential to "significantly affect the circumstances, behavior or choices of the individuals concerned; have a prolonged or permanent impact on the data subject or at its most extreme, lead to the exclusion or discrimination of individuals."[415]

Third, the A29WP stated that the controller ought to provide *meaningful* information. To do so, they should "find *simple ways* to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision."[416] The information should however "be sufficiently comprehensive for the data subject to understand the reasons for the decision."[417] To make the explanation meaningful and understandable, "real, tangible examples of the type of possible effects should be given."[418]

---

407 *Id*., Art. 22(2)(c).
408 *Id*., Art. 22(3).
409 Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purpose of Regulation 2016/679 (3 October 2017, las revised on 6 February 2018), available on: https://ec.europa.eu/newsroom/article29/items (last accessed on 20 Jan. 2023).
410 *Id*., at 19.
411 *Id*., at 20.
412 *Id*., at 21.
413 *Ibid*.
414 *Ibid*.
415 *Id*., at 25.
416 *Id*., at 21 (emphasis added).
417 *Id*., at 25.
418 *Id*., at 25.

If a given data processing can qualify as 'automated decision' under Article 22 GDPR (as interpreted by the A29WG), there seem to be two types of requirements that stem from the right to human explanation. On the one hand, the explanations should be *holistic*, meaning that they can, or even should extend to all stages (before, during, after) of an automated decision process.[419] Wachter and Floridi[420] espoused this *holistic view*, arguing that Article 22 GDPR generated the following duties for the data processor: to give explanation *ex ante* (on how an AI system's functionalities), to give explanation *ex post* (on the rationale of a system's output) and to comply with existing legal obligations.

On the other hand, the A29WP seems to suggest the standard of *clarity* (and by that, understandability) warranted by Article 12 GDPR which mentions 'efficient and succinct' communication. The Working Party also coheres with the 'basic' epistemology of explanations by virtue of which, explanatory goodness depends on the level of understandability delivered which, of course, presupposes clarity of the explanation as such, and a satisfactory level of comprehensiveness on the side of the explainees.[421] Most importantly, and in line with the 'holistic' reading of Article 22 GDPR, the Working Group, as well as scholarship, seem to suggest that said Article should include both *ad hoc* and *pos hoc* explanations: a data subject should ideally understand a system's functionalities and the 'reasoning' pattern(s) it applied in the course of automated data processing.

B. 'It's about Understanding Why (A System is Accurate)' - Litigants Said

A shift from *understanding-how* (a system worked) to *understanding-why* (a system was relied upon) can be seen in the previously mentioned *Pickett*, *Loomis* and *Ewert*,[422] which is the Canadian *pendant* of *Loomis*. The appellant in *Ewert* challenged the use of five psychological an actuarial risk assessment tools used by the Correctional Service of Canada to assess an offender's psychopathy and risk of recidivism, on the basis that they were developed and tested on predominantly non-Indigenous populations and that no research confirmed that they were valid when applied to Indigenous persons. He claimed, therefore, that reliance on these tools in respect to Indigenous offenders breached the Corrections and Conditional Release Act. One of the issues raised in this case was that of 'reasonable steps' taken to produce accurate information about the risk of recidivism of indigenous people. The appellant argued that Canadian authorities had long been aware of concerns regarding the possibility of AI exhibiting cultural bias and yet took no action to confirm their validity, continuing to use them in respect to Indigenous offenders, despite the fact that research

---

419 See Article 29 Working Party, Guidelines on transparency under Regulation 2016/679, *cit. supra*, at 7. This 'holistic view' is also supported by Art. 68(c) (post-compromise) AI Act *cit. supra*, relative to the right to explanation of individual decision-making. Par. 1 of this provision states that "any affected person subject to a decision which is taken by the deployer on the basis of the output from an high-risk AI system listed in Annex III (…), and which produces legal effects or similarly significantly affects him or her in a way that they consider to adversely impact their health, safety and fundamental rights *shall have the right to request from the deployer clear and meaningful explanations on the role of the AI system in the decision-making procedure and the main elements of the decision taken*" (emphasis added).

420 Sandra Wachter, Luciano Floridi, Brent Daniel Mittelstadt, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation" (2017) 2 *Int'l Data Priv'y L.*, 1, at 3.

421 See our discussion on understandability *supra*, Sub-Section 2.1.2.

422 *Ewert vs. Canada*, 2018 SCC 30, File n° 37233, 13 June 2018.

would have been feasible. There is systemic discrimination against Indigenous offenders; for the correctional system to operate fairly and effectively - the appellant argued - the assumption that all offenders can be treated fairly by being treated the same way must be abandoned.[423]

The arguments in *Ewert* confirm the '*give me the reason*s' trend we also observed, namely in *Pickett*. The appellant essentially criticized the inertia of the Canadian authorities, arguing that they consistently relied on automated recidivism decisions, without even seeking to find evidence of their accuracy. We thus detect a plea for an explanation apt at delivering understanding of the *reasons* why a system should be viewed as accurate and reliable. The Canadian courts' evidentiary assessment was, however, stringent. To establish that the reliance on the automated tools violated the principle of "fundamental justice against arbitrariness" said courts argued that the appellant "had to show on a balance of probabilities that the (authorities') practice of using the impugned tools with respect to Indigenous offenders had no rational connection to the government objective."[424] The courts found he had not done so: "there was no evidence before the trial judge that how the impugned tools operate in the case of Indigenous offenders is likely to be different from how they operate in the case of non-Indigenous offenders that their use in respect of the former is completely unrelated to the government objective.' The trial judge could not have found, "on the evidence before him" that the impugned tools overestimate the risk posed by Indigenous inmates or lead to harsher conditions of incarceration or the denial of rehabilitative opportunities because of such an overestimation.[425] In other words, the appellant did not meet the standard of proof required to support his claims.

Ewert, like *Loomis*, is noteworthy. Though both cases include *requests to understand* the reasons justifying (human) reliance on AI output, they also showcase a harsh court scrutiny over the reality of the alleged harm. Whether it be gender discrimination in *Loomis*, or ethnic discrimination in *Ewert*, the courts required that the claimants present arguments (and explanations) going beyond mere suspicions or assertions. They requested that the claimants argue - ideally based on 'strong' evidence - that the systems concerned were, in fact, inaccurate. In both cases, the claimants failed to meet the standards of proof and of persuasion. Is this due to the fact that in both *Loomis* and *Ewert* a public interest (*i.e.* the functioning of national correctional systems) was at stake? Who knows. The lesson for the EU we can draw from both cases is that, in the future, defendants - which may be public or private persons - are likely to be called to: 1. give reasons for their reliance on AI output; 2. provide evidence that justify those reasons; 3. that evidence can include general expertise as well as explanations (e.g. local explanations) on a system's functionalities.

Another takeaway from the cited caselaw caselaw is that the reasons for reliance on AI output ought to be given when that output no human intervention/involvement in producing that output can be discerned. In the EU, the meaning of 'absence of human involvement' in connection to the concept of 'automated decisions' within the meaning of Article 22 GDPR, was open for debate. Finally, the *Schufa* case came along, dealing with a credit scoring system having refused the plaintiff's loan application based on the low probability that they might be able to reimburse the loan. In his Opinion, Advocate

423 *Id.*, at 169.
424 *Ibid.*
425 *Ibid.*

General (AG) Pikamäe[426] considered that the decision in this case could, indeed, qualify as automated: Article 22 GDPR does not specify *the form* that the decision should have, though its automatized nature should appear as a distinctive feature.

AG Pikamäe's position on this point is - dare we say - a reasonable one: according to him, the automated nature of a decision depends on the rules and practices of the credit establishment which should leave *no margin of appreciation* as regards the use of (and presumably, the reliance on) automated assessment tools of loan applications. In other words, automated decisions are 'automated' when they imply both *means of automated data processing* and *automatic human reliance*. The CJEU's ruling[427] however was rather laconic though generally converging with AG Pikamäe's Opinion. The Court stated that it was "common ground" that the activity of the loan-assessing private entity in *Schufa*, met the definition of profiling, as per Article 4(4) GDPR, because the automated establishing of a probability value pertaining to a person's credit related to a specific person and to that person's ability to repay a loan.[428] Interestingly, the CJEU seems to have interpreted the 'automated' portion of the 'automated decision' concept as pertaining to the *means* of personal data processing, without placing much emphasis on the 'absence of human involvement' part. In that regard, AG Pikamäe's Opinion is more elaborate.

Assuming that the AG had the right intuition on the automated human reliance aspect of automated decisions, it should be noted that the AI Act prescribes a duty of human control and oversight *prima facie* hinting to the fact that reliance should *never* be automatic. The point on which AG Pikamäe should probably have focused is the *possibility and effectiveness for ex post* human control, the relevant questions of fact being the following: 1. is a given automated decision the *determining factor* in making a final decision (e.g. approving loans)?; 2. would the human agent's decision been the same if no AI system was used? If the answer to both questions is 'yes' a decision could qualify as automated because it would be made in the absence of other relevant factors that could imply a decision different from that made by an AI system.

Our double test for 'reasoned automated reliance' will be mentioned further in this study. Presumably, integrating such a test in the AILD/R-PLD framework would reveal a can of worms that neither the EU legislature nor the CJEU are keen on opening. Indeed, to inquire if a human agent would have made the same decision as an AI system in a given circumstance presupposes that there be a standard (say, a variant of the reasonable person test) serving as referent for the assessment of this type of *ex hypothesi* reasoning. The discussion on the possibility for such a test to emerge is beyond the scope of this paper and will, no doubt, be developed in a future study. May it suffice stressing at this stage that, if 'automated decision' within the meaning of Article 22 GDPR means *automatic reliance on AI output* (slavish or reasoned) the *effectiveness* of the right to explanation would depend on a data subject's *ability to prove* and *explain* that reliance. If the data subject fails to do so, they might not be able to exercise the right to explanation because the decision at stake would not be considered as automated.

---

426 CJEU (Opinion - AG Pikamäe), 16 March 2023, *Schufa Holding et al*., case C-634/21, EU:C:2023:220.

427 CJEU, 7 December 2023, *Schufa Holding et al*., case C-634/21, EU:C:2023:957.

428 *Id*., pt 47.

Our goal here is not to suggest a 'new' normative interpretation of Article 22 GDPR. In the future, both the CJEU and scholarship will no doubt enlighten us more on what 'automated decisions' are in connection to the GDPR. With explanatory accuracy as *fil rouge* of this paper, our brief comment on said Article 'merely' serves the purpose of canvassing the key features expected from explanations in the context of automated data processing. The feature to keep in mind for the remainder of this article is - again - the *holistic* nature of explanations: these should concern *all the stages* of a given data processing and deliver *ad hoc* and *post hoc* understanding to the data subject.

Assuming that the GDPR is a useful referent for the explanations provided under the EU's AI regulation (AI Act, AILD and R-PLD), the victims of harm associated with high-risk AI systems should be entitled to request explanations on the transparency/explainability constraints embedded in the system (*ad hoc* explainability) as well as on the concrete unfolding of a given decisional process (*ad post* explainability). However, the procedural EU regulation of AI creates systems of evidence that only support *ad hoc* explainability. What matters is that the human agents (programmers, users, deployers, importers etc) be able to explain that they did all they could to create well-performing (transparent, robust, explainable etc) AI technologies. These are no doubt important explanations. But shouldn't the victims be the ones to decide what they need to know? If the cited North-Amercian caselaw shows us anything, it is that litigants do have the tendency to require *post hoc* explanations that is, information on how an AI system actually arrived at a decision *in concreto* (*i.e.* in their particular case). Under the relevant EU instruments, it is not a given that the disclosure of such information will be authorized, because victims are restricted as regards the *types of evidence* they can ask to have access to. As will be argued, the 'holistic' concept of explanation the GDPR seems to warrant is imperfectly (because partially) translated in AI-specific instruments like the AILD.

C. 'It's about Understanding if (Technical Standards were Observed)' - Said No One… Except the EU Legislature

A paradox characterizes the EU' forthcoming regulation of AI liability that is, the AILD and R-PLD. On the one hand, we observe *openness*: both instruments 'open up' a procedural pathway for victims of harm through the right to request disclosure of evidence. Ideally, this right is meant to provide victims with the understanding necessary for them to establish and explain the causal link between an AI system and a harm suffered, thus increasing their chances of justifying compensation. On the other hand however, we detect a *restriction*: the evidence that victims can request disclosure of is quite limited in scope. Indeed, if disclosed, that evidence can only support *ad hoc* explainability, providing understanding on whether *a priori* technical standards were complied with. When exercised, the right to request disclosure does not make available any meaningful or relevant information on a system's functionalities or decision-making processes having actually resulted in the suffering of harm (*post hoc* explainability).

The limitation to *ad hoc* explainability is, no doubt, useful because, by virtue of the cited instruments' provisions, that explainability calls for evidence based that the EU legislature deems as necessary to presume fault or defectiveness (**Sub-Section 4.3.1.**). However, a closer look at the systems of evidence in the AILD and R-PLD reveal a series of inconsistencies, which beg the question of whether the procedural

rights these instruments laudably recognize can, in practice, be conducive to an *effective*, *truly meaningful participation* in, and *fair adjudication* of AI liability disputes (**Sub-Section 4.3.2.**).

1. The Right to Request Disclosure of Evidence

The AILD creates a fault-based system, placing on the claimant the burden to prove the defendant's fault. 'Fault' is defined as "*human act or omission* which does not meet a duty of care under Union law or national law that is directly intended to protect against the damage that occurred."[429] From the perspective of liability scholarship, this definition is unsurprising: it assumes that 'faulty' behavior is equivalent to unlawful behavior which only a human agent can be accused of.

The AILD pursues a double regulatory objective: first, it seeks to establish common rules on the *disclosure* of evidence on high-risk AI systems in view of enabling claimants to "substantiate a non-contractual fault-based civil law claim for damages."[430] Second, it regulates the overall "burden of proof in the case of non-contractual fault-based civil law claims brought before national courts for damages caused by an AI system."[431]

It can be argued that the right to request disclosure of evidence in the AILD gives a specific procedural expression to the right to transparency and human explanation, originally enshrined in the GDPR. In the Directive, the beneficiaries from said right are victims of harm caused by *high-risk* AI systems. That benefit is not automatic: a claimant cannot - merely - rely on their status of (alleged) victim to request that evidence be disclosed by the defendant. On the contrary, they carry the burden of proving the merits of the case by establishing that, prior to fact disclosure request brought before a court, they had undertaken all proportionate attempts to "gather the *relevant* evidence from the defendant."[432] Only when those attempts fail, may the victim go before a national court and ask that it order the disclosure requested.

When the court finds it plausible to issue such an order, the disclosure should be "necessary and proportionate," taking into consideration the legitimate interests of all parties, in particular any limitations that might stem from the protection of trade secrets within the meaning of Directive 2016/943,[433] as well as of any confidential information related to, say, public or national security. If, after the issuing of such an order, a defendant (user or provider) fails to comply, national courts shall - and here's the kicker - "*presume their non-compliance with a relevant duty of care*,"[434] this

---

429 AILD *cit. supra*, Preamble, pt 22.
430 *Id*., Art. 1(a).
431 *Id*., Art. 1(b).
432 *Id*., Art. 1(2) (emphasis added).
433 Directive 2016/943 of the European Parliament and of the Council, of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, *OJ L 157, 15.6.2016, p. 1. Pursuant to Article 2(1) of this Directive, a 'trade secret' is interpreted as information which - cumulatively - meets three requitements: it is a secret in the sense that it is not, as a body or in the precise configuration and assembly of its components, generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question (a); it has commercial value because it is secret (b); it has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret (c).*
434 AILD, *cit. supra*, Art. 3(5) (emphasis added).

presumption being essentially justified by another presumption that "the evidence requested was intended to prove for the purposes of the relevant claim for damages."[435]

Article 3 AILD is echoed *mutatis mutandis* in Article 8 R-PLD which also recognizes a right to request disclosure of evidence. Under the R-PLD, an injured party claiming compensation for damages caused by a defective product (such as a biased AI) may bring their disclosure request before a national court. The claimants acting under the R-PLD - much like those relying on the AILD - are required to present "facts and evidence sufficient to support the plausibility of the claim for compensation."[436] Here again, national courts are bound by a principle of proportionality and the legitimate interests of the parties[437] while being mindful of any confidentiality restraints related to, say, the possibility to disclose trade secrets.[438] If the defendant refused to comply with the order to disclose evidence, the defectiveness of the product will be presumed.[439]

Though much can be said based on the sheer comparative reading of Articles 3 AILD and 8-9 R-PLD, we will limit our comments to two key points: first, the *effectiveness* of the right to request disclosure of evidence; second - and more importantly - the *conditions for the formation* of the presumptions of fault and defectiveness.

Regarding the first point, there is little doubt that, on paper, the cited Articles are laudable. They finally recognize a procedural right to access evidence, which part of scholarship has been adamantly pleading for since the early days of AI's regulatory discourse.[440] However, the *effectiveness* with which this right will or should be exercised remains unclear, mainly because of the national courts' discretion in the instruments considered. Indeed, both the AILD and R-PLD admittedly introduce minimal harmonization, not seeking to reduce or eliminate the Member States' discretionary powers. This of course comes at the risk of enhancing the disparity regarding the conditions under which disclosure of evidence can be granted: neither the AILD nor the R-PLD offer any guarantee that, say, French and German courts when applying their respective national laws, will order said disclosure *in the same* conditions.

To illustrate this risk of disparity, consider the following automated recruitment scenario. As we have argued elsewhere[441] it follows from the CJEU's caselaw that in 'ordinary' (non-automated) recruitment cases, the recruiters are under *no obligation* to disclose information on the criteria used to select job applicants.[442] Let us then imagine an applicant who suspected biased automated recruitment, following which they decided to request, from the recruiter, information on the algorithm's functionalities as well as on the profiles of the job applicants shortlisted for an interview. Indeed, to be able to argue, say, ethnic bias, a job applicant of color would need to access the selected shortlist, whose racial background would support (or not) that applicant's suspicion of

---

435 *Id*., Art. 3(5).
436 R-PLD, *cit. supra*, Art. 8(1).
437 *Id*., Art. 8(2).
438 *Id*., Art. 8(3).
439 *Id*., Art. 9(1).
440 Ljupcho Grozdanovski, "In search of effectiveness and fairness in proving algorithmic discrimination in EU law" (2021) 58 *CMLRev.*, 99.
441 *Ibid*.
442 CJEU, 19 April 2012, *Meister*, case C-415/10, EU:C:2012:217.

being discriminated against. However, recruiters are not often keen on making transparent their candidate lists and, in EU law, they have not obligation to do so, as confirmed by the CJEU in the *Meister case.*[443]

In our automated recruitment scenario, suppose the recruiter refused to disclose the information requested, pushing the applicant to request that disclosure before a court. The court's decision could go in one of two ways. On the one hand, the national judge can refer to the CJEU's *Meister* case concluding that, under EU non-discrimination law, recruiters are, indeed, not required to share information on the conditions under which recruitments had been performed. Based on this caselaw, the court could consider that: 1. bearing in mind the exceptions listed in the AILD, it would be within the employer's *legitimate interest* not to make known the criteria and procedures they followed in selecting applicants; 2. in EU non-discrimination law, recruiters are, anyway, not bound by an obligation to disclose such information. In such circumstances, it is not unreasonable to assume that a victim's request for disclosure on the grounds of AILD/R-PLD would be rejected.

On the other hand, however, the court could refer to Annex III of the AI Act which lists access to labour as a sector where high-risk systems are used.[444] To verify if the recruiter in fact complied with the AI Act, it might order that they disclose the evidence requested by the claimant… even if this meant going against the CJEU's longstanding caselaw on the recruiters' (non-existent) obligation to share recruitment information with unsuccessful job applicants.

Considering that the AILD is not yet binding, these are of course speculative observations. But they do allow us to make an important point: national courts will be left with considerable freedom to assess the grounds on which they order (or not) disclosure of evidence, the danger being that the benefit from the *right* to request such disclosure may vary from one national law to another. In the absence of specific guidelines in the AILD, the national courts' decisions may be based on a variety of criteria, ranging from the type of evidence at stake, the national procedural and data protection requirements, EU data sharing and data protection requirements, to national or the CJEU's constant caselaw in the sector(s) concerned. The vagueness of those criteria might have the effect of not always providing claimants with the *effective* possibility to access the evidence they need to launch proceedings, which is of course alarming. What if an HR system was indeed biased, but a national court decided against ordering any disclosure of evidence relative to that system? Should we accept that, due to the differences between national procedural laws, there will be cases of AI liability that will go undetected and unsanctioned?...

Second, the presumptive mechanism in Articles 3 AILD and 8-9 R-PLD is surprising from a perspective of fairness: *the defendant's refusal to disclose information seems to be interpreted as a confession of guilt* of sorts. The reasoning

---

443 *Id*., pts 13 seq.

444 AI Act, *cit. supra*, Annex III, pt 4 (post-compromise): "AI systems intended to be used for recruitment or selection of natural persons, notably for placing targeted job advertisements, screening or filtering application, evaluating candidates in the course of interviews or tests; (b) AI systems intended to be used to make or materially influence affecting the initiation, promotion and termination of work-related contractual relationships, task allocation based on individual behavior or personal traits or characteristics, or for monitoring and evaluating performance and behavior of persons in such relations."

seems to go as follows: if the defendant did not wish to share information, it must be because they 'have something to hide' in terms of their compliance with a legally prescribed duty to care or applicable safety requirements. In other words, non-compliance with a *procedural duty* (to disclose information) constitutes the basic fact (*indicium*) that gives rise to the presumption of fault *i.e*. non-compliance with a *substantive duty* (to observe applicable technical legislation). This procedural-to-substantive leap is rather 'light': it is similar to presuming that when a person skips lunch, it is because they have an eating disorder (which might be the case, but additional evidence would be needed for this inference to hold).

The peculiarity of the presumptive reasoning in the AILD and R-PLD does not end there: when a presumption of fault or of defect is established, the claimant - we might think - is discharged from further adducing any evidence of fault or defectiveness. Interestingly, this is not the case. In the AILD, the burden of *proving* fault reappears in Article 4 relative to the *presumption of causation*.

2. The Exercise of the Right to Request Disclosure of Evidence

The 'incoherence' in the exercise of the right to request disclosure of evidence finds two main expressions. In the AILD, the evidentiary status of fault is peculiar. When a victim seeks to establish it, fault can, under certain conditions, be presumed. When the victim seeks to establish causation, they are required to give several types of evidence which include… *proof* of fault. The question then becomes the following: how can a victim establish fault when fault is presumed (*i.e*. is not based on any solid evidence of *indicia*) (**A**)?

Much like the AILD, the R-PLD has an incoherence of its own. This incoherence pertains to the proof of defectiveness. Essentially understood as a failure to meet reasonable expectations of a normal functioning of an AI system (whatever 'normal' is),' defectiveness can be presumed in the same conditions as those under which fault is presumed in the AILD (*i.e*. refusal to disclose evidence requested). This begs the following question: when we presume defectiveness under the R-PLD, do we *ipso facto* presume fault under the AILD (**B**)?

a. Fault in the AILD: a Fact First Presumed Then Proven

Article 4 AILD habilitates national courts to presume the *causal link* between the fault of the defendant and a given output (or the absence thereof) by the AI system when *three cumulative conditions* are met: the *claimant has proven the fault of the defendant*,[445] it can be considered *reasonably likely* that the fault has influenced the output produced by the AI system (or the failure to produce an output),[446] the claimant has proven that the output produced by the AI system has given rise to the harm suffered.[447] Similarly, Article 9 R-PLD (titled 'Burden of proof') states that the presumption of defectiveness is established when: 1. the claimant *proves* that a defendant refused to comply with the obligation to disclose 'relevant evidence' upon a court order;[448] 2. they establish that the product did not comply with mandatory safety

445 AILD *cit. supra*, Art. 4(1)(a).
446 *Id*., Art. 4(1)(b).
447 *Id*., Art. 4(1)(c).
448 R-PLD, *cit. supra*, Art. 9(2)(a).

requirements laid down in Union law or national law, intended to protect against the risk of ham occurring;[449] 3. they establish that the harm was caused by an *obvious malfunction* of the product during normal use or under ordinary circumstances.[450]

There is much to unpack from these provisions. Let us begin by highlighting the - intentionally? - vague wording of the AILD: how could a claimant prove the 'reasonable likelihood' that the defendant's fault was causally connected to the harmful output of a given system? From the perspective of liability doctrines, the proof needed in the context of a 'reasonable likelihood' situation would involve demonstrating that the defendant's actions played a *contributing role* in (*i.e*. was a contributing cause to) a harm materializing. Judging by the wording alone of Article 4 AILD, the standard of proof seems to be low - 'reasonable likelihood' as opposed to *conclusiveness* (in civil cases, preponderance of evidence). Bearing in mind the minimal level of harmonization stemming from the AILD, we can assume that that national courts will assess 'reasonable likelihood' in reference to the standards of evidence contained in their national laws which - as argued earlier - might differ from one Member State to another, adversely affecting the effectiveness of the claimants' procedural abilities. Setting aside the disparity between the Member States' laws of evidence, let us, in an *élan* of prospection, anticipate a claimant's explanatory and evidentiary strategy in establishing this 'reasonable likelihood' standard.

Take the following hypothetical: a biometric identification system is used by a Member State's authorities to assess asylum applications. Nationals from a specific country notice they are systematically refused asylum, pushing them to suspect that the system disregards applications submitted by citizens of that country. Suppose that they decided to launch an action of discrimination on the grounds of nationality, requesting that the competent authorities disclose information about the system's accuracy. Imagine the authorities refused, pushing the national court to presume their fault under Article 3 AILD. So far, so good: by virtue of this presumption, the victim would be discharged from their duty to establish the *cause* of their harm (*i.e.* fault). The story does not stop there, however.

Under Article 4 AILD, the victim should *further argue (and prove) causation* and *harm*. To do so, they would need to *positively prove* fault. The million-dollar question is thus the following: *what is the point of presuming fault if a victim still needs to establish it when proving causation*? In other words, how can a victim prove that the defendant's conduct 'reasonably likely' impacted a system's output, if the latter refused to disclose any relevant evidence that the victim might use to argue causation?

The fact that the claimant's burden to establish fault is not really removed in the AILD, is confirmed in Article 4(2) which goes on to specify the *relevant facts* to be established by the claimant, depending on whether the defendant is a provider or a user. When the defendant is a provider, said Article states that the conditions pertaining to the proof of causation shall be met, *only where the complainant has demonstrated* that the provider or, where relevant, the person subject to the provider's obligations, failed to comply with any of the requirements laid down in Chapters 2 and 3 of Title III of the AI Act.

---

449 *Id*., Art. 9(2)(b).
450 *Id*., Art. 9(2)(c).

The claimant is called to - somehow - give evidence that supports *ad hoc* explanations, aimed at showing that if harmful output was produced, it was essentially because an AI system was ill-designed since its inception. For example, a claimant is held to present proof (and explanation) that an AI system was not developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in Article 10 (2-4) AI Act;[451] that the system "was not designed and developed" in a way that meets the transparency requirements laid down in Article 13 AI Act;[452] that it did not allow for an effective oversight by natural persons during the period in which it was in use pursuant to Article 14 of the AI Act,[453] and that it did not achieve an appropriate level of accuracy, robustness and cybersecurity pursuant to Article 15 and Article 16, point (a), of the AI Act.[454] The claimant may also establish that the necessary corrective actions were not immediately taken to bring the AI system in conformity with the obligations laid down in Title III, Chapter 2 of the AI Act or to withdraw or recall the system, as appropriate, pursuant to Article 16, point (g), and Article 21 of the AI Act.[455]

Alternatively, when the defendant is a user of an AI system, causation will be *presumed* if the claimant managed to prove that their adversary did not comply with their obligations to use or monitor the AI system in accordance with the accompanying instructions of use or, where appropriate, suspend or interrupt its use pursuant to Article 29 of the AI Act,[456] exposed the AI system to input data under its control which is not relevant in view of the system's intended purpose pursuant to Article 29(3) of the AI Act.[457]

The design of the burden for claimants in the AILD is peculiar. It allows for fault to be presumed while also requiring proof thereof so that causation can be presumed. The practical difficulty which ensues is the that of a litigant being unable to give evidence of the defendant's fault, in cases where fault was presumed *precisely because* the defendant refused to disclose evidence. It will be interesting to see how the Member States' and EU courts will deal with what appears to be a congenital incoherence of the AILD's system of evidence.

The EU legislator did foresee two circumstances where the claimants should not struggle as much for the presumption of causation to be established. First, the scenario where evidence is available, despite the defendant's refusal to give access to relevant information. Article 4(4) AILD states that, for high-risk systems, a national court shall not presume causation in cases where "the defendant demonstrates that *sufficient evidence and expertise is reasonably accessible* for the claimant to prove the causal link."[458] Presumably, this Article's refers to expert evidence similar to that used in cases like *Pickett*. To refer to our biometric identification hypothetical: the claimant could establish causation if they had access to publicly available expert reports confirming that the system used to vet asylum applications was notoriously biased. Article 4(4) AILD may be applied in line with the *factum* to *fama* shift, we discussed

451 AILD, *cit. supra*, Art. 4(2)(a).
452 *Id*., Art. 4(2)((b).
453 *Id*., Art. 4(2)(c).
454 *Id*., Art. 4(2)(d).
455 *Id*., Art. 4(2)(e).
456 *Id*., Art. 4(3)(a).
457 *Id*., Art. 4(3)(b).
458 Emphasis added.

earlier in this paper:[459] if they cannot access case- and system-specific evidence (and explanation) of causation, they could *faute de mieux* refer to *general* expert opinions which may confirm, or not the plausibility of that causation.

The second exception to the presumption of causation concerns cases dealing with systems that are not high-risk. For those, the presumption of causation shall only apply where national courts find "it *excessively difficult* for the claimant to prove the causal link."[460] Pity that the excessive difficulty exception is limited to non-high-risk systems only…

Finally, when a claim for damages is brought against a defendant who used an AI system in the courts of personal, non-professional activity, the presumption of causality shall apply only where "the defendant *materially inferred* with the conditions of the operation of the AI system if the defendant was required and able to determine the conditions of operation of the AI system and failed to do so."[461]

b. Presuming Defectiveness (Ergo Fault?) in the R-PLD

i. Defining Defectiveness: the Ambiguity of the 'Expectations of Safety'

Neither the PLD nor the revised version thereof (R-PLD) include a system of evidence organized around the notion of fault. As already mentioned, the relevant fact (*probandum*) in this instrument is *defect*, the presence of which is - in principle - independent from the manufacturer's intentional or unintentional failure to meet a legal standard of product safety.

In this context, Article 6 of the 'original' PLD defines *defectiveness* in reference to the *level of safety* consumers are entitled to expect from a product. This expectation may pertain to the presentation of the product,[462] its reasonably expected use[463] and the time when the product was put into circulation.[464] The R-PLD is slightly more elaborate on the definition of defectivenes. In the amended version of Article 6, the key referent continues to be the level of expectation of safety; however, in addition to the presentation/use/time of market placement triptych (inherited from the 'original' PDL), R-PLD includes other grounds for safety expectations which can be clustered into two families: 1. the security precautions that the manufacturer has control over and 2. the security precautions that can be 'reasonably' expected to be taken by the users.

The security precautions falling within the scope of the manufacturer's control are those that pertain to the disclosure under a "technical standardization legislation" (like the AI Act). The requirements found in this 'family' include the instructions for installation, use and maintenance;[465] where the manufacturer retains control over the product after the moment it was placed in the market, the moment in time when the

459 See *supra*, Sub-Section 2.2.2.
460 AILD, *cit. supra*, Art. 4(5) (emphasis added).
461 *Id*., Art. 4(7) (emphasis added).
462 Directive 85/374 (PLD), *cit. supra*, Art. 6(1)(a).
463 *Id*., Art. 6(1)(b).
464 *Id*., Art. 6(1)(c).
465 R-PLD, *cit. supra*, Art. 6(1)(a).

product left the control of the manufacturer;[466] product safety requirements, including safety-relevant cybersecurity requirements[467] and any intervention by a regulatory authority or by an economic operator referred to in Article 7 relating to product safety.[468]

Regarding the security precautions taken by the users, they are defined in reference to the *reasonably foreseeable use* and *misuse* of a given product;[469] the effect on the product of any ability to continue to learn after deployment;[470] the effect on the product of other products that can reasonably be expected to be used together with the product; [471] the specific expectations of the end-users for whom the product is intended.[472]

The requirements included in both families of safety expectations essentially aim at *elucidating the origin of defectiveness*. Much like the criteria for explanatory 'goodness,' defectiveness under the R-PLD is assessed against *objective criteria* (compliance with technical standards) and *subjective* ones (consumers' expectations of safety). The latter are evidentially tricky. To argue that a product had failed to meet safety expectations is to, essentially, prove a perceptible and verifiable deviation from that product's normal or intended use. Though the 'normalcy' and 'intentionality' of that use varies from case to case, the CJEU seems to - usually - consider the level of safety that a product warrants *generally* and the level of safety that consumers expect *in a specific case*. The *Boston Scientific*[473] case provides an interesting example here.

A US manufacturer of pacemakers and cardioverter defibrillators imported and marketed its products in Germany. A quality control performed after those products were released in the German market revealed the risk of premature battery depletion, resulting in loss of telemetry and/or loss of pacing output "without warning."[474] Pacemakers already used on patients were promptly replaced. However, a German insurance company assigned Boston Scientific before the German courts, requesting the payment of compensation in respect of the costs related to the implantation of the potentially defective devices. The German judges submitted questions for a preliminary ruling to the CJEU, asking if a defect could be considered as established under Article 6 PLD, if a group of products presented - merely - a risk of defectiveness (*i.e.* the defect has not yet materialized). In its response, the CJEU confirmed that the level of safety that a consumer is entitled to 'reasonably expect' is a key referent for the assessment of defectiveness.[475] With regard to medical devices, the Court stressed that "in light of their function and the particularly vulnerable situation of patients using such devices, the safety requirements for those devices which such patients are entitled to expect are particularly high."[476] Against the backdrop of this high level of expected safety, the CJEU concluded that, when there is evidence showing that a group of products *may be*

---

466 *Id*., Art. 6(1)(e).
467 *Id*., Art. 6(1)(f).
468 *Id*., Art. 6(1)(g).
469 *Id*., Art. 6(1)(b) (emphasis added).
470 *Id*., Art. 6(1)(c).
471 *Id*., Art. 6(1)(d).
472 *Id*., Art. 6(1)(h) (emphasis added).
473 CJEU, 5 March 2015, *Boston Scientific*, joined cases C-503/13 and C-504/13, EU:C:2014:2306.
474 *Id*., pt 14.
475 *Id*., pt 37.
476 *Id*., pt 39.

defective, "it is possible to classify as defective *all the products in that group or series*, *without there being any need to show that the product in question is defective*."[477]

The CJEU's ruling in *Boston Scientific* is noteworthy: the defect at issue in this case was considered proven, *based on the risk* that a group of products *might share* (as opposed to 'do share') the same defect. The Court thus recognized that there *may be a discharge* from the duty to adduce positive evidence in the presence of a strong enough presumption of defectiveness. The 'strength' of this presumption seems to be function of the type of product (pacemakers), the market in which that product is used (medical devices) and the expectations that consumers normally have in that market (high level of safety).

Assuming that *Boston Scientific* is a useful referent for the future application of the R-PLD, one cannot help but wonder if the CJEU would rely on a similar presumption of defect if it had to adjudicate a case like, say, *Loomis*? Would the Court consider COMPAS defective because of the risk - highlighted in several studies - of that system developing a bias? Intuitively, applying the *Boston Scientific* logic in *Loomis* would be an overstretch: the fact that COMPAS *may* express a bias does not mean that it will… But this was exactly what the Court ruled in *Boston Scientific*.

In principle, the discovery of a *high probability* for a defect in one pacemaker does not strongly warrant the belief that all pacemakers of a series share the same level of risk of defectiveness. Of course, the devices in *Boston Scientific* were not intelligent, performing personalized blood-pumping based on a patient's individual health chart. They were *automated*, manufactured according to standardized procedures and essentially performing the same function. The presumption of defectiveness in the cited case seems to stem from a logic that roughly goes as follows: 1. *in principle*, safe pacemakers are manufactured following rigorous protocols and high safety standards; 2. the risk of defect in one pacemaker is likely due to non-compliance with those protocols and standards; 3. it is likely that this non-compliance characterized the manufacturing of all the pacemakers in the same series; 4. a cost-benefit reasoning also shows that it is less costly to withdraw, from the market, the pacemakers from that series; 5. in light of these premises, it *may be* presumed that an entire series of pacemakers shares the same level of risk of defectiveness. Presented in this way, the CJEU's premise-to-presumption leap in *Boston Scientific* is not perfect but at least seems plausible. This plausibility is essentially warranted by the fact that pacemakers' operating and use are automated (as opposed to intelligent), which means that they present a certain level of *predictability*.

There is some doubt on whether the presumptive reasoning in *Boston Scientific* - as we presented it - can apply to high-risk AI systems for the simple reason that these can be technical standard conforming *and still be unpredictable*. A biometric-identification system performs one key function *i.e.* identification of individuals. However, the variables it might rely on for that purpose might be outside any reasonable (human) foresight. While a system may be trained in scrupulous observation of applicable technical standards, its outputs may vary depending on the contexts in which it operates. If the same system was used, by public authorities, in the screening of asylum seekers and in crime-preventing public surveillance, in the former scenario, the system may express, say, a racial bias whereas in the latter scenario, it may be perfectly

477 *Id*., pt 41 (emphasis added).

bias-free or express another bias (like gender or age). In other words, in the case of pacemakers, the *proof of a probable defect* (premature battery depletion) renders the risk of harm somewhat *predictable* and *verifiable*. In the case of a biometric identification system (or any high-risk system for that matter), the same level of predictability/verifiability cannot be applied.

Considering that the unprovability of defectiveness entails the unpredictability of AI systems' performance, the regulatory reflex in the EU was to reinforce *a priori* technical standardization in view of releasing, in the market, systems that can be plausibly - though not definitely - predictable. A term often used in the EU's regulatory jargon as referent for what might be a tolerable level of (un)predictability is the 'reasonably expected use' and 'misuse' of AI.

The European Parliament's (EP) Resolution on civil liability rules for AI, defined the notion of 'high risk' as a significant potential in an autonomously operating AI-system to "cause harm or damage to one or more persons in a manner that is random and goes beyond *what can be reasonably expected*; the significance of the potential depends on the interplay between the severity of possible harm or damage, the degree of autonomy of decision-making, the likelihood that the risks materializes and the manner and the context in which the AI system is being used."[478] For the EP, high-risk is synonymous with unpredictability ('is random and goes beyond what can be reasonably expected'). It is also an issue of degree ('significance and potential'). Intolerable levels of unpredictability are measured against several probabilities: the severity of the harm (provided it can be foreseen), the degree of autonomy and the likelihood of a risk materializing. These are, of course, general evidentiary guidelines, the concrete meaning and application of which being no doubt determined on a case-by-case basis.

The AI Act, focused on prevention of harm, mentions the *reasonably foreseeable misuse* of AI, defined as the use of a system in a way that is not "in accordance with its *intended purpose*, but which may result from *reasonably foreseeable human behavior* or *interaction* with other systems."[479] This instrument thus assumes two things: 1. that a system has a known or knowable ('intended') purpose, generating an expectation that it should operate in accordance with that purpose (e.g. recruiting workers on the basis of skill alone); 2. in light of that purpose, the system warrants a reasonably foreseeable human conduct. Both factors essentially tie into a standard understanding of human control and oversight: a predictable AI system is one that remains *within the scope of the purpose defined or intended and the risks foreseen by a human agent* (programmer or user). This observation is supported by the reading of the AI Act's provisions on risk detection and management. The risk management systems consist in integrative processes that run through the entire lifecycle of those system, and which may entail regular systematic updating. These systems include the identification and analysis of any *known* and *foreseeable* risks associated with high-risk systems;[480] estimation and evaluation of the risks that may emerge when those systems are used in accordance with their intended purpose and under conditions of "reasonably

---

478 European Parliament Resolution of 20 October 20202 with recommendations to the Commission on a civil liability for Artificial Intelligence (2020/2014(INL), *OJ C 404, 6.10.2021, p. 107,* Art. 3 of the proposed Regulation.

479 AI Act, *cit. supra*, Art. 3(13) (emphasis added).

480 *Id*., Art. 9 (2)(a).

foreseeable misuse,"[481] evaluation of other possibly arising risks based on the analysis of the data gathered from the post-market monitoring system[482] and the adoption of suitable risk management measures.[483] The risk management measures should be such that any "residual risks" (whatever those are) associated with a hazard and overall residual risk of high-risk AI systems "is judged acceptable, provided that the high-risk AI system is used in accordance with the intended purpose or under conditions of reasonably foreseeable misuse. The residual risks shall be communicated to the used."[484]

The key takeaway from the risk identification and management systems is that the so-called high risks can never be fully eliminated, but can at least be reduced to an *acceptable level*, defined in reference to that which a human can *reasonably foresee*.[485] It remains however that human *foresight* in this context is reasonable, not panoptic: harm may occur without a human agent being able to foresee the (risk of) defect which might cause it. In light of this, the R-PLD introduces a lightening of the burden to prove defectiveness using a well-known evidentiary device used in contexts of uncertainty. Enter the presumption of defectiveness.

ii. Presuming Defectiveness

A reading of the system of evidence in the R-PLD shows a multifaceted *onus probandi*. To be entitled to compensation, Article 9(1) requires that the claimant prove the defectiveness of a given product, the damage suffered and the causal link between the two. The system of evidence in said Article does not structurally differ from that defined in Article 4 PLD.[486] The novelty in the R-PLD is that it establishes a *presumption of defectiveness* when any of the following conditions (*ergo* not all of them cumulatively) are met: 1. the defendant has failed to comply with an obligation to disclose relevant evidence at their disposal;[487] 2. the claimant establishes that the product does not comply with mandatory safety requirements laid down in Union law or national law, intended to protect against the risk of the harm suffered;[488] 3. the claimant establishes that the harm was caused by an obvious malfunction of the product during the normal use or under ordinary circumstances.[489]

In the first two cases, the normative kinship between the R-PLD and the AILD is apparent: the presumption of defectiveness seems to be formed under the *same conditions* as the presumption of fault. Like fault, defectiveness is presumed when a defendant refuses to disclose evidence requested by the claimant which brings up an interesting question: *where there is presumption of fault, is there also a presumption of defectiveness*? Imagine a case of biased automated access to social benefits (a high-risk

---

481 *Id*., Art. 9(2)(b).
482 *Id*., Art. 9(2)(c).
483 *Id*., Art. 9(2)(d).
484 *Id*., Art. 9(3).
485 This is the gist of human oversight: risks to health, safety or fundamental rights should be limited to uses in accordance with a system's intended purpose of under conditions of *reasonably foreseeable misuse*, in particular when such risks persist notwithstanding the application of other requirements set out in the AI Act. See AI Act *cit. supra*, Art. 14(2).
486 PLD, *cit. supra*, Art. 4: "The injured person shall be required to prove the damage, the defect and the causal relationship between defect and damage."
487 R-PLD, *cit. supra*, Art. 9(2)(a).
488 *Id*., Art. 9(2)(b).
489 *Id*., Art. 9(2)(c).

sector in the AI Act[490]) was brought before a Member State's court. Suppose the social services concerned refused to disclose evidence on, say, compliance with the human oversight standard. That refusal would be a basic fact for both the presumption of fault *and* the presumption of defectiveness. But does this mean, in future caselaw, that the AILD and R-PLD will *apply jointly*? Only time will tell. At this stage, we can but observe that the evidentiary rationale of both instruments is the same: proof of non-compliance with technical standardization is the decisive *indicium* for both the presumption of fault and the presumption defectiveness to stand.

Second, defectiveness is presumed when the claimant shows an 'obvious malfunction of the product during the normal use or under ordinary circumstances.' Intuitively, this seems reasonable. Procedurally, it opens questions, chief among them being the proof of 'obvious malfunction.' Considering - as we did earlier - that the so-called high risks, and corresponding harms, are hardly predictable, in which circumstance would a system's malfunction be obvious? The existing caselaw shows that harm becomes manifest when it is too late *i.e*. when it had already materialized. The *Arkansas Department of Human Services v. Ledger Wood et al*.[491] case gives a good example of this.

The appellees were low-income individuals with serious physical disabilities. They were beneficiaries of a Medicaid program that provides home-based and community-based services. Registered nurses made individual assessments of the beneficiaries' needs and based on those, determined the number of hours of homecare per week. The DHS implemented a reassessment system (Resource Utilization Groups system - RUG), based solely on a set of complex computer algorithms. These algorithms took patient information gathered from 286-question ArPath assessment and placed the beneficiaries into one of twenty-three RUG tiers. It is important to stress that once a beneficiary was assigned to a tier, the nurses *had no discretion* in moving them to another tier.

It soon became apparent that the system was disastrously flawed, leaving patients without adequate care: many remained without food, in soiled clothes, were not bathed, missed key exercises, treatments and turnings, faced an increased risk of failing, became more isolated in their homes and generally suffered worsened medical conditions due to the lack of care. They brought an action under the Administrative Procedure Act (APA), arguing that the DHS did not comply with the latter. Without much difficulty, the circuit court found that the plaintiffs provided the evidence necessary to prove merits (*i.e*. the likelihood of their claims for damages being

---

490 AI Act, *cit. supra*, Annex III (post-compromise), pt 5: "(a) AI systems intended to be used by or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, including healthcare services and essential services, including but not limited to housing, electricity, heating/cooling and internet, as well as to grant, reduce, revoke, increase or reclaim such benefits and services, (b) AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud; (c) AI systems intended to be used for making decisions or materially influencing decisions on the eligibility of natural persons for health and life insurance; (d) AI systems intended to evaluate and classify emergency calls by natural persons or to be used to dispatch, or to establish priority in the dispatching of emergency firs response services, including by police and law enforcement, firefighters and medical aid, as well as of emergency healthcare patient triage system."

491 Supreme Court of Arkansas, 9 November 2017 (Opinion Delivered - Appeal from the Pulaski County Circuit Court, N° 60CV-17-442), *Arkansas Department of Human Services v. Bradley Ledger Wood et al.*, No. CV-17-183.

successful). In the appeals judgment, the appellants contested this, arguing their adversaries' failure to prove irreparable harm. Usurpingly, this argument was not found convincing. Indeed, in US caselaw, harm is 'irreparable' when it "cannot be adequately compensated by money damages or redressed in a court of law."[492] Considering the evidence adduced, the Arkansas Supreme Court found that the appellees "have provided a sufficient showing of irreparable harm to justify the circuit court's issuance of a temporary restraining order."[493]

However - and here is the interesting part - the *cause of that harm* was not the fact that the algorithm 'messed up.' It was that the DHS *made automatic reliance on its output mandatory*. This is an important point to keep in mind: the emerging caselaw shows that victims of harm are not always hostile to the use of AI systems. Their criticism is often turned toward the *level of reliance* on those systems. What they seem to look for is understanding on why a human agent presumed that an AI output was accurate and therefore reliable. Based on the explanation received (or not) they then construct, as best as they can, their own causal explanations. In *Arkansas Department of Human Services v. Ledger Wood et al.* the root of the matter was not - what the AILD would define as - fault. *No one* in this case (parties, courts) felt the need to discuss if the system used complied with relevant technical legislation, the 'fault' deriving from the reliance on the system, not its non-compliance with manufacturing standards!

With the exception of cases like *Arkansas Department of Human Services v. Ledger Wood et al.*, there will be cases (possibly the majority of them) where harm will not be as manifest. Take the topical example of a credit scoring AI: a system developed a bias against ethnic minorities, by basing its decisions namely on the applicants' places of residence.[494] Noticing - to the extent that AI can 'notice' - that credit-approved applicants historically reside in 'white areas,' the system's approval of residents in those areas was much greater than that of those living in ethnically mixed ones. In a case like this, little is self-evident both as regards the harm and the malfunction having caused it. Typically, in such a case, the best a claimant can do is *suspect* discrimination which would push them to require disclosure of evidence of that harm, allowing them to move forward with judicial proceedings.

It follows that, the systems of evidence in the AILD and R-PLD are so designed that they do not include any evidence supporting *post hoc* epxlainability. As previously mentioned, this is due to the fact that both instruments are procedural expressions of an understandable but insufficiently justified normative belief: *lawful* conduct (*i.e.* compliance with technical standards) *cannot* be the source of harm.

The 'web of presumptions' that the AILD and R-PLD establish is indeed convenient from the perspective of procedural economy but is open to criticism from the perspective of basic procedural fairness in two regards. First, there is the issue of the 'meaningfulness' of the explanations: do the AILD and R-PLD, as currently designed, support the litigants' *meaningful participation* in the resolution of AI liability disputes? Second, there is the *equality of arms* principle. When we think about AI

---

492 *Id*., at 9.

493 *Id*., at 10.

494 See Will Douglas Heaven, "Bias isnt' the only problem with credit scores – and no, AI can't help" (2021) *MIT Tech'y Rev*., available on: https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/ (last accessed on 20 Jan. 2024).

liability, we tend to focus on the victim and their ability to prove and explain causation. However, we ought not forget the *defendants i.e.* the agents who, by virtue of the AILD and R-PLD, will be presumed responsible. They too have a right to meaningfully participate in the evidentiary debate and provide the explanations necessary to make their views known. The 'hermetic nature' of the evidence systems in the AILD and R-PLD invites various critiques in terms of fairness.

IV. CRITIQUE OF THE AILD'S AND R-PLD'S EVIDENTIARY HERMETISM

To sketch out ways in which - what we call - the evidentiary hermetism of the AILD and R-PLD can be 'relaxed,' let us revisit the idea of explanatory facticity:[495] explanations, including causal ones, are fact- and context bound. Let us also recall that liability law is, in essence, a *corpus* of rules and principles that crystalized *in practice first*: presumably, people dealt with causal problems long before codified law came along to instruct litigants and courts on how to address those problems. In other words - and as already stressed - causal explanations aim at accuracy (and require evidence) so that the (fair) resolution of a dispute *can be informed*. If factual accuracy were not a prerequisite for procedural fairness, we might readily consider resolving disputes through the simple act of coin tossing.

The word of advice for the future application of the AILD and R-PLD is: *presume less, prove more, and more effectively*. In this perspective, we hinted in the Introductory portion of this paper,[496] shift from a *law-based* to a *needs-based* approach, in an attempt to 'reconnect' said instruments with the procedural needs of litigants. In this context, and based on the relevant caselaw in AI liability, one point seems beyond doubt: *post hoc explainability matters* and is even paramount for the evidence and explanations given by victims of AI-related harm (**Sub-Section 5.1.**).

As for defendants, they too should benefit from the procedural ability to receive *post hoc* explanations on a system's decisional processes. This is relevant in cases where harm occurs without the defendant having intended it, or without them having been directly involved in its occurrence. The ability to request access to evidence should - for the sake of the equality of arms principle - extend to defendants as well (**Sub-Section 5.2.**).

A. The Explanations Claimants Need: Not on Compliance with the Law, But on the Accuracy and Trustworthiness of Harmful AI Output

Bearing in mind the presumptive mechanisms enshrined in both the AILD and R-PLD, it is safe to assume that the evidentiary debates which will unfold under those instruments will largely focus on the compliance or non-compliance with the AI Act (*ad hoc* explanations). This 'straightjacketing' the debate on evidence by designating the relevant cause-harm interrelationship is a textbook example of what we earlier called *underdeterministic causal labelling*.[497] The downside is, of course, that such labelling narrows the scope of the discovery of relevant evidentiary facts, restricting the

---

495 See *supra*, Sub-Section 2.1.1.
496 See *supra*, 1 - Introduction.
497 See *supra*, Sub-Section 2.2.1.

litigants' procedural ability to give evidence and explanation *other than* that required by law. When the law declares (labels) a causal truth, it usually is dismissive of the discovery of different 'truth(s),' even if they are perhaps more accurate representations of reality than that retained by a providential legislator. This is the gist of Spinoza's 'refuge of ignorance' metaphor: a causal explanation viewed as *normative* or *nomic* will, however logically 'thin,' always trump any attempt to question its truth from the vantage point of reality. Does this mean that the EU legislature prefers the convenience of *ad hoc* explainability over the fact-accuracy that *post-hoc* explainability has the potential to provide?

Take the topical example of biased AI. In a 'wrongfulness' scenario, the parties would seek to determine if a system's output to, say, approve loans to white applicants only was due to a bias already present in the system's training data or was one the system autonomously developed. With the but-for test in mind, the question that the victim would seek to answer by giving evidence (and corresponding explanations) would be the following: "*had the system not used as criterion the applicants' place of residence, would the credit-approved applicants be the same*?"

To answer this question, they would necessarily require both *ad hoc* and *post hoc* explanations in order to have a plausible (or at least, plausibly correct) idea of what actually caused the bias. Presumably, no such debate will unfold under the AILD and R-PLD: by prescribing *unlawfulness* as a 'necessary and sufficient cause'[498] of harm, both Directives conveniently circumvent any meaningful discussion on a system's *in concreto* functioning (that is, its functioning *at the time* when the harm materialized). In short, they do seem to create a 'refuge of ignorance' in the sense that uncovering factual (causal) accuracy does not seem to be their primary concern. The AILD and the R-PLD do not offer litigants the procedural possibility to prove wrongful conduct *other than unlawfulness*. A provider's record keeping might be enlightening on the data they used to program a system but may not uncover the system's specific variable-association having resulted in, say, ethnic minorities being labelled as less likely to finish college or even get into one. *That* association is the actual cause of ethnic discrimination! Not the provider's failure to neatly keep records.

Is *post-hoc* explainability necessary at all under the AILD and R-PLD? Suppose in an 'algorithmic discrimination' scenario, experts managed to reverse-engineer biased AI output, identifying the stage in a system's decisional process where the 'glitch' happened. What would be the added value of that information for the claimant? Presumably none, in the current regulatory landscape in the EU. Neither the AILD nor the PLD give the possibility of proving machine-learnt bias through evidence showing that *no human* could be reasonably associated with a case of algorithmic discrimination.

Bearing in mind our analysis of explanatory epistemology,[499] the relevant question is the following: would the claimants need to understand how a system worked and if so, should the systems of evidence in the AILD and R-PLD include ex post explainability? For the purpose of providing fact-based causal explanations, the answer is 'yes.'

---

498 The concept of necessary and sufficient cause was discussed *supra*, 1 - Introduction.

499 See *supra*, Section 2.

Moving forward, the EU legislature and courts should probably relax their obsession with the proof of unlawfulness (*i.e*. non-compliance) and focus instead on *what litigants require* in terms of evidence and evidentiary explanations. The primary justification for this is the trend becoming apparent in the emerging caselaw on AI liability: it is not about proving (human) compliance with the law, it is about *giving reasons for (human) reliance* on harm-causing (because inaccurate) AI output. Indeed, whatever the sector concerned (tax fraud, medical misdiagnosis, [500] judicial functioning[501]) litigants look to uncover and discuss the rationales of *two interrelated decisions*: *that of the AI* and *that of the human having chosen to rely on the AI*. Explanations pertaining to AI decisions address the following question: *are there reasons justifying the belief that a system's output is accurate*? The answer to this question necessarily calls for *post hoc* explanations, delivered - as confirmed by the caselaw cited in this paper - by any means available: reverse engineering, local explainability, general explainability, general expertise on a system's accuracy…

Regarding the second (human) decision calling for explanations, the relevant question is the following: *are there reasons to justify a human agent's reliance on a given AI ouput*? To answer this question, courts tend to look at human conduct, both *ad hoc* and *post hoc*. *Ad hoc* explanations - as mentioned earlier - provide information on the (legal) standards and duties imposed on human agents in view of increasing the *trustworthiness* of a system. *Post hoc* explanations provide information on an agent's reasons to consider a system trustworthy and reliable, once output is produced.

The *Loomis* case[502] gives a good example on the necessity for both *ad hoc* and *post hoc* explanations, not only because causal explanatory epistemology requires this, but because what is at stake is the exercise of a constitutional right *i.e*. the right to be presumed innocent and not be sentenced wrongfully or based on inaccurate information.[503] Indeed, the defendant in *Loomis* contended that, unless he could review how factors were weighed and risks scored, "the accuracy of the COMPAS assessment cannot be verified."[504] He further argued that "even if statistical generalizations based on gender are accurate, *they are not necessarily constitutional*."[505]

The defendant's argument in *Loomis* is interesting: his first line of defense was to say that COMPAS's decision was inaccurate, since there was no evidence to show otherwise, *in his specific case*. It is, however, his ancillary argument that is more compelling: *even if* the decisions were found to be accurate, their application should be viewed as *unconstitutional* since reliance on those decisions would violate a fundamental right. The implication in *Loomis* is that AI output should *always* be subject to some form of *ex post* control and oversight, as well as to a comprehensive statement of reasons explaining why a human agent considered that the output was trustworthy and reliable.

---

500 See Supreme Court of Arkansas, 9 November 2017 (Opinion Delivered - Appeal from the Pulaski County Circuit Court, N° 60CV-17-442), *Arkansas Department of Human Services v. Bradley Ledger Wood et al.*, No. CV-17-183.

501 Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra.*

502 *Ibid*.

503 *Id*., pt 34.

504 *Id*., pt 53.

505 *Id*., pt 79 (emphasis added).

It is also interesting to note that in *Loomis*, neither the sentencing court, nor the Minnesota Supreme court appeared hostile to the courts' use of COMPAS. On the contrary, the sentencing court's stance was that the risk assessment performed by that system could be used as a relevant factor for (1) diverting low-risk prison bound offenders to a non-prison alternative; (2) assessing whether an offender can be supervised safely and effectively in the community; (3) imposing terms and conditions of probation, supervision, and responses to violations.[506] In this context, the sentencing court considered that risk assessment performed by COMPAS may be used to "*enhance a judge's evaluation*, weighing, and application of the other sentencing evidence in the formulation of an individualized sentencing program appropriate for each defendant."[507] However - the court cautioned - the use of a COMPAS *must be subject to limitations*.[508] Risk- and needs-assessment information should be "used in the sentencing decision to inform public safety considerations related to offender risk reduction and management. It *should not be used as an aggravating or mitigating factor in determining the severity of an offender's sanction*."[509] The court's ruling on this point is enlightening in its suggestion to distinguish between (human) *decisions* and *decisive factors* for those decisions. AI systems are decision-supporting tools, not decision-makings entities! Even when they are assumed to be accurate, decision-making power should never be fully delegated to them. In many ways, their output can be assimilated to 'standard' expertise: as any type of expert evidence, AI output should be informative, relevant, support informed decisions, but never replace human decision-making power. If a human chose to base their decisions on AI output alone, *Loomis* tells us that they would need to *give reasons* on why that choice was justified.

An emerging assessment standard of the justification of human reliance on AI is a hypothetical counterfactual test which answered the following question: *what content would a human decision have, had it not involved AI use?* This is, in essence, a question the Minnesota Supreme Court sought to answer in *Loomis*, ultimately finding that even without the use of COMPAS, the circuit court would have imposed "the exact same sentence" on the defendant. As mentioned earlier,[510] this is a counterfactual reasoning typical of the but-for test. However, the risk with such a reasoning is that it might be overly hypothetical. There is a fine line between *hypothesizing* and *presuming*[511] how a human agent would have acted, without an AI system being included in the decisional process. Elucidating the exact impact an AI had on a human decision is a complex issue, deserving of a separate study. For the purpose of this paper, may it suffice stressing that *Loomis* is perhaps foretelling of what we qualified as a *needs-based explanatory approach* to AI liability. This approach consists in

506 *Id*., pt 88.
507 *Id*., pt 92 (emphasis added).
508 *Ibid*.
509 *Ibid* (emphasis added).
510 *Ibid*.
511 The difference between a hypothesis and a presumption resides in their evidentiary status and the 'strength' of the inference each presuppose. We argued elsewhere that presumptions are (indirect) evidence, the object of which are facts which, in a normal state of affairs, appear to be a probable and a plausible substitute for a fact for which direct proof is sought, but is unavailable or difficult to adduce. For presumptive inferences to hold, they require probing evidence of *indicia* (basic facts) that support the strength (and truth value) of the presumptive inference. Unlike presumptions, hypothesis do not have the status of evidence. They pertain to *possible* states of affairs which, not needing to play the role of evidence, do not need to respond to evidentiary standards like those that *indicia* must meet, in connection to presumptions. See Ljupcho Grozdanovski, « Le Probable, le plausible et le vrai. Contribution à la théorie Générale de la présomption en droit » (2020) 84-1 *RIEJ*, 39, at 71.

providing evidence and explanations on *why* there are reasons to believe that a given AI output was accurate and why the reliance on that output was justified.

This accuracy/reliance schema is not only becoming visible in cases dealing with COMPAS, but can be also seen in disputes involving other AI systems. For example, in *Cahoo v. Fast*,[512] Michigan's Unemployment Insurance agency (UIA) had used a system to detect and punish individuals having submitted fraudulent unemployment insurance claims. The plaintiffs contented that UIA detected fraud where none existed and sent little or no notice to the plaintiffs, precluding them from launching administrative appeals in the authorized delays (30 days after receiving notice). In its defense, UIA gave a *negative evidence argument*, stating that the plaintiffs had failed to demonstrate injury-in-fact because their claims were not entirely adjudicated by the Michigan Integrated Data Automated System (MiDAS).

Indeed, MiDAS performed so-called *auto-adjudication* - a process beginning with the automated generation of a flag, resulting in the automated generation of questionnaires. It then created determination based on logic trees, followed by a notice of fraud, eventually conducive to collection of taxes due.[513] Admittedly, MiDAS is not a "marvel of artificial intelligence"[514] given that a human could perform any of those activities, except the generation of the fraud questionnaire.[515] Once a default fraud determination had been made, MiDAS automatically issued three notices: 1. a primary notice of determination which confirmed overpayment from the UIA, without providing any explanation on the reasons underlying that decision;[516] 2. another notice of determination which generally informed the claimant that their actions "misled or concealed information to obtain benefits and announced that benefits were terminated on any active claims;[517] 3. a list of overpayments, accompanied by a statutory penalty for fraudulent misrepresentation of two-to-four times the amount of overpayments.[518] MiDAS made a number of errors. One of the plaintiffs in *Cahoo* argued that she had been unaware of the fraud determination and did not learn about it until she had filed for bankruptcy 'months later' (even though, she admitted to not closely following the electronic communication sent to her by the Michigan social services).

Interestingly, like in *Loomis*, the litigants in *Cahoo* presented their grievances along two lines of reasoning. First came their arguments on MiDAS' *inaccuracy*, the allegation being that the fraud determinations were "wrongfully adjudicated based on MiDAS's rigid application of the UIA's logic trees, which led to 'automated' decisions."[519] Then came the *unjustified reliance argument*: like in *Loomis*, the plaintiffs in *Cahoo* contended that UIA had wrongfully relied on the output produced by MiDAS.

Unlike *Loomis* however, in *Cahoo*, the evidentiary debate on causation was slightly different: the court did not require a *post-hoc* explanation on MiDAS'

---

512 US District Court (Eastern District of Michigan – Southern Division), *Cahoo et al. v. Fast Enterprises et al*., case n° 17-10657.
513 *Id*., at 3.
514 *Ibid.*
515 *Id*., at 3.
516 *Id*., 4.
517 *Id*., at 4.
518 *Ibid*.
519 *Id*., at 19.

(in)accuracy. It found that the plaintiffs had sufficiently demonstrated an injury-in-fact stemming from MiDAS's rigid application of logic trees, coupled with inadequate notice procedures that are "fairly traceable" to FAST's and CSG's conduct."[520] The court's operative assumption seems to have been that *the proof of harm was, itself, proof that the AI output was inaccurate*.

Cahoo marks a teachable moment for our prospections on future AI liability cases in the EU. First, it bears repeating that the assumption in *Cahoo* is that it is AI inaccuracy that causes harm, not *non-compliance with technical standards*. Based on the elements of fact (absence of notice and of explanations on the reasons for tax fraud, violation of the right to property), it was apparent that MiDAS did not perform well, rendering plausible the assumption that harm was, indeed, the consequence of *inaccurate* output (again, viewed as a wrongful act of the system, not an unlawful act of its programmer).

Second, and based on that assumption, the evidentiary debate in *Cahoo* focused on the *allocation of liability* as the court sought to identify the agent who could be plausibly seen as responsible for MiDAS's inadequate functioning. Two candidates were considered: the provider and the user. To determine which of the two was the culprit, the court applied the '*fairly traceable*' test[521] used - as the but-for test and its variants[522] - to infer, from the evidence available, the agent who should bear the responsibility of compensating harm.

In the "nebulous land of 'fairly traceable' where "causation means more than speculative but less than but-for."[523] The allegation was, essentially, that UIA's system functioned the way it did because of its *provider's* injurious actions.[524] In an attempt to shield itself from liability, the latter asserted it merely followed the State's instructions.[525] The key criterion for identifying the liable party then became an agent's *level of discretion* and *intentionality* in the programming and/or use of MiDAS. Providing advice to a third party - the court stated - that voluntarily injures another "is *constitutionally insufficient* to expose one to liability, whereas actively participating in the injury is sufficient."[526] Taking into account the elements of fact, the court found that the harm was 'fairly traceable' *to both the provider and the user*.[527]

The *Cahoo* case clarifies aspects of *Loomis.* The basic evidentiary debates in both cases revolve around the *accuracy of the AI output* and *human reliance* on that output. However, each case deals with a different variant of that debate. *Loomis* is a good example of a debate focused on proving the reliance on (in)accurate AI decision of a *public* (*judicial) authority*. As already discussed, the Minnesota Supreme Court's reasoning can be criticized, namely for the application of the hypothetical sentencing test (seeking to determine the decision a court would have reached without the use of AI). Though in *Loomis*, the Supreme Court found no automatic reliance on COMPAS's

---

520 *Id*., at 27.
521 *Id*., at 21.
522 See *supra*, Sub-Section 2.2.2. (B).
523 US District Court (Eastern District of Michigan - Southern Division), *Cahoo et al. v. Fast Enterprises et al*., *cit. supra*, at 22.
524 *Ibid*.
525 *Id*., at 23.
526 *Id*., at 24 (emphasis added).
527 *Id*., at 27.

output, the evidence it considered to assess both the system's accuracy and the reasons for reliance[528] leave us wondering if the Court's level of scrutiny would have been higher, had the allegations been made against private parties or public bodies other than courts. After all, accusing a court of being a 'slave to the algorithm' would imply total delegation of the legal/judicial decision-making, which is a troubling and alarming thought.[529]

But *which test* should we use to determine if a court was justified in automatically relying on AI output? *Loomis* does not answer this question. Future caselaw - perhaps of the CJEU - will hopefully shed more light in this regard. In *Cahoo*, the violation of a fundamental right was also attributed to a public authority. However, unlike the Minnesota courts' use of COMPAS in *Loomis*, the Michigan unemployment agency in *Cahoo* played a more *active role* in shaping the use it wished to make of MiDAS.

An interesting thought comes to mind: are we witnessing the emergence of an *active human involvement* test? This test would seek to trace back an AI-related harm to an active (intentional) human act having had a decisive impact on a system's performance. The already mentioned *Coscia* case[530] is relevant here. Seeking proof of intent-to-harm (in the case of a high-speed trading algorithm capable of spoofing), the court's approach in *Coscia* is perhaps a precursor to a more generalized, future judicial practice. In essence, the court required that proof be adduced *until a human culprit could be found*. In *Coscia*, that human turned out to be the user. Indeed, similar to *Cahoo*, it was the programmers' testimonials in *Coscia* who confirmed that the user had instructed them to create a system able to make profit… Be it at the price of spoofing.

The 'active human impact/involvement' test, performed in cases of standard Business-to-Customer (B2C) or Business-to-Business (B2B) connections, is - and has been - characteristic of cases where those connections are made possible *via* online platforms. The *Force v. Facebook*[531] case gives an interesting example here. Several US citizens argued that Facebook provided Hamas (considered in the US as a terrorist organization) with a platform that enabled attacks in Israel. Facebook did not review or edit the posts made by its users. Its terms of service explicitly stated that the users owned all the content and information posted, and exercised control over how this information was shared through users' privacy and application settings.

The liability issue in this case was, of course, whether Facebook was responsible for the content published on its platform. To address this issue, the evidentiary debate focused on determining (*i.e.* proving and explaining) if Facebook was the 'publisher' or - merely - the 'speaker' of the content provided by Hamas. To this end, it was necessary to uncover *how* Facebook used its algorithms.[532]

---

528 See Supreme Court of Wisconsin, 13 July 2016 (decided), *State of Wisconsin v. Eric L. Loomis*, *cit. supra*

529 For an analysis of the use of automation in dispute resolution, see Bastiaan van Zelst, *The end of justice(s)?: perspectives and thoughts on (regulating) automation in dispute resolution* (Eleven Int'l Publishing, 2018).

530 See US Court of Appeals for the 7th Circuit, *US* v. *Coscia*, *cit. supra*.

531 US Court of Appeals (2d Circuit), *Force v. Facebook* (2018), n° 18-397.

532 *Id*., at 22.

The plaintiffs argued that this use fell outside the scope of publishing because "the *algorithms automate Facebook's editorial decision-making*."[533] That argument did not convince the courts who asserted that 'so long as a third party willingly provides the essential published content, the interactive service provided receives full immunity regardless of the specific edit(orial) or selection process."[534] Facebook could therefore not qualify as publisher of information, but acted as mere 'speaker' of content. Though making information more available is, indeed, an essential part of traditional publishing, it does not amount to 'developing' that information as a publisher would.[535]

Even though *Cahoo*, *Coscia* and *Force v. Facebook* address legal different issues, they share the common thread of the above-mentioned accuracy/reliance evidentiary schema, as well as a test of active (intentional) involvement of a programmer or user in shaping a system's functionalities and objectives. *This* is what litigants in cases involving AI seem to *need evidence on*! To adduce that evidence, 'systems of presumptions,' such as those in the AILD and R-PLD will not cut it. Contrary to this, North American caselaw indicates that, similar to any debate involving the proof of fault, AI liability cases demand thorough fact-finding, as exemplified by trends such as *Coscia*'s 'prove until a human is identified.' This need for proper fact-finding is understandable from the standpoint of the right to a fair trial. First, for a fair adjudication, causation must, indeed, be established through fact-based explanations, ensuring compensation is awarded based on convincing information about the reality of the harm suffered. Second, fair trials maintain their 'fairness' by guaranteeing the equality of arms *for both parties*, including those presumed liable under AILD and R-PLD.

B. The Forgotten Actors in AI Liability Trials: the Rights of Defendants

According to the CJEU, the equality of arms principle is an important "corollary" the right to a fair trial.[536] In essence, this principle presupposes a level of procedural symmetry between the parties, in particular in three regards: 1. the *allocation of procedural duties* (burdens, standards of proof); 2. the *access to relevant information* and *knowledge* (in other words, evidence) able to support of their claims; 3. *equal opportunity* to make their views known and respond to the adversary's arguments. The CJEU has recognized that, in some instances, the procedural parity between the parties in a dispute may not be absolute. Admitted limitations to the right to access evidence may pertain to the content of the evidence concerned and the safeguard of constitutional

533 *Id*., at 38 (emphasis added).
534 *Id*., at 38 (emphasis added).
535 *Id*., at 49.
536 See, *inter alia*, Gen. Court, 16 July 2014, *Isotsis v. Commission*, case T-59/11, EU:T:2014:679, pt 262.

principles like the good administration (of ongoing administrative procedures or pending trials).[537]

It remains however, that save in exceptional circumstances, the parties' equal procedural footing should be observed, allowing them to benefit from the same level of - what procedural scholars have termed - *fitness to plead*.[538] The big question is, of course, if the AILD and R-PLD comply with this (constitutionally required) level of equality? To answer this question, let us bring forth the already discussed procedural postulate both instruments share: the defendant's refusal to disclose evidence in connection to their compliance with technical legislation is enough to generate a presumption of responsibility. But which evidence could they provide in order *to rebut* that presumption?

Between the AILD and the R-PLD, the former is by far the more laconic. Indeed, Article 4(7) AILD states that "the defendant shall have the right to rebut the presumption laid down in paragraph 1." This *pro forma* recognition of the right to defense points to the fact that the AILD is largely focused on regulating the burden of the claimants, though it does not pay much attention to the *feasibility* of that burden, for the reasons previously mentioned.[539] The only point where feasibility is taken into consideration is in cases dealing with the proof of causation in connection to AI systems which do not qualify as high-risk under the AI Act. For those, Art. 4(5) AILD states that said presumption shall apply only where "the national court considers it *excessively difficult* for the claimant to prove the causal link."[540]

The AILD's assumption on defendants seems to be that they, as primary bearers of the legal duty to comply with instruments like the AI Act, are *necessarily in possession of the evidence* the claimants may request access to, and that the defendants themselves might use in their defense. This of course suggests that unlike claimants, defendants cannot request that evidence be disclosed. And why would they? As argued earlier,[541] the evidentiary debates under the AILD - and by extension, the R-PLD - will revolve around *ad hoc* explainability and be limited to debates on whether the defendants complied with relevant legislations like the AI Act. The AILD appears somewhat oblivious to the *procedural needs* of the defendants, failing to consider the

---

537 The issue of the scope of the right to access evidence has, in particular, been raised in connection to the right to access documents issued by the EU institutions - namely in the context of dispute-resolution procedures - requested by third parties (*i.e.* entities not directly concerned by a disputed involving an EU institution and adjudicated on the grounds of EU law). See e.g. CJEU, 21 September 2010, *Sweden v. API and Commission et al.*, joined cases C-514/07 P, C-528/07 P and C-532/07 P, EU:C:2010:541. A journalist association based in Sweden requested the disclosure of documents relative to infringement proceedings brought by the EC against that State. The disclosure was refused, considering that the case was still pending and that the disclosure was requested by an entity that was not party to the proceedings. Analyzing the Member States' practice on the scope of the right to give generalized and unconditional public access to evidence, AG Maduro noted, in his Opinion, that not all States recognize such access, especially when the documents requested pertain to a pending case. In practice, the exercise of this right is characterized by a search for balance between ensuring the transparency of adjudicatory procedures (including the ways in which evidence is given) and the safeguard of legitimate interests (of the parties involved in the administrative or judicial procedures concerned). See *Id.*, Opinion delivered on 1 October 2009, EU:C:2009:592, para. 29.

538 See, *inter alia*, Ronnie Mackay, Warren Brookbanks, *Fitness to Plead: International and Comparative Perspectives* (OUP, 2018).

539 See our observations on the AILD, *supra*, Sub-Section 4.3.2.

540 Emphasis added.

541 See *supra*, Sub-Section 4.3.

possibility that, like claimants, they may also require a deeper understanding of the system they have used. In other words, they might also need *post hoc* explanations to exercise the right to defense. Nevertheless, given that the evidence system in the AILD does not permit the solicitation or provision of such explanations, defendants might find themselves devoid of the practical opportunity to present evidence and articulate their perspectives. This predicament arises particularly in cases where they may not comprehend the reasoning behind their system's detrimental decision-making processes.

In contrast to the AILD, the R-PLD gives a more prominent place to defendants. In its Preamble, the R-PLD stresses that the Member States' courts should presume causation where "notwithstanding the defendant's disclosure of information, it would be excessively difficult for the claimant, in light of the technical or scientific complexity of the case, to prove its defectiveness or the causal link, or both."[542] In the interest of a fair apportionment of risk - the R-PLD continues - economic operators should be exempted from liability "*if they can prove the existence of specific exonerating circumstances*."[543]

The R-PLD indeed contains several grounds for defense. As per Article 10, the defendant can escape liability if they can prove *any* of the following: 1. if they are manufacturers or importers, they should establish that they did not place the product on the market or put it into service;[544] 2. if they are distributors, they should prove that they did not make the product available on the market;[545] 3. if it is probable that the "defectiveness that caused the damage did not exist when the product was placed in the market, put into service or, in respect to a distributor, made available on the market, or that this defectiveness came into being after that moment;"[546] 4. the defectiveness is due to compliance of the product with mandatory regulations issues by public authorities;[547] 5. when the defendant is a manufacturer, "the objective state of scientific and technical knowledge at the time when the product was placed on the market, put into service or in the period in which the product was within the manufacturer's control was not such that the defectiveness could be discovered."[548] All exemptions converge in their demand for evidence of awareness (or foreseeability) regarding the risk of harm. In scenarios (1) and (2), the defendant should prove that they were not responsible for the commercialization of a 'defective' AI, arguing their *lack of relevant knowledge* on any existing or potential risks of harm. In scenario (3), the defendant should prove that the risk of harm was unforeseeable, having emerged after the system's release in the market.

---

542 R-PLD, *cit. supra*, Preamble, pt 34. The 'technical and scientific complexity' is - according to the R-PLD - a case-by-case issue and depends on various factors such as the complex nature of a product (e.g. an innovative medical device), the complex nature of the technology use (e.g. machine learning), the complex nature of the information and data to be analyzed by the claimant and the complex nature of the causal link (e.g. the link between a pharmaceutical or food product and the onset of a health condition, or a link that, in order to be prove, would require the claimant to explain the inner workings of an AI system). See *ibid*.

543 R-PLD, *cit. supra*, pt 36 (emphasis added).

544 *Id*., Art. 10(1)(a).

545 *Id*., Art. 10(1)(b).

546 *Id*., Art. 10(1)(c).

547 *Id*., Art. 10(1)(d).

548 *Id*., Art. 10(1)(e).

Scenario (4) is peculiar because it alludes to the case - not mentioned in the AILD - of *harm occurring in spite of a manufacturer's lawful conduct* (*i.e.* compliance with mandatory technical standards). By including this, the R-PLD fills a gap in the AILD regarding actions for compensation of harm occurred in the presence of evidence showing the defendant's *lawful* conduct. Here again however, the element of knowledge/foreseeability comes into play: the defendant would presumably seek to establish that their compliance with the AI Act warranted the assumption that a system was risk-free or that the technical standards followed did not allow for a risk of harm to be reasonably foreseen.

Finally, scenario (5) makes a clear allusion to expert evidence. Referring to the 'state of scientific and technical knowledge,' a defendant could escape liability by offering expertise likely to convince a court that the risk of harm was undetectable. In our opinion, and judging by the caselaw cited throughout this paper, *expert evidence* will most certainly play a prominent role in the future evidentiary debates on AI liability in the EU. In applying the R-PLD, the Member States' and Union courts will, no doubt, be called to define the probative value of the expertise brought forth by the parties. The *Pickett* and *Loomis* cases give a glimpse into a possible 'battle of experts' which will likely become exacerbated as AI technologies continue to evolve. For each expert opinion confirming the general accuracy, reliability and trustworthiness of an AI system, there will likely be a competing study arguing the contrary. We can expect to see, in the EU, the emergence of a *probative value test* which may include criteria similar to those included in the previously discussed Bradford-Hill test.[549]

In this context, one important question remains open as regards the right to effective defense: mirroring the right of claimants to request disclosure of evidence, should the grounds for defense in the R-PLD, and even the AILD, be interpreted as including a right, for defendants, to ask for independent experts, possibly for the purpose of reverse-engineering a given AI output? It is too early to tell, namely because the cited instruments are not yet binding. However, if a defendant sought to argue that a defect (like a bias) occurred *after a system had left their sphere of control*, they would naturally need to somehow prove this. The most probative evidence here would be the opening of the 'black box' which, as *Pickett* shows, can be an arduous, time-consuming process.

The deeper question is, of course, if the systems of evidence in the instruments considered should be more permissive to *post hoc* explainability, as a set of explanatory methods and techniques conducive to *understanding* of how specific systems worked (their compliance with the AI Act notwithstanding). For the sake of ensuring high levels of fairness of future AI liability cases, we might argue that *post hoc* explainability does indeed appear to be necessary, if the aim is to allow both parties to exercise their constitutional rights with equal effectiveness. Not only should claimants be able to understand the stages of causation having resulted in harm, but defendants might, depending on the facts of a case, also require such understanding: consider a recruitment algorithm displaying an unfair bias, with neither its programmer, user and potential victim having understood the reasons and methods behind the development of that bias.

---

549 See Susan Haack, "Correlation and causation. The 'Bradford Hill criteria' in epidemiological, legal and epistemological perspective," *cit. supra*.

It is yet to be seen if, when confronted with the difficult access to certain forms of *post hoc* explainability - such as reverse-engineering - the EU courts will align with their North-American counterparts, as regards the types of expert evidence they might view as admissible when direct evidence of causation is unavailable. The shift of focus to *general expertise* on specific AI systems is, as previously discussed, open to criticism: general expert opinions can support a belief in the overall trustworthiness of an AI system, but they prove nothing on that system's performance in connection to a specific harm. To resolve this conundrum, the available caselaw points to an alternative: true, general expert opinions do not establish *in concreto* (local) AI accuracy but can justify the defendant's *reasons to rely* on that system's output. The accuracy/reliance schema reappears again; we have discussed it earlier and will not revisit it here. May it suffice stressing that there is little doubt that explanations on a system's 'inner workings' are the preferred evidence, when understanding causation in AI liability cases is concerned. What litigants *need* are not statistics on Tesla cars' performance in the last five years, nor do they need to know if the manufacturing standards of Tesla cars were complied with. What they need is understanding on why *in their case*, the car made a right instead of a left turn.

However, if that type of understanding is impossible because the evidence is not accessible, the emerging caselaw reveals a shift in the explanatory enterprise from 'understanding the machine' to 'understanding the human using the machine.' The inevitability of human agency brings us back to Spinoza: considering our observations on the EU's regulation of AI liability, are we ensnared in a refuge of ignorance?

CONCLUDING REMARKS: THE AILD, THE R-PLD AND THE REFUGE OF IGNORANCE THEY BUILT

Do the AILD and the R-PLD offer a refuge of ignorance when we grapple with causal knowledge and explanation in the field of AI liability? To answer this question, we must consider the type of knowledge about facts these instruments are conducive to.

Can litigants rely on them to request the evidence and gain the understanding *they need* to causally explain the harms suffered? Alas, no. Neither of the cited instruments includes the possibility for the parties to engage in discovery proper, for the purpose of determining if a given AI-harm association was correlative or causal.

Why is it that the AILD and R-PLD fail to support *proper* discovery and explanation of causality? We have already mentioned a key component of the answer: the 'cognitive disturbance' in acquiring causal knowledge about AI lies in the potential revelation that a harm may be causally linked to an intelligent system rather than a human agent.

The AILD and the R-PLD both grapple with the current dilemma in liability doctrines, which involves choosing between liability regimes designed around *criteria for allocation of liability* and regimes designed around *criteria of discovery*. Historically, those sets of criteria were not mutually exclusive because, prior to the advent of AI, causal truths derived from discovery would reliably trace back to human culprits. However, under the influence of AI, the long-standing belief in the responsible human can be brought into question, since it no longer holds universally (*i.e.* in all cases). In spite of this, we continue to be - so to speak - *discovery-phobic*, preferring not to delve too much into facts and, with by doing so, take the risk of uncovering that

an AI system had acted without apparent human intervention. Consider the consequences of such a discovery: if evidence showed that an intelligent system caused harm *by itself*, we would need to rethink the concept of agency as cornerstone of liability in law (criminal and tort, civil and contractual).

Given our reluctance to acknowledge that AI systems can display signs of agency, we understandably cling to what we've always known to be true: only *human* agency can, directly or indirectly, be conducive to harm. To refer back to Spinoza: our preference for the human agency principle is, in many ways, not different from *choosing to believe* that stones fall from roofs because God wants them to, not because of a combination between factors like the stone's weight, the speed of the wind and gravity.

At the end of the day, the AILD and R-PLD are really not avant-garde. Consider the operative assumptions of their systems of evidence: 1. compliance with the AI Act's provisions (especially those targeting high-risk AI systems) is enough to reduce or eliminate the risks of harm; 2. if harm does ensue, it is because (and *only* because) the AI Act (or similar legislation) was not observed; 3. agents who refuse to share information on their compliance with the AI Act - in a way - confess to being at fault or to the defectiveness of a system used. Based on these assumptions, said systems of evidence are designed in such a way that, whatever evidence and explanations the parties request and give, the resulting 'knowledge' will always showcase that a human (dis)obeyed the law, rather than uncover the factors that played into an AI system acting in the way it did.

From the perspective of the epistemology of knowledge, the AILD and R-PLD are not perfect but their underlying motives are certainly understandable. The trickier question is whether their design is *procedurally fair*, from the litigants' standpoint. This entire paper is dedicated to arguing why the answer to this question is 'no.'

As mentioned earlier, procedural fairness translates - or ought to translate - to frameworks of abilities which give tangible expression to the principle of equality, namely in the ways in which litigants give and receive evidence and (causal) explanations. Ideally, the exercise of these entitlements should support the litigants' *meaningful participation*. This concept of 'meaningfulness' - *from the perspective of individuals, not legislators*! - is a recurring theme across the points raised in this paper: we contended that a crucial element in enhancing the believability of explanations lies in their level of significance to those receiving them.[550] We also argued that the meaningfulness of evidentiary debates is largely function of how effective the litigants' abilities are in accessing the evidence and giving explanations that *they consider* as relevant for the expression of their views.

Through the prism of this idea of meaningfulness - specifically referring to the litigants' 'meaningful participation' in trials - the AILD and R-PLD are open to criticism. Following up on our *needs-based approach* to AI liability, we examined what we consider to be topical examples of the emerging caselaw, revealing a trend which shows that, from the litigants' perspective, *explanations about causation do matter*. While legal compliance is important, it is the last thing litigants (and even courts) are likely to flag as a key explanatory factor in AI liability cases. As previously argued, the emphasis

[550] See *supra*, Sub-Section 2.1.2.

in what litigants 'need to understand' is underscored in two aspects: first, the accuracy of a specific AI output (requiring explanations related to *all the factors* influencing the system's output, both *ad hoc* and *post hoc*), and second, the rationale behind why human agents believe that the output was genuinely accurate and justified reliance. In essence, litigants seek to understand the rationalities involved in a case of AI use having resulted in harm: on the one hand, the *rationality behind the automated decision*, on the other hand, the *rationality behind the human decision* to rely on it. This suggests that, for the purpose of causally explaining AI-related harm, human and non-human behaviors are viewed as components of a *single causal chain*.

In summary, *proving* and *explaining causation* is crucial for the adjudication of AI Liability cases. For the sake of accuracy, meaningful participation (of litigants) and fairness (of judicial decisions), *post-hoc* explanations should be incorporated into the causal explanations and evidence presented under the forthcoming procedural regulation in the EU. The rationale behind that integration is simple: "we don't want theories. We want facts!"[551] - a statement which holds even more weight when we consider that it is evidence and *post-hoc* explanations that provide the best opportunity for dispute resolution in the field of AI liability to be informed and by that, more fair.

[551] Doris Lessing, *The Grass is Singing* (Fourth State, ed. 2013), at 22.