FLI Position on adapting Non-Contractual Civil Liability Rules to Artificial Intelligence

AI Liability Directive* – Executive Summary

View the full version:
futureoflife.org/ai-liability-full

28 November 2023

Angelica Fernandez
policy@futureoflife.org

* Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence, COM(2022) 496 final, 28.9.2022 (AILD)

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

In March of this year, we released an open letter *that sparked a global debate on AI safety.*

Introduction

The Future of Life Institute (FLI) welcomes the opportunity to provide feedback on the European Commission's proposal to adapt non-contractual civil liability rules to artificial intelligence (AILD). Liability can play a key role in catalysing safe innovation by encouraging the development of risk-mitigation strategies that reduce the likelihood of harm before products or services are deployed into a market. Moreover, an effective liability framework protects consumers' fundamental rights and can increase their trust in and uptake of new technologies. Safety and liability are intertwined concepts. Keeping AI safe requires a coherent and strong liability framework that guarantees the accountability of AI systems.

In light of the ongoing AI Act negotiations and the discussions around the adoption of a revised Product Liability Directive (PLD)[1], we considered it timely to update the recommendations of our 2022 AI Liability Position Paper.

The European Commission proposal on non-contractual civil liability rules for AI (AILD) establishes a fault-based liability framework for all AI systems, regardless of their risk, under the proposed AI Act. The AILD covers non-contractual fault-based civil liability claims for damages caused by an output, or the absence of an output, from an AI system. A fault-based claim usually requires proof of damage, the fault of a liable person or entity, and the causality link between that fault and the damage. However, AI systems can make it difficult or impossible for victims to gather the evidence required to establish this causal link. The difficulty in gathering evidence and presenting it in an explainable manner to a judge lies at the heart of claimants' procedural rights. The AILD seeks to help claimants' to fulfill their burden of proof by requiring disclosure of relevant evidence and by mandating access, under specific circumstances, to defendants' information regarding high-risk AI systems that can be crucial for establishing and supporting liability claims. It also imposes a rebuttable presumption of causality, establishing a causal link between non-compliance with a duty of care and the AI system output, or failure to produce an output, that gave rise to the damage. This presumption aims to alleviate the burden of proof for claimants. Such a mechanism is distinct from a full reversal of the burden of proof in which the victim bears no burden, and the person presumed liable must prove that the conditions of liability are not fulfilled. Moreover, the AILD specifically addresses the burden of proof in AI-related damage claims, while national laws govern other aspects of civil liability. In this sense, the AILD focuses on the procedural aspects of liability consistent with a minimum harmonisation approach, which allows claimants to invoke more favourable rules under national law (e.g., reversal of the burden of proof). National laws can impose specific obligations to mitigate risks, including additional requirements for users of high-risk AI systems.

AILD's shortcomings are hard to overlook.[2] It falls short of what is expected for an effective AI liability framework in three crucial aspects. First, it underestimates the black box phenomena of AI systems and, therefore, the difficulties for claimants and sometimes defendants to

[1] For the proposed revision see Proposal for a Regulation of the European Parliament and of the Council on General Product Safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council, COM/2021/346 final (PLD proposal); For the original text see 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 (OJ L 210, 7.8.1985, p. 29).

[2] For a detailed analysis on the main shortcomings of the AILD and its interaction with the PLD framework, see Hacker, Philipp, The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future (November 25, 2022). Available at http://dx.doi.org/10.2139/ssrn.4279796

understand and obtain relevant and explainable evidence of the logic involved in self-learning AI systems. This situation is particularly evident for advanced general purposes AI systems (GPAIS). Second, it fails to make a distinction between the requirements for evidential disclosure needed in the case of GPAIS versus other AI systems. In a case involving GPAIS, claimants' ability to take their cases to court and provide relevant evidence will be severely undermined under a fault-based liability regime. Third, it does not acknowledge the distinct characteristics and potential for systemic risks and immaterial harms stemming from certain AI systems. It's time to acknowledge these shortcomings and work towards enhanced effectiveness (an effective possibility for parties to access facts and adduce evidence in support of their claims) and fairness (implying a proportionate allocation of the burden of proof).

To remedy these points, FLI recommends the following:

1. Strict liability for GPAIS to encourage safe innovation by AI providers.

FLI recommends a strict liability regime for general purpose AI systems (GPAIS). Strict or no fault-based liability accounts for the knowledge gap between providers, operators of a system, claimants, and the courts. It also addresses the non-reciprocal risks created by AI systems. This model incentivises the development of safer systems and placement of appropriate guardrails for entities that develop GPAIS (including foundation models) and increases legal certainty. Furthermore, it protects the internal market from unpredictable and large-scale risks.

To clarify the scope of our proposal, it is important to understand that we define GPAIS as "An AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained."[3] This definition underscores the unique capability of a GPAIS to accomplish tasks beyond its specific training. It Is worth noting that the AI Act, in its current stage of negotiation, seems to differentiate between foundation models and GPAIS. For the sake of clarity, we use "General purpose AI systems" as a future-proof term encompassing the terms "foundation model", and "generative AI". It provides legal certainty for standalone (deployed directly to affected persons) and foundational GPAIS (provided downstream to deployers or other developers). Moreover, we consider GPAIS to be over a certain threshold, allowing us to bring into scope currently deployed AI systems such as Megatron Turing MLG, Llama 2, OPT-175B, Gopher, PanGu Sigma, AlexaTM, and Falcon, among other examples. Furthermore, GPAIS can be used in high-risk use cases, such as dispatching first response services or recruiting natural persons for a job. Those cases are under the scope of high-risk AI systems. But GPAIS serves a wide range of functions not regulated by Annex III of the AI Act, which presents serious risks, for example, they can be used to develop code or create weapons.[4]

Given their emergent and unexpected capabilities, unpredictable outputs, potential for instrumental autonomous goal development, and low level of interpretability, GPAIS should be explicitly included in the scope of the AILD. GPAIS opacity challenges the basic goal of legal evidence, which is to provide accurate knowledge that is both fact-dependent and rationally

[3] This definition includes unimodal (e.g., GPT-3 and BLOOM) and multimodal (e.g., stable diffusion, GPT-4, and Dall-E) systems. It contains systems at different points of the autonomy spectra, with and without humans in the loop

[4] These risks have been acknowledged by the Hiroshima process. See, OECD (2023),'G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI.

construed.[5] This barrier triggers myriad procedural issues in the context of GPAIS that are not resolved by the mechanisms established in Art. 3 and 4 AILD. It also disproportionately disadvantages claimants who need to lift the veil of opacity of GPAIS logic and outputs. Moreover, GPAIS creates non-reciprocal risks even if the desired level of care is attained; only strict liability is sufficient to incentivise a reduction of harmful levels of activity.

There are three compelling reasons for adding strict liability to GPAIS:

1. Strict liability mitigates informational asymmetries in disclosure rules for cases involving GPAIS, guaranteeing redress and a high level of consumer protection.
2. The necessary level of care to safely deploy a GPAIS is too complex for the judiciary to determine on a case-by-case basis, leading to a lack of legal certainty for all economic actors in the market.
3. Disclaimers on liability issues and a lack of adequate information-sharing regimes between upstream and downstream providers place a disproportionate compliance burden on downstream providers and operators using GPAIS.

RECOMMENDATIONS:

- Specify the wording in Art. 1(1)(a) of the AILD so that the Directive will be applicable to GPAIS, whether or not they would otherwise qualify as high-risk AI systems.
- Include GPAIS in the definitions in Art. 2 of the AILD, and clearly define GPAIS that will be subject to strict liability.
- Add a provision to the AILD establishing strict liability for GPAIS.
- Establish a joint liability scheme between upstream and downstream developers and deployers. In order to ensure consumers are protected, all parties should be held liable jointly when a GPAIS causes damage, with compensation mechanisms allowing the injured person to recover for the total relevant damage. This is in line with Art. 11, and 12 of the PLD, and the legislator can be inspired by the GDPR and the way responsibilities for controllers and processors of data are allocated.
- Specify that knowledge of potential harm should be a standard when allocating responsibility to the different links of the value chain, whether the harm has occurred or not. Model cards on AI systems should be regarded as a standard of the knowledge of harm a GPAIS provider has on the deployment of their system in order to allocate risk.
- Clearly link the forthcoming AI Act obligations on information sharing to GPAIS in the AILD to mitigate informational asymmetries between (potential) claimants and AI developers.
- Specify that neither contractual derogations, nor financial ceilings on the liability of an AI corporation providing GPAIS are permitted. The objective of consumer protection would be undermined if it were possible to limit or exclude an economic operator's liability through contractual provisions. This is in line with Recital 42 of the PLD proposal.

[5] For a procedural law perspective on the admissibility of evidence in courts regarding AI systems cases, see Grozdanovski, Ljupcho. (2022). 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. Réseaux. N° 232-233. 99-127. 10.3917/res.232.0099; Grozdanovski, Ljupcho. (2021). In search of effectiveness and fairness in proving algorithmic discrimination in EU law. Common Market Law Review. 58. 99-136. 10.54648/COLA2021005.

2. Include commercial and non-commercial open-source[6] AI systems under the liability framework of the AILD to encourage a strong and effective liability framework.

The term "open source" is being applied to vastly different products without a clear definition.[7] The business model of some AI systems labelled as open source is also unclear. Finally, there is no consensus on which elements can be determined to characterise commercial or non-commercial open source in this new regulatory landscape. Open-source AI systems are not directly addressed in the scope of the AILD. However, there are three crucial reasons to include commercial and non-commercial open-source AI systems explicitly under the liability framework of the AILD, regardless of whether they are considered GPAIS or narrow AI systems:

1. Unlike with open source software, there is no clarity of what "open source" means in the context of AI. This introduces loopholes for unsafe AI systems to be deployed under the banner of 'open source' to avoid regulatory scrutiny.
2. Deploying AI systems under an open source license poses irreversible security risks and enables misuse by malicious actors. This compromises the effectiveness and legal certainty of the whole AI liability framework. The decentralised control of open-source systems means that any misuses or unintended consequences that arise will be extremely challenging, if not impossible, to cut off by the upstream provider. There is no clear mechanism to control the open distribution of high-risk capabilities in the case of advanced AI systems and models once they are distributed or deployed.
3. If open-source AI systems are allowed to be deployed in the market without being subject to the same rules as other systems, this would not only create an unequal playing field between economic actors but also devoid the AI liability framework of its effectiveness. It would suffice to be branded open-source to escape liability, which is already a market dominance strategy of some tech behemoths. By going the route of explicitly including all open-source AI systems in the AILD framework, this ex-post framework would contribute indirectly to the enforcement of the AI Act provisions on risk mitigation and the application of sectoral product safety regulation that intersects with the products under the scope of the EU AI Act.

RECOMMENDATIONS:

- Explicitly include in the scope of the AILD both commercial and non-commercial open-source AI systems.
- Define the elements to be considered commercial open-source AI systems in collaboration with the open-source AI community to enhance economic operators' legal certainty. Llama 2 is an example of a commercial open source, even though it is mostly not sold and its source code was not released. Therefore, it should be under the scope of the AILD.
- Carefully review and justify based on evidence if exemptions for open source are needed. If yes, explicitly address non-commercial open-source AI systems exemptions, in line

[6] For ease of understanding the term "open-source" is used as a colloquial term to refer to models with public model weights. As briefly discussed in this paper, so-called open-source AI systems don't actually provide many of the benefits traditionally associated with open-source software, such as the ability to audit the source code to understand and predict functionality.

[7] Widder, David Gray and West, Sarah and Whittaker, Meredith, Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (August 17, 2023). http://dx.doi.org/10.2139/ssrn.4543807

with other EU law instruments. For example, through licensing agreements, there could be a limited exemption in the liability framework for exclusively academic researchers, so long as they do not proliferate the liability-emitting artefacts to third parties and there are obligations to subject these systems to rigorous physical and cybersecurity access controls to prevent the deliberate or accidental leaking or proliferation of model weights. They should also be subject to external audits, red-teaming, and information-sharing obligations.

3. Establish a fault-based liability with reversed burden of proof for non-general purpose high-risk AI system.

FLI agrees with the AILD proposal in that some high-risk AI systems should fall under a fault-based liability regime. This will be the case with non-general purpose high-risk AI systems.[8] However, the presumption of fault should lie on the provider of an AI system. Pursuing this course of action would ease the burden for claimants and increase their access to justice by minimising information asymmetry and transaction costs. Providers of AI systems can rebut this presumption of fault by proving their compliance with and observance of the required level of care or by the lack of a causal link between the output and the damage. Non-compliance liability relies on the AI Act as the "backbone" of AI safety legislation for the liability framework.

As mentioned earlier, several specific characteristics of AI can make it difficult and costly for injured parties to identify and prove the fault of a potentially liable entity in order to receive compensation.[9] Harmed individuals are subject to significant information asymmetry with respect to the AI systems they interact with because they may not know which code or input caused harm. The interplay between different systems and components, the multitude of actors involved, and the increasing autonomy of AI systems add to the complexity of proving fault.[10] In this case, liability will be placed on the AI provider, the party that can reduce harm at the lowest cost.

FLI believes that a fault-based liability regime with a reversed burden of proof for non-general purpose high-risk AI systems is a sufficient and balanced approach. Following the risk-based approach of the AI Act, it seems sensible to have less stringent requirements than strict liability for these AI systems, which do not necessarily exhibit the self-learning and autonomous capabilities of GPAIS. Moreover, most of the use cases for these systems are defined narrowly in Annex III and will be subject to rigorous requirements under the AI Act. However, some non-general purpose AI systems might not be captured by Annex III of the AI Act, for this reason, we propose that the liability regime is not dependent on the high-risk categorisation of the AI Act, but that has a broader scope to fully capture risks for harm from AI providers and offer the effective possibility of redress to claimants.

RECOMMENDATIONS:

[8] Non-general purpose AI systems sometimes referenced to as high-risk narrow AI systems. As indicated before GPAIS will be subject to strict liability.

[9] Such characteristics are autonomous behaviour, continuous adaptation, limited predictability, and opacity - European Commission (2021), Civil liability – adapting liability rules to the digital age and artificial intelligence, Inception Impact Assessment, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Civil-liability-adapting-liability-rules- to-the-digitalage-and-artificial-intelligence_en.

[10] Buiten, Miriam and de Streel, Alexandre and Peitz, Martin, EU Liability Rules for the Age of Artificial Intelligence (April 1, 2021). Available at SSRN: https://ssrn.com/abstract=3817520 or http://dx.doi.org/10.2139/ssrn.3817520; Zech, H. Liability for AI: public policy considerations. ERA Forum 22, 147–158 (2021). https://doi.org/10.1007/s12027-020-00648-0.

- Modify Art. 3 (1) AILD to include a reversed burden of proof for non-general purpose high-risk AI systems.
- Establish a clear distinction between non-general purpose high-risk AI systems (also sometimes referenced to as high-risk narrow AI systems) and GPAIS in the AILD.
- Create a mechanism that aligns the AI Act regulatory authorities, such as the AI Office, with the liability framework. For example, regulatory authorities under the AI Act could also become a "one-stop shop" for AI providers, potential claimants, and lawyers seeking to obtain evidence on high-risk systems and their compliance with their duty of care under the AI Act. They will be placed advantageously to assess prima facie the level of compliance of a given non-general purpose high-risk AI system and support potential claimants' evidence requests. This "one-stop-shop" mechanism could mirror some of the features of the mechanisms under GDPR that allow for cross-border enforcement cooperation between data protection authorities.

4. Protect the fundamental rights of parties injured by AI systems by including systemic harms and immaterial damages in the scope of the AILD.

FLI calls for compensable damages to be harmonised across the EU and include immaterial and systemic harms. This recommendation is without prejudice to the liability frameworks from EU Member States and the minimum harmonisation approach that the AILD aims to achieve. FLI argues that (a) immaterial and systemic harms stemming from AI systems should be in the scope of recoverable damages, and (b) in order to ensure consistent protection of fundamental rights across Member States, immaterial, societal, and systemic harms produced by an AI system should be defined by EU law and not by national laws.

Addressing "systemic risk" and, by extension, societal-level harms, is not a new concept for the EU legislator,[11] as it has been addressed in the context of the Digital Services Act (DSA).[12] Some of the risks that AI poses are relatively small or unlikely on a per-incident basis, but together, can aggregate to generate severe, impactful, correlated, and adverse outcomes for specific communities or for society as a whole. Adding a systemic risk dimension to the proposed liability framework in the AILD, therefore, reflects fundamental rights considerations.

Along with systemic harms, we also propose that immaterial harms (also referred to as "non-material harms" or "non-material damages") be covered within the scope of the AILD. Immaterial harms refer to harms that are challenging to quantify in monetary terms, as the damage itself is of a "qualitative" nature and not directly related to a person's physical health, assets, wealth, or income. Covering immaterial harms is necessary to account for the particular nature of damages caused by AI systems, including "loss of privacy, limitations to the right of freedom of expression, human dignity, [and] discrimination, for instance in access to employment."[13] It is reasonable to consider that risks associated with AI systems can quickly scale up and affect an entire society. However, the proposed Directive leaves it up to Member

[11] Interestingly, Recital 12 of the AILD acknowledges systemic risks under the DSA framework.

[12] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277, 27.10.2022, p. 1–102.

[13] European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

States to define the damages covered. This could mean that a person discriminated against by a credit-scoring AI system could claim damages for such discrimination in one Member State but not in another.

Scholars have also proposed attaching compensation for immaterial harms to a model of non-compliance liability when deployers and operators engage in prohibited or illegal practices under the AI Act.[14] This model could fit easily into existing non-discrimination, data protection, and consumer protection legislation. For example, Article 82 of the GDPR[15] provides for the liability of a controller or processor where this entity violates their obligations under the GDPR. In this sense, the scope of application for recoverable immaterial damages will not be too broad, countering the idea that including immaterial damages disproportionately broaden liability provisions.

Explicitly including immaterial damages and systemic harms in the recitals and definitions of the AILD would enhance the protective capacity of the framework and solidify the links between the AI Act and the AILD. Notably, given that Recital 4 of the AI Act[16] explicitly recognises "immaterial" harms posed by AI, both in the European Commission and Council text. The European Parliament's mandate for the AI Act further highlights immaterial harms, mentioning "societal" harm specifically.[17] This resolution already proposed the concept of 'significant immaterial harm.'

RECOMMENDATIONS:

- Modify Recital 10 AILD to include systemic harms and immaterial damages as recoverable damages.
- Include a definition of immaterial harm in the AILD based on the AI Act, and the European Parliament's resolution.
- Include a notion of systemic risk in the AILD based on the DSA.

[14] See Wendehorst, C. (2022). Liability for Artificial Intelligence: The Need to Address Both Safety Risks and Fundamental Rights Risks. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives (Cambridge Law Handbooks, pp. 187-209). Cambridge: Cambridge University Press. doi:10.1017/9781009207898.016; Hacker, Philipp, The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future (November 25, 2022). Available at http://dx.doi.org/10.2139/ssrn.4279796

[15] Art. 82(1) establishes that "Any person who has suffered material or non-material damage as a result of an infringement of this Regulation shall have the right to receive compensation from the controller or processor for the damage suffered." Regulation (EU) 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/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1–88.

[16] The European Commission's initial proposal of the AI Act as well as the Council mandate both include in Recital 4 the wording: "Such harm might be material or immaterial."

[17] 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, 2021/0106(COD), version for Trilogue on 24 October, 2023.