EJLT European Journal of Law and Technology

The EU AI Liability Directive (AILD): Bridging Information Gaps

Marta Ziosi,* Jakob Mökander, Claudio Novelli,*** Federico Casolari,**** Mariarosaria Taddeo,***** Luciano Floridi******

Abstract

The proposed European AI Liability Directive (AILD) is an important step towards closing the 'liability gap', i.e., the difficulty in assigning responsibility for harms caused by AI systems. However, if victims are to bring liability claims, they must first have ways of knowing that they have been subject to algorithmic discrimination or other harms caused by AI systems. This 'information gap' must be addressed if the AILD is to meet its regulatory objectives. In this article, we argue that the current version of the AILD reduces legal fragmentation but not legal uncertainty; privileges transparency and disclosure of evidence of high-risk systems over knowledge of harm and discrimination; and shifts the burden on the claimant from proving fault to accessing and understanding the evidence provided by the defendant. We conclude by providing four recommendations on how to improve the AILD to address the 'liability gap' and the 'information gap'.

Keywords: artificial intelligence; EU AI Liability Directive; informational asymmetry; legal uncertainty; policy recommendations; right of access to evidence

* Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK; email marta.ziosi@sant.ox.ac.uk.
** Digital Ethics Center, Yale University, 85 Trumbull St, New Haven, CT 06511, US
*** Department of Legal Studies, University of Bologna, Via Zamboni, 27/29, 40126, Bologna, IT.
**** Department of Legal Studies, University of Bologna, Via Zamboni, 27/29, 40126, Bologna, IT.
***** Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK; Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK.
****** Digital Ethics Center, Yale University, 85 Trumbull St, New Haven, CT 06511, US; Department of Legal Studies, University of Bologna, Via Zamboni, 27/29, 40126, Bologna, IT.

1. The EU AI Liability Directive: Background, Scope and Purpose

The EU has produced several initiatives to regulate artificial intelligence (AI),[1] foremost amongst which is the proposed Artificial Intelligence Act (AIA)[2] in April 2021. The AIA proposes a comprehensive regulatory framework and an oversight structure to prevent harms from AI-related risks. However, it leaves open the question of what happens when harm occurs. For instance, the AIA does not specify redress mechanisms for cases where algorithmic decision-making leads to discriminatory outcomes, such as a hiring algorithm that systematically disadvantages candidates from certain ethnic backgrounds, thereby perpetuating racial biases in employment. Since 1999, the European Commission's liability regime relied on the EU Product Liability Directive (PLD). However, a 2018 evaluation report[3] of the PLD identified important shortcomings regarding AI; specifically, whether AI software would count as a product, new types of AI-related risks (e.g. cybersecurity breaches), and specific barriers to proving harm due to specific aspects of AI technologies (e.g. complexity, autonomy and opacity). As a result, the proposed AI Liability Directive (AILD)[4] and a revised Product Liability Directive[5] were published in September 2022.

The two directives are designed to be complementary to each other.[6] The PLD is concerned with strict liability[7] and it encompasses both physical goods and software, including AI. It is applicable to manufacturers and other entities within the supply chain, such as remanufacturers and businesses that significantly alter products, provided certain conditions are met. It also extends its coverage to defective products leading to instances of physical injury, property damage, and data loss. On the other

[1] Other examples are the revision of the General Product Safety Directive and Machinery Directive.
[2] We shall consider the first Draft (COM/2021/206 final) and the draft of compromise amendments to the same (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)).
[3] COMMISSION STAFF WORKING DOCUMENT Evaluation of Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products Accompanying the document Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the Application of the Council Directive on the approximation of the laws, regulations, and administrative provisions of the Member States concerning liability for defective products (85/374/EEC) 2018.
[4] European Commission, DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) 2022.
[5] European Commission, Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on liability for defective products 2022.
[6] Philipp Hacker, 'The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future' [2022] SSRN Electronic Journal <https://www.ssrn.com/abstract=4279796> accessed 9 January 2023.
[7] A defendant held to strict liability must compensate for damages regardless of whether they adhered to or violated a specified standard of conduct. The action that causes harm is all that is required for liability to be triggered.

hand, the AILD is concerned with fault-based liability[8] and it focuses specifically on AI systems. It encompasses claims against manufacturers, but also professional (providers) and non-professional (consumers) users. The AILD may also involve the violation of fundamental rights and primary financial loss.

The AILD will transform the legal landscape for companies designing and deploying AI systems. It aims to:

(1) harmonise legal regimes and reduce legal uncertainty;
(2) prevent liability gaps between providers and users of AI systems; and
(3) make the process of compensation for injured parties easier and more effective.

The so-called 'liability gap' refers to the difficulty of ascribing responsibility for harm caused by AI systems. Sometimes, it may be challenging to determine whether and how to allocate various types of liabilities for an AI system's misconduct among designers, developers, deployers, or users. Consider the case of algorithmic discrimination against protected categories such as sex and race. For example, it may be difficult to explain why or how a woman is shown lower-wage jobs than men by automated job ads or why people of colour are more likely to be labeled at high risk of default by credit scoring algorithms. This difficulty derives from those properties of AI systems, like complexity, autonomy and opacity, coupled with the fact that these systems frequently function within intricate socio-technical environments.[9] These factors collectively weaken accountability chains and make it difficult to determine who is responsible for discrimination.[10] Closing this gap is essential, but it is not the only relevant concern. A victim of algorithmic discrimination may be unaware that an AI system has produced that outcome or may not know that they have been discriminated against.[11] The person of colour who is denied a loan, for example, might not be aware that the bank relied on algorithms for credit scoring. The woman who is not shown the same job opportunity as a man might not even realise the difference. We refer to this lack of knowledge as the 'information gap'. As this information is essential to ascribe responsibility, we argue that this gap is primary and necessary in order to fill the accountability gap and achieve the AILD's regulatory objectives.

In theory, the AILD should close both the liability gap by reducing the 'burden of proof' for claimants and the information gap by granting claimants the right to access information about high-risk AI systems to prove faults. In practice, however, the AILD

[8] Fault-based liability is based on the notion of ascribing responsibility for a specific wrongdoing. Ascribing responsibility rests on proving that the defendant has been 'at fault', i.e., having failed to adhere to specific standards of conduct.
[9] Claudio Novelli, Mariarosaria Taddeo and Luciano Floridi, 'Accountability in Artificial Intelligence: What It Is and How It Works' [2023] *AI & SOCIETY* <https://doi.org/10.1007/s00146-023-01635-y> accessed 31 July 2023.
[10] Zoe Porter and others, 'Distinguishing Two Features of Accountability for AI Technologies' (2022) 4 *Nature Machine Intelligence* 734; Simon P Rowland and others, 'Digital Health Technology-Specific Risks for Medical Malpractice Liability' (2022) 5 *npj Digital Medicine* 1.
[11] Filippo Santoni de Sio and Giulio Mecacci, 'Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them' (2021) 34 *Philosophy & Technology* 1057.

fails to adequately address the information gap, leaving questions about informational relevance, distribution and asymmetries unanswered. The central question is whether the AILD shifts the 'burden of evidence' and grants 'the right of access to evidence' in ways that sufficiently reduce information asymmetry between claimants (e.g. users) and defendants (e.g., developers). We argue that the AILD reduces legal fragmentation but not legal uncertainty; privileges transparency and disclosure of evidence of high-risk systems over knowledge of harm and discrimination; and shifts the burden on the claimant from proving fault to accessing and understanding the evidence provided by the defendant. Therefore, we contend that the AILD – as currently conceived – is unlikely to meet its regulatory objectives and conclude with some recommendations for improving it.

2. Harmonisation and Legal (Un)Certainty

The AILD's first objective is to harmonise non-contractual civil law claims[12] for damages caused by AI across the EU to avoid legal fragmentation and reduce legal uncertainty. Its Explanatory Memorandum states that 'in the absence of EU harmonised rules for compensating damage caused by AI systems, injured persons would be faced with 27 different liability regimes' and that, currently, 'if a victim brings a claim, national courts may apply existing rules on an ad hoc basis to come to a just result for the victim in ways that cause legal uncertainty'.[13] However, the AILD's solutions still expose victims to uncertainty about legal treatment and the scope of the legal regime.

First, the AILD can achieve harmonisation only regarding procedural aspects of AI systems' liability.[14] It states that 'this Directive should not harmonise general aspects of civil liability which are regulated in different ways by national civil liability rules, such as the definition of fault or causality [and] the different types of damage that give rise to claims'.[15] Therefore, national regulators can introduce laws favouring claimants, such as complete reversals of the burden of proof. However, minimum harmonisation could lead to different treatment for claimants depending on which state they are in. It also incentivizes tech providers to act in countries with less stringent liability laws, exposing users to higher likelihoods of harm or lower guarantees for compensation.

Second, uncertainty remains regarding the AILD's material scope. Several of the specifications on the scope of application of the AILD's rely on the AI Act.[16] At the time

[12] Judicial claims unrelated to contract breaches. They include personal injury, defamation, property damage, and similar issues where one party believes it has been wronged and seeks compensation.
[13] European Commission DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) (n 4).
[14] Hacker (n 6).
[15] AILD, Explanatory Memorandum.
[16] Hacker (n 6).

of writing, the finalised version of the AIA is still pending,[17] despite the EU Parliament, the Council and the Commission entering the last stages of negotiations.[18] Primarily technical adjustments are envisioned from here onwards. However, this could indirectly affect the material scope of the AILD. For example, the AILD requirement for providers to disclose evidence applies only to 'high risk systems' as defined in Article 6 of the AIA. In one of the latest agreements, the high-risk category has been expanded to include harm to the environment. AI systems used for critical infrastructures such as water management systems or energy grids ought to be categorised as high risk where they entail a severe environmental risk.[19] This indirectly extends the AILD scope. Finally, the definition of AI in the AIA refers to the Organisation for Economic Co-operation and Development's (OECD) official definition of AI. In November 2023, EU policymakers decided to freeze discussions about it considering the OECD's decision to update it.[20] On the one hand, this increases uncertainty for companies claiming to use AI in their business.[21] Companies might not know whether their products can legitimately be marketed as AI systems under evolving definitions. On the other hand, this creates confusion for users who seek legal action upon having been wronged. For example, when a user faces biased decisions in loan rejections from a financial institution, they may be unclear whether this experience falls under algorithmic discrimination as per the AI definition, or whether it should be addressed under traditional consumer protection laws related to unfair practices.

3. AI Systems' Complexity, Autonomy and Opacity

The AILD's second objective is to close the liability gap created by AI systems' complexity, autonomy and opacity. Specifically, it enables potential claimants, who previously questioned the AI system provider without success, to request courts to 'order disclosure of evidence about specific high-risk AI systems that are suspected of having caused damage'.[22] Following disclosure, a presumption of causality for the damage can be triggered by showing a defendant's 'lack of compliance with a duty of care under Union or national law'. The defendant can rebut this presumption,

[17] Luca Bertuzzi, 'MEPs Seal the Deal on Artificial Intelligence Act' (www.euractiv.com, 27 April 2023) <https://www.euractiv.com/section/artificial-intelligence/news/meps-seal-the-deal-on-artificial-intelligence-act/> accessed 30 April 2023.
[18] Luca Bertuzzi, 'EU Policymakers Enter the Last Mile for Artificial Intelligence Rulebook' (www.euractiv.com, 25 October 2023) <https://www.euractiv.com/section/artificial-intelligence/news/eu-policymakers-enter-the-last-mile-for-artificial-intelligence-rulebook/> accessed 12 November 2023.
[19] Luca Bertuzzi, 'MEPs Seal the Deal on Artificial Intelligence Act' (n 17).
[20] Luca Bertuzzi, 'OECD Updates Definition of Artificial Intelligence "to Inform EU's AI Act"' (www.euractiv.com, 9 November 2023) <https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/> accessed 12 November 2023.
[21] Hacker (n 6).
[22] AILD, Article 3(1).

typically by demonstrating that such lack of compliance could not have resulted in the reported damage.[23]

While the order to disclose evidence applies only to high-risk systems, the AILD extends it to non-high-risk systems that are considered opaque, and it relieves from it high-risk systems on which sufficient evidence is already available through documentation pursuant to the AIA.[24] Collectively, these measures incentivise defendants to disclose information about AI systems. However, they still do not address the information gap in cases where individuals do not know that they are interacting with an AI system, let alone where they have been discriminated against.

The draft AIA requires the notification of users whenever they are interacting with an AI system (unless it is obvious from the context) or exposed to emotion recognition systems, biometric categorisation systems, foundation models or deepfakes.[25] However, it remains unclear how users should be notified. This is a critical limitation, given that the effectiveness of such communication largely depends on when and how it takes place.

To submit claims, users must either know or reasonably suspect harm and provide adequate facts and evidence to validate the likelihood of a damages claim. However, such knowledge is not necessarily easily obtained. For example, users may not know, or suspect, whether an AI system's decision results from algorithmic bias that creates unlawful discrimination, and they cannot typically access the required information from the system's output logs. Some cases, such as the refusal of a loan application, might create incentives to enquire. In other cases, the impact of discrimination is more subtle, and the outcome of the AI decision may never be questioned. When discrimination simply amounts to a woman systematically not being shown the same level of job opportunities as a man, discrimination amounts to the absence of an opportunity, rather than a denial of it. How can users be aware of and claim discrimination in such cases? While the AILD privileges transparency as a solution, it leaves a high degree of opacity on the potential victim's side.

---

[23] AILD, Article 3(5).

[24] AILD, Article 4(5) and (4).

[25] European Commission, 'Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS' (21 April 2021) <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN> accessed 16 March 2023; European Parliament, 'Compromise Amendments on the Draft Report – Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts'; Christopher Ferguson and others, 'The Regulation of Artificial Intelligence in Canada and Abroad: Comparing the Proposed AIDA and EU AI Act' (2022) <https://www.fasken.com/en/knowledge/2022/10/18-the-regulation-of-artificial-intelligence-in-canada-and-abroad> accessed 8 December 2022; Isabelle Hupont and others, 'The Landscape of Facial Processing Applications in the Context of the European AI Act and the Development of Trustworthy Systems' (2022) 12 *Scientific Reports* 10688.

4. The Burden of Evidence

The AILD's third objective is to make the compensation process easier and more effective by granting claimants the right to the disclosure of evidence about AI systems suspected of having caused harm. This right is conditional on the claimant presenting 'facts and evidence sufficient to support the *plausibility* of a claim for damages' (AILD, Article 3(1)) (our italics). Further, claimants have the right of access only to evidence deemed '*necessary* and *proportionate* to support a potential claim' (AILD, Article 3(4)) (our italics). Alongside the presumption of causality, these conditions shift the problem from proving fault to accessing and interpreting evidence.

Consider, for example, '*plausible* evidence' (our italics), where the term 'plausible' is left unspecified. Indeed, the term might be left intentionally vague to allow courts the flexibility to subjectively evaluate evidence during the AILD's application, fostering adaptability rather than rigid adherence to fixed rules. At the same time, it might be hard for a non-expert user to consider which amount or type of evidence can be considered plausible. On the one hand, this might discourage them from making a claim. On the other hand, it might leave room for a defendant with access to and knowledge of their AI system documentation to further deter them from doing so, arguing for its implausibility. This endangers the process and demands further regulatory guidance. Moreover, the AILD specifies that courts shall only order defendants to disclose evidence *if* the claimant has made all proportionate attempts to gather it from the defendant (AILD, Article 3(2)). These caveats, while preventing claim overload, presuppose victim awareness of potential harm; however, without such awareness, protections against reckless litigations might curtail the advantages of evidence disclosure and burden reversal.

Once the claimant has established their right to evidence disclosure, they must prove fault resulting from non-compliance with the AIA's requirements and obligations. For example, a presumption of causality is triggered if a high-risk AI system: (i) was not trained, validated, and tested on data sets that meet quality criteria; (ii) does not meet transparency requirements regarding its design and intended use; or (iii) does not allow for effective human oversight.[26] The AIA specifies that information about AI systems should be 'relevant, accessible and comprehensible to users'.[27] However, the AIA's requirements are both technical in nature and legal in form. Even when information about black-box algorithms is transparently disclosed, inadequacy in technical and legal literacy may make it difficult for claimants to evaluate defendants' compliance with rules.[28]

---

[26] European Commission (n 25).
[27] ibid.
[28] Samar Abbas Nawaz, 'The Proposed EU AI Liability Rules: Ease or Burden?' (*European Law Blog*, 7 November 2022) <https://europeanlawblog.eu/2022/11/07/the-proposed-eu-ai-liability-rules-ease-or-burden/> accessed 12 April 2023.

As mentioned, according to the AILD, the disclosure of evidence should be limited to what is *necessary* and *proportionate* to uphold a potential claim. This aims to balance claimants' rights with the protection of third parties' trade secrets and confidential national security information. However, it opens the door to the possibility of withholding information under vague pretences of confidentiality, which is specifically relevant given the difficulty of establishing a necessary and proportionate threshold for evidence of discrimination before it is proven. This leaves room for abuse and adds to the burden of knowledge on the claimant, a burden of ignorance.

5. Recommendations

Given the concerns raised above, we propose four recommendations to improve the AILD. R2 and R3 address the information gap more directly, and they also positively influence the liability gap due to their interrelated nature. R1 and R4 focus on broader aspects of the directives' effective implementation.

R1: *Establish targeted measures to promote coherence and complementarity in the EU-wide application of the AILD and PLD.*

As the full integration of the AILD and PLD seems unlikely,[29] clarificatory guidance and support should be provided to claimants, defendants and courts to harmonise their scope and implementation. For example, identifying contexts where it is advantageous to apply strict liability rules to AI systems, and contexts where it is more cost-effective to prevent harm by focusing on the quality of care of the provider. The directive's next evaluation should thus consider promoting the complementarity of the two Directives through targeted reviews and revisions.

R2: *Create incentives to develop and deploy AI systems that are less complex, less autonomous, and less opaque ex-ante, without over-relying on transparency requirements ex-post.*

Research shows that in high-stakes environments (e.g. medicine), inherently interpretable models should be preferred over black-box ones.[30] We recommend that EU policymakers set European-wide standards for the interpretability of high-risk models and incentivise their adoption by, for example, granting a presumption of conformity with the AIA or simplifying the conformity procedures for providers who implement them.[31] Fundamentally, this strategy recommends that EU policymakers

[29] Hacker (n 6); Luca Bertuzzi, 'The New Liability Rules for AI' (www.euractiv.com, 30 September 2022) <https://www.euractiv.com/section/digital/podcast/the-new-liability-rules-for-ai/> accessed 8 December 2022.

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

[31] A general presumption of conformity with the AIA given adherence to standards is already under discussion Huw Roberts and others, 'A Comparative Framework for AI Regulatory Policy' <https://ceimia.org/wp-content/uploads/2023/02/Comparative-Framework-for-AI-Regulatory-Policy.pdf>.

adopt a proactive stance towards accountability. This means implementing measures that anticipate and prevent failures before they occur, rather than adopting a reactive stance that focuses on redressing failures after they have happened.[32]

R3: *Strengthen the right of access to evidence for the claimant by incentivising notification of harm.*

It has been suggested that external validation of models by trusted third parties can ensure the reproducibility of results and surface biases.[33] We recommend that EU policymakers introduce incentives for providers to share information about their high-risk AI systems with trusted third parties (e.g., through independent audits)[34] to allow for the notification of specific groups about their higher likelihood of exposure to bias, discrimination or errors.

R4: *Establish a real holistic approach in defining the supranational legal framework of AI.*

As seen above, the proposed Directives heavily rely on the AIA content. Considering that, the EU legislator should adopt a holistic approach, shedding light on definitions and requirements enshrined in the AIA that are still vague or unclear, thus promoting a clearer interaction between the three different pieces of legislation. The subject-oriented nature of the proposed liability Directives, for example, may complement and perhaps further specify the wide-ranging content of the AIA.

6. Conclusion

The AILD, designed to complement the PLD, aims to close the 'liability gap' associated with harms caused by AI systems. However, our analysis reveals significant shortcomings, particularly in addressing the crucial 'information gap'. The complexity, autonomy, and opacity of AI systems create challenges in ascribing liability for algorithmic discrimination, exacerbating the difficulty of proving fault. Furthermore, victims of discrimination may be unaware that AI systems have produced unfavourable outcomes, introducing an additional layer termed the 'information gap'. The proposed AILD represents a crucial stride in bridging the 'liability gap' associated with AI-induced harms. However, our analysis highlights the pressing need to concurrently address the 'information gap' to ensure the directive achieves its intended regulatory goals.

The current iteration of the AILD, though enhancing legal harmonisation, introduces heightened legal intricacies, tilts towards prioritising transparency and evidence

[32] Novelli, Taddeo and Floridi (n 9).
[33] Benjamin Haibe-Kains and others, 'Transparency and Reproducibility in Artificial Intelligence' (2020) 586 *Nature* E14.
[34] Gregory Falco and others, 'Governing AI Safety through Independent Audits' (2021) 3 *Nature Machine Intelligence* 566; Jakob Mökander, 'Auditing of AI: Legal, Ethical and Technical Approaches' (2023) 2 *Digital Society* 49.

disclosure over recognising harm and discrimination, and places a burden on claimants to navigate and comprehend the evidence presented by defendants. To fortify the AILD and effectively close both the 'liability gap' and the 'information gap', we recommend the four key improvements outlined above. By implementing these recommendations, the AILD can evolve into a more robust framework that not only facilitates liability claims for AI-induced harms but also empowers individuals to navigate the complexities of algorithmic discrimination, thereby fostering a more just and accountable AI landscape in Europe.