Yale Journal of Law & Technology
Volume 26, Issue 3

Special Issue:
Yale Information Society Project
Digital Public Sphere Series

Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond

Sandra Wachter[*]

Predictive and generative artificial intelligence (AI) have both become integral parts of our lives through their use in making highly impactful decisions. AI systems are already deployed widely—for example, in employment, healthcare, insurance, finance, education, public administration, and criminal justice. Yet severe ethical issues, such as bias and discrimination, privacy invasiveness, opaqueness, and environmental costs of these systems, are well known. Generative AI (GAI) creates hallucinations and inaccurate or harmful information, which can lead to misinformation, disinformation, and the erosion of scientific knowledge. The Artificial Intelligence Act (AIA), Product Liability Directive, and the Artificial Intelligence Liability Directive reflect Europe's attempt to curb some of these issues. With the legal reach of these policies

[*] Oxford Internet Institute, University of Oxford, 1 St. Giles, OX1 3JS. Correspondence: sandra.wachter@oii.ox.ac.uk.

This work has been supported through research funding provided by the Wellcome Trust (grant No. 223765/Z/21/Z), Sloan Foundation (grant No. G-2021-16779), the Department of Health and Social Care, and Luminate Group to support the Trustworthiness Auditing for AI project and Governance of Emerging Technologies research program at the Oxford Internet Institute, University of Oxford. The funders had no role in the decision to publish or the preparation of this Essay.

I would like to express my gratitude to Professor Brent Mittelstadt, Dr. Daria Onitiu, Professor Philipp Hacker, and Chaitanya Rawat for their thoughtful and extensive comments on this Essay. I am also indebted to Elisabeth Paar, Gilad Abiri, and the *Yale Journal of Law & Technology* editorial team for their insightful and detailed feedback and their support during the production process. Their insights have immensely improved the quality of the Essay.

going far beyond Europe, their impact on the United States and the rest of the world cannot be overstated. In this Essay, I show how the strong lobbying efforts of big tech companies and member states were unfortunately able to water down much of the AIA. An overreliance on self-regulation, self-certification, weak oversight and investigatory mechanisms, and far-reaching exceptions for both the public and private sectors are the product of this lobbying. Next, I reveal the similar enforcement limitations of the liability frameworks, which focus on material harm while ignoring harm that is immaterial, monetary, and societal, such as bias, hallucinations, and financial losses due to faulty AI products. Lastly, I explore how these loopholes can be closed to create a framework that effectively guards against novel risks caused by AI in the European Union, the United States, and beyond.

Essay Contents

Introduction

Predictive and generative artificial intelligence (GAI) have both become integral parts of our lives through their use in making highly impactful decisions. Predictive AI (PredAI) systems are already deployed widely—for example in employment, healthcare, insurance, finance, education, public administration, and criminal justice. They are used to give people loans, admit them to universities, send them to prison, or hire and fire them.

GAI refers to artificial intelligence (AI) systems used to create or generate media, such as text, images, sound, or video.[1] GAI systems aided by large language models (LLMs), such as OpenAI's ChatGPT or Google DeepMind's Gemini, produce text and answer questions. Diffusion models, such as Open AI's DALL·E, Midjourney, and Stability AI's Stable Diffusion, can generate images and videos in response to user prompts. And generative adversarial networks can create images, voice profiles, and videos.

PredAI and GAI introduced both shared and distinct social, ethical, and legal risks. Ethical and legal challenges concerning discrimination and bias, explainability, misinformation, free speech, and data protection have been widely explored in relation to PredAI.[2] GAI exhibits similar issues to traditional AI relating to environmental impact, data protection, impact on employment and workplace automation, cybersecurity, bias, and discrimination.[3] But GAI also presents new questions relating to inaccurate and offensive content, misinformation,

---

[1] Of course, some GAI can also be predictive. The point is to distinguish between "traditional" AI (e.g., classification systems) and GAI (e.g., LLMs).

[2] *See, e.g.*, Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOC'Y, Dec. 1, 2016, at 1, 4-12; Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV. 1149, 1154 (2018); John Bowers & Jonathan Zittrain, *Answering Impossible Questions: Content Governance in an Age of Disinformation*, HARV. KENNEDY SCH. MISINFORMATION REV. (Jan. 14, 2020), https://misinforeview.hks.harvard.edu/article/content-governance-in-an-age-of-disinformation/ [https://perma.cc/MD43-BTZG].

[3] Laura Weidinger et al., *Taxonomy of Risks Posed by Language Models*, FACCT '22: PROC. 2022 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 214, 216-21 (2022).

harmful information, hallucinations, and violation of intellectual property rights.[4]

Following an intensive three-year period of negotiations, the forthcoming European Union (EU) AI Act (AIA) is set to govern PredAI and GAI models and systems.[5] The AIA is not, however, the only regulatory instrument that is set to govern AI. Rather, it will be complemented by harmonized technical standards and will stand next to two liability frameworks currently under negotiation: the updated Product Liability Directive (PLD) and the Artificial Intelligence Liability Directive (AILD).[6]

---

[4] *Id.*

[5] For an overview of the history of the AI Act and critiques of previous drafts, see generally Hannah Ruschemeier, *AI as a Challenge for Legal Regulation – The Scope of Application of the Artificial Intelligence Act Proposal*, 23 ERA F. 361 (2022); Martin Ebers et al., *The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)*, 4 MULTIDISCIPLINARY SCI. J 589 (2021); Michael Veale & Frederik Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach*, 22 COMPUT. L. REV. INT'L 97 (2021); Urs Gasser, *An EU Landmark for AI Governance*, 380 SCIENCE 1203 (2023); Mattis Jacobs & Judith Simon, *Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission*, 1 DIGIT. SOC'Y, no. 6, July 30, 2022, at 1; Meeri Haataja & Joanna J. Bryson, *The European Parliament's AI Regulation: Should We Call It Progress?*, 4 AMICUS CURIAE 707 (2023); Christiane Wendehorst, *The Proposal for an Artificial Intelligence Act COM (2021) 206 from a Consumer Policy Perspective*, FED. MINISTRY OF SOC. AFFS., HEALTH, CARE & CONSUMER PROT., REPUBLIC OF AUSTRIA 22-173 (2021), https://www.sozialministerium.at/dam/sozialministeriumat/Anlagen/Themen/Konsumentenschutz/Konsumentenpolitik/The-Proposal-for-an-Artificial-Intelligence-Act-COM2021-206-from-a-Consumer-Policy-Perspective_dec2021__pdfUA.pdf [https://perma.cc/HKC3-A9WZ].

[6] *European Parliament Legislative Resolution of 12 March 2024 on the Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products*, P9_TA (2024) 0132 (Mar. 12, 2024) [hereinafter *PLD*]; *see also Defective Products: Revamped Rules to Better Protect Consumers from Damage*, EUR. PARLIAMENT (Mar. 12, 2024), https://www.europarl.europa.eu/news/en/press-room/20240308IPR18990/defective-products-revamped-rules-to-better-protect-consumers-from-damages [https://perma.cc/2KTR-KUC6] (summarizing changes adopted in the revised text); *Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual*

Although these are EU laws, their effects will go beyond the geographic boundaries of the European Union. 448.4 million people, spanning twenty-seven countries, live in the European Union, making the EU one of the world's biggest markets.[7] If other countries do not want to lose access to this market, they will have to comply with these rules. Thus, the AIA will apply to companies in third-party countries, including companies that operate in the United States, that wish to place AI products on the market or that produce AI systems whose outputs are used in the EU.[8] Furthermore, the so called "Brussels Effect" will make it very likely that the EU frameworks will act as a blueprint for other regulations around the world, as was the case with the General Data Protection Regulation (GDPR), which has now risen to a global standard.[9] What is more, the EU issued one of the first comprehensive and legally enforceable frameworks worldwide. From a business perspective, and in the interest of streamlining, it will make sense for businesses to adapt their operations to comply with the strictest laws rather than to have fragmented standards across operations. The global effect of the AIA and the liability directives cannot be overstated.

In this Essay, I will explore how, despite very laudable efforts by European lawmakers, most of the aforementioned concerns about AI are not sufficiently addressed in the AIA and current AI liability directives. The Essay proceeds in four

---

Civil Liability Rules to Artificial Intelligence (AI Liability Directive), COM (2022) 496 final (Sept. 28, 2022) [hereinafter *AILD*].

[7] *Population and Population Change Statistics*, EUROSTAT (July 6, 2023), https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_and_population_change_statistics [https://perma.cc/MK2H-R5NU]; ANU BRADFORD, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD 25-66 (2020).

[8] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EcEC No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, 2024 O.J. (L) 1 [hereinafter *AI Act*].

[9] *See generally* BRADFORD, *supra* note 7; Charlotte Siegmann & Markus Anderljung, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, CTR. FOR GOVERNANCE A.I. (Aug. 2022), https://cdn.governance.ai/Brussels_Effect_GovAI.pdf [https://perma.cc/7UAQ-42VY].

parts. I begin by assessing the weaknesses and loopholes in the AIA as it applies to "narrow" PredAI in Part I and "general-purpose" GAI models and systems in Part II. Recognizing the many flaws and ambiguities of the AIA—and the general lack of practical, clear requirements for AI providers and developers—I turn my attention to the PLD and AILD in Part III. There, I reveal similar limitations at the foundations of these frameworks, which focus predominantly on material harms and monetary damages while ignoring the immaterial, financial, and societal harms of AI. I end in Part IV with recommendations on how to improve the status quo and make these legal instruments truly effective at governing AI and its full array of harms and benefits.

I. Predictive AI and the EU AI Act

The scope of the AIA is defined in Article 2,[10] according to which the law applies to two types of entities consisting of natural or legal persons, public authorities, agencies, or other bodies: (1) providers, meaning entities that place an AI system or service on the market, and (2) deployers, meaning entities that use an AI system under their authority, except where the AI system is used in the course of a personal, non-professional activity.[11] The AIA applies to both private and public providers and deployers inside the European Union, as well as to those in other countries that place an AI system on the EU market or that use a system impacting parties in the EU.[12]

A. Risk-Based Approach

For PredAI, the AIA follows a "risk-based" approach, under which obligations for providers and deployers are assigned according to the risk of their AI systems or services. Obligations are set out for systems and services carrying unacceptable risk,[13] high risk,[14] minimal, and no risk,[15] and there

[10] *See AI Act*, *supra* note 8, art. 2.
[11] Article 2 also has rules for importers and distributors, product manufacturers, users, and authorized representatives of providers. *See id.*
[12] *Id.* art. 2(1).
[13] *Id.* ch. II.
[14] *Id.* ch. III.
[15] *Id.* ch. X.

are also specific transparency obligations for certain AI systems, including general-purpose AI models (GPAI).[16]

1. Unacceptable Risks

Systems with unacceptable risks are "prohibited,"[17] although somewhat misleadingly, this does not constitute a full "ban" of such systems. Instead, the AIA features several wide-reaching and alarming exemptions, which fueled heated political debate during trilogue negotiations.[18]

In fact, the sections on bans and GPAI were the main sticking points that almost prevented the final agreement between the EU member countries from reaching the finish line in late December 2023. Even though a political agreement had already been reached in early December 2023,[19] the provisions on biometric categorization for law enforcement caused renewed debates. France was leading the charge,[20] and it may have been motivated by its plans to host the Olympics in 2024 and its desire to use AI-powered threat detection software for safety reasons.[21] Some civil society organizations worry such uses may lend themselves to mission creep towards biometric categorization.

Additional pressure to reach an agreement was fueled by the fact that the Spanish presidency of the Council of the

---

[16] *Id.* ch. IV.
[17] *Id.* art. 5.
[18] *See, e.g.*, Luca Bertuzzi, *AI Act: EU Parliament's Discussions Heat Up Over Facial Recognition, Scope*, EURACTIV (Oct. 7, 2022), https://www.euractiv.com/section/digital/news/ai-act-eu-parliaments-discussions-heat-up-over-facial-recognition-scope [https://perma.cc/FQR6-L4YG].
[19] *Commission Welcomes Political Agreement on AI Act*, EUR. COMMISSION (Dec. 9, 2023), https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473 [https://perma.cc/63L5-3A6H].
[20] Luca Bertuzzi, *AI Act: EU Policymakers Nail Down Rules on AI Models, Butt Heads on Law Enforcement*, EURACTIV (Dec. 12, 2023), https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-policymakers-nail-down-rules-on-ai-models-butt-heads-on-law-enforcement [https://perma.cc/6CMJ-FC2J].
[21] Juliette Jabkhiro & Julien Pretot, *Explainer: Olympics-How France Plans to Use AI to Keep Paris 2024 Safe*, REUTERS (Mar. 8, 2024), https://www.reuters.com/sports/olympics-how-france-plans-use-ai-keep-paris-2024-safe-2024-03-08 [https://perma.cc/75MR-GM3F].

European Union was due to end in January 2024.[22] The presidency of the Council is responsible for driving forward EU legislation. And the outgoing presidency wanted to strike a deal before leaving office, partly out of fear that the new Belgian presidency would shelve the whole legislative process and leave the Union with no agreement after three years of negotiation. These pressures led to this imperfect solution. After a thirty-six-hour negotiation marathon, a compromise was reached that left many lawmakers unsatisfied.[23]

According to Article 5 of the AIA,[24] systems with unacceptable risks are:

- "subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques" (Article 5(1)(a)).
- "an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation" (Article 5(1)(b)).
- "biometric categorisation systems that categorise individually natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation," except for uses in the area of law enforcement (Article 5(1)(g)).
- "social scor[ing]" AI systems used for "evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics" (Article 5(1)(c)).
- "'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law

---

[22] *The Presidency of the Council of the EU*, EUR. COUNCIL (2024), https://www.consilium.europa.eu/en/council-eu/presidency-council-eu [https://perma.cc/8X3L-QTBZ].

[23] Luca Bertuzzi, *European Union Squares the Circle on the World's First AI Rulebook*, EURACTIV (Feb. 15, 2024), https://www.euractiv.com/section/artificial-intelligence/news/european-union-squares-the-circle-on-the-worlds-first-ai-rulebook [https://perma.cc/DA5B-29PV].

[24] *AI Act*, *supra* note 8, art. 5.

enforcement," unless strictly necessary for certain objectives (Article 5(1)(h)).

- "risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics" (Article 5(1)(d)).[25]
- "AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage" (Article 5(1)(e)).
- "AI systems to infer emotions of a natural person in the areas of workplace and education institutions, except where the use of the AI system is intended to be put in place or into the market for medical or safety reasons" (Article 5(1)(f)).

The final list of prohibited systems leaves much to be desired. This would have been a good opportunity to ban, for instance, biometric categorization systems, "real-time" and ex-post remote biometric identification in public spaces, predictive policing, and emotion recognition in high-risk areas.

It is well established that remote biometric identification has abysmal accuracy rates (returning false matches some 80 percent of the time),[26] that predictive policing systems can generate racist and sexist outputs,[27] and that emotion

[25] According to Article 5(1)d, this "prohibition shall not apply to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity." *AI Act*, *supra* note 8, art. 5.

[26] Big Brother Watch Team, *Understanding Live Facial Recognition Statistics*, BIG BROTHER WATCH (May 22, 2023), https://bigbrotherwatch.org.uk/blog/understanding-live-facial-recognition-statistics [https://perma.cc/7KVE-CB2R]; *see also* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 1, 11 (2018).

[27] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 84-105 (2017); *COMPAS Recidivism Risk Score Data and Analysis*, PROPUBLICA DATA STORE (May 2024), https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis [https://perma.cc/TQ8W-AJ9S] (data last updated in 2017).

recognition software has little to no ability to objectively measure reactions.[28] Additionally, these systems have vast impacts on the rights to free speech, privacy, protest, and assembly.[29]

Further, a prohibition of EU-based institutions selling banned AI systems and services to third-party countries was not introduced, despite being discussed at negotiations.[30] This was a missed opportunity. Such a ban could have complemented the Commission's Proposal for a Regulation on Prohibiting Products Made with Forced Labour on the Union Market[31] and the EU Corporate Sustainability Due Diligence Directive (CSDDD).[32] These frameworks take a sensible step toward preventing companies from profiting off forced labor, and they introduce ethical business practices that protect the environment and do not violate human rights. A ban on the sale of prohibited AI to third-party countries would have fit well within these current regulatory proposals.

2. High Risks

Article 6 identifies two categories of high-risk applications: (1) AI systems that are "intended to be used as a safety component of a product, or the AI system is itself a product" (Annex I),[33] and that have to undergo a third-party conformity assessment (e.g., toys, medical devices, in vitro diagnostic medical devices, etc.),[34] and (2) eight other high-risk

---

[28] Lisa Feldman Barrett et al., *Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements*, 20 PSYCH. SCI. PUB. INTEREST 1, 47 (2019).

[29] Marissa Gerchick & Matt Cagle, *When It Comes to Facial Recognition, There Is No Such Thing as a Magic Number*, ACLU (Feb. 7, 2024), https://www.aclu.org/news/privacy-technology/when-it-comes-to-facial-recognition-there-is-no-such-thing-as-a-magic-number [https://perma.cc/EUG5-MLHA].

[30] Bertuzzi, *supra* note 23.

[31] *Proposal for a Regulation of the European Parliament and of the Council on Prohibiting Products Made with Forced Labour on the Union Market*, at 36-52, COM (2022) 453 final (Sept. 14, 2022).

[32] *Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and Amending Directive (EU) 2019/1937*, at 54-115, COM (2022) 71 final (Feb. 23, 2022).

[33] *AI Act*, *supra* note 8, art. 6(1)(a).

[34] *See id.* annex I.

applications listed in Annex III.[35] These high-risk applications are allowed but must follow certain ex-ante rules. They fall into the following areas:

1. biometrics including emotion recognition;
2. critical infrastructure;
3. education and vocational training;
4. employment, workers management and access to self-employment;
5. access to and enjoyment of essential private services and essential public services and benefits;[36]
6. law enforcement;[37]
7. migration, asylum, and border control management;[38] and
8. administration of justice and democratic processes.[39]

While a helpful start, this list omits other important areas, such as AI used in media, recommender systems, science and academia (e.g., experiments, drug discovery, research, hypothesis testing, parts of medicine), most of finance and trading, most types of insurance, and specific consumer-facing applications, such as chatbots and pricing algorithms, which pose significant risk to individuals and society.[40]

Before placing a high-risk system on the market, providers must follow requirements specified in Chapter III of the AIA. This includes duties such as establishing risk-assessment systems; ensuring data governance; keeping technical documentation and records; and maintaining transparency,

---

[35] *Id.* annex III.

[36] This includes AI used for risk assessment and pricing for life and health insurance for natural persons. *See id.* annex III(5)(c).

[37] Insofar as their use is permitted under relevant EU or national law. *See id.* annex III(6).

[38] Insofar as their use is permitted under relevant EU or national law. *See id.* annex III(7).

[39] This includes AI systems intended to be used for influencing the outcome of an election or referendum or the voting behavior of natural persons. *See id.* annex III(8)(b).

[40] For a taxonomy of both observed and anticipated risks of large language models, see Weidinger et al., *supra* note 3. On risks to academia, see Brent Mittelstadt et al., *To Protect Science, We Must Use LLMs as Zero-Shot Translators*, 7 NAT. HUM. BEHAV. 1830, 1830-32 (2023); On financial risks, see Talia B. Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175, 1204-19 (2022).

human oversight, accuracy, cybersecurity, and robustness.[41] High-risk systems must also be registered in a public database,[42] are subject to post-market monitoring by providers and market surveillance authorities,[43] and (as discussed below) require fundamental human rights impact assessments for certain Annex III applications.[44]

In contrast, deployers have very limited obligations, which include human oversight, recordkeeping, and monitoring duties, and for some deployers, fundamental rights impact assessment and registration duties.[45]

3. Transparency Obligations for Specific AI Systems

Article 50 in Chapter IV specifies transparency obligations for specific AI systems. The AIA requires that users must be made aware that they are interacting with (e.g., for chatbots or emotion recognition)[46] or viewing outputs from (e.g., for GPAI or deepfakes) certain types of AI.[47] These obligations leave much to be desired given the well-established harms that such systems may cause. For example, in the past, chatbots have advised users to take their own lives,[48] given dieting tips to

---

[41] *See AI Act*, *supra* note 8, arts. 9-15.

[42] Unless the AI system is used for law enforcement and migration, in which case only the supervisory authority will have access. *See id.* art. 49(4).

[43] See *id.* arts. 72, 74-76.

[44] *See id.* art. 27 ("Once the impact assessment has been performed, the deployer shall notify the market surveillance authority of the results of the assessment . . . .").

[45] *Id.* arts. 26-27, 49(3).

[46] *Id.* art. 50(1), 50(3).

[47] *Id.* art. 50(2).

[48] Chloe Xiang, *"He Would Still Be Here": Man Dies by Suicide After Talking with AI Chatbot, Widow Says*, VICE (Mar. 30, 2023, 3:59 PM), https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says [https://perma.cc/4834-X8QC]. Similar concerns related to patient safety have been raised in relation to Virtual Mental Health Assistants (VMHA). *See* Surjodeep Sarkar et al., *A Review of the Explainability and Safety of Conversational Agents for Mental Health to Identify Avenues for Improvement*, 6 FRONT. A.I., Oct. 12, 2023, at 1, 4-5 ("People have been proven to be particularly forthcoming about their mental health problems while interacting with conversational agents, which may increase the danger of '*agreeing with those user utterances that imply self-harm*[.]'").

people battling eating disorders,[49] and produced reputation-damaging outputs (e.g., false sexual assault charges against innocent people).[50] Transparency alone is insufficient to address these issues.[51]

Finally, according to Article 95 in Chapter X, AI systems posing minimal or no risk are not required to adhere to *any* of the AIA's obligations or harmonized standards. Providers and deployers of such systems can, if they so choose, adhere to AIA requirements or voluntary codes of conducts.[52]

B. Pre-Market Risk Assessment for High-Risk AI

The list of high-risk systems in Annex III, and the procedure for classifying the risk level of systems, have been heavily debated during the three-year negotiation process. Prior drafts of the AIA specified that high-risk obligations apply by default as soon as an AI system or service is developed for one of the high-risk sectors.[53] The final version—pushed by the European Parliament and Council, though criticized by the Parliament's legal office[54]—complicates matters by introducing a complex pre-market risk assessment as a precondition for high-risk obligations to apply. In other words, providers of AI systems that would have been considered high-risk by default can

---

[49] Kate Wells, *An Eating Disorders Chatbot Offered Dieting Advice, Raising Fears About AI in Health*, NPR (June 9, 2023, 6:59 AM ET), https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea [https://perma.cc/7LRQ-S7NL].

[50] Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 555 (2023).

[51] Of course, other frameworks, such as those on speech regulation and other tort laws, could cover some of these issues, though most of these are ex-post mechanisms that enter after the damage is already done.

[52] *See AI Act*, *supra* note 8, art. 95.

[53] Title III and Annex III of the April 2021 draft provided the relevant language. *See 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*, at 45-68, COM (2021) 206 final (Apr. 21, 2021).

[54] Luca Bertuzzi, *AI Act: EU Parliament's Legal Office Gives Damning Opinion on High-Risk Classification 'Filters'*, EURACTIV (Oct. 24, 2023), https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-parliaments-legal-office-gives-damning-opinion-on-high-risk-classification-filters [https://perma.cc/XC3S-TNG2].

instead specify the system's intended use, conduct an internal assessment, and claim that no significant risk of harm to the health, safety, or fundamental rights of natural persons is expected. In this scenario, providers would be exempt from Chapter III obligations for high-risk AI systems.

Article 6(3) sets out criteria that exempt from high-risk designation systems used or intended to:

- "perform a narrow procedural task";
- "improve the result of a previously completed human activity";
- "detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review";
- "perform a preparatory task."[55]

However, systems that perform profiling of natural persons are always high-risk.[56]

This assessment-based loophole is unfortunate. The exemptions are very vague and far reaching. From a human-rights perspective, a clear-cut approach that assumes risk levels based on intended uses would have been preferable to one introducing a complicated internal assessment procedure that will allow providers to argue that their systems and services do not cause fundamental-rights concerns. At a minimum, the new provisions create a significant enforcement burden for the AI Board established under Article 65 and for the Commission that is responsible for monitoring assessment efficacy and changing its criteria over time.

This approach is even more problematic because providers can place systems onto the market as soon as they have conducted an assessment. They need not wait for approval to verify the risk level of their systems and services.[57] This is quite

---

[55] *AI Act*, *supra* note 8, art. 6(3).

[56] *Id.*

[57] On the history of these provision and the debates around it, see Luca Bertuzzi, *AI Act: Leading MEPs Revise High-Risk Classification, Ignoring Negative Legal Opinion*, EURACTIV (Oct. 25, 2023), https://www.euractiv.com/section/artificial-intelligence/news/ai-act-leading-meps-revise-high-risk-classification-ignoring-negative-legal-opinion [https://perma.cc/SA4W-TAXW].

uncommon for the EU, which is usually very comfortable having a centralized approval mechanism in place. The AIA's deviation from this norm might be the result of lobbying efforts, but that remains speculative at this point.

Additionally, the Act requires registration of AI systems in a public database,[58] but the risk assessment and its methodology are not publicly accessible.[59] Upon the request of the national competent authorities, providers must provide this documentation.[60] But again, this creates a significant additional enforcement burden for competent authorities that may have neither the time nor the resources to test provider claims and systematically audit systems.

C. Fundamental Rights Impact Assessment for High-Risk AI Systems

Article 27 also establishes a mandatory fundamental rights impact assessments for high-risk applications listed in Annex III.[61] First, it is unclear what added benefit this assessment will have because for a system to be classified as high-risk, a pre-market assessment on potential impacts on fundamental rights must already be conducted by the AI provider. That said, Article 27 applies to AI deployers rather than providers (who are covered by Article 6(2)), so the intended effect may be to ensure that multiple bodies conduct relevant assessments based on their unique knowledge of the system, service, or intended use case and affected populations.

Second, Article 27 clearly states that not every deployer must undertake a fundamental rights impact assessment; AI used in critical infrastructure is exempt.[62] Article 27 also only targets deployers that are bodies governed by public law, private operators providing public services, or private entities acting on behalf of public bodies. Private companies only need to conduct an impact assessment if they are deploying high-risk systems referred to in Annex III(5)(b), relating to

---

[58] *AI Act*, *supra* note 8, art. 71.
[59] *Id.* art. 6(4).
[60] *Id.*
[61] *See id.* art. 27.
[62] *Id.*

creditworthiness and credit scoring, and (5)(c), relating to life and health insurance.[63]

This means that a large range of systems are not covered, including AI used in employment, education, critical infrastructure, and possibly even the financial sector. While creditworthiness, credit scores, and life and health insurance are covered, the broader financial sector could potentially be exempt from some of the rules of the AIA.[64] Financial authorities in the member states can assess whether existing laws governing the financial sector already have sufficient safeguards in place to deal with similar risks. If financial authorities determine that existing laws sufficiently address and integrate the AI Act's obligations, private companies would not be required to follow additional AIA rules.[65]

Finally, an overarching doctrinal issue is that human and fundamental rights do not bind the private sector.[66] Rather, human rights are created to limit the powers of state actors. It is therefore questionable how useful fundamental rights assessments will be to mitigate the risks of high-risk AI outside of the public sector.

D. Duties Under the AIA

Chapters II and III of the AIA define the responsibilities of providers and deployers. Given the intended purpose of the AIA as a horizontal regulatory framework, one may expect a wide range of clearly defined duties that explain how AI should be developed and deployed. Unfortunately, Articles 9 through 27 within Chapter III rarely provide clear expectations and requirements.

Article 10 serves as an example of this ambiguity and its potentially harmful effects. Article 10 is the main provision dealing with data and data governance, including issues relating

---

[63] *See also* Philipp Hacker, Comments on the Final Trilogue Version of the AI Act 9, 11 (Jan. 23, 2024) (unpublished manuscript) (available at: https://ssrn.com/abstract=4757603 [https://perma.cc/5ATM-GXYJ]).

[64] *See AI Act*, *supra* note 8, recital 158.

[65] *See id.*

[66] Eleanor Spaventa, *Fundamental Rights in the European Union*, *in* EUROPEAN UNION LAW 243, 247, 249 (Catherine Barnard & Steve Peers eds., 3d ed. 2020). For this and for the very few exceptions, see ROBERT SCHÜTZE, EUROPEAN UNION LAW 488-91 (3d ed. 2021).

to bias and discrimination.[67] Article 10(2) explains that training, validation, and testing data need to be subject to governance practices, including documentation of the relevant design choices, annotation and enrichment, formulation of information about assumptions and representativeness, examination of bias, and identification of shortcomings and measures to detect, prevent, and mitigate possible biases.[68] Article 10(3) requires training, validation, and testing datasets to be relevant, representative, and, to the best extent possible, free of errors and complete.[69] While these are reasonable standards, the AIA does not give any indication of how they should be achieved in practice.

The AIA does not define bias or discuss how to measure it. It also does not discuss acceptable levels of bias, mitigation strategies, expected behavior if bias cannot be detected or mitigated, or how to prevent biases. It lacks examples of positive or desirable bias (e.g., positive/affirmative action), bias related to ground truth (e.g., in relation to health), or how bias differs culturally yet often has a Western-oriented view.[70] Yet researchers in recent years have developed a wide range of technical mechanisms that can be useful for detecting and mitigating biases.[71]

Clear guidance on this issue is pivotal. As I have argued elsewhere, the choice of fairness metrics to measure bias is not

---

[67] For a more in-depth discussion of Article 10 of the AIA, see Marvin van Bekkum & Frederik Zuiderveen Borgesius, *Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the GDPR Need a New Exception?*, 48 COMPUT. L. & SEC. REV., Apr. 2023, at 1, 8-12, and Philipp Hacker, *A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act*, 13 L. INNOVATION & TECH. 257, 258-59 (2021).

[68] *See AI Act*, *supra* note 8, art. 10(2).

[69] *Id.* art. 10(3).

[70] Nithya Sambasivan et al., *Re-Imagining Algorithmic Fairness in India and Beyond*, FAccT '21: PROC. 2021 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 315, 316-17 (2021); *see also* Sandra Wachter et al., *Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI*, 41 COMPUT. L. & SEC. REV., July 2021, at 1, 19-24;

[71] Sandra Wachter et al., *Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law*, 123 W. VA. L. REV. 735, 744 (2021); Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, FAIRWARE '18: PROC. INT'L WORKSHOP SOFTWARE FAIRNESS 1, 1-2 (2018).

a neutral decision. Most bias tests (thirteen out of twenty of the most popular metrics) do not live up to the standards of European and U.K. non-discrimination law.[72] One reason is because the vast majority of bias tests and standards have been developed in the United States under fundamentally different anti-discrimination laws.

Similarly, I have revealed the harmful impact of enforcing many "group fairness" measures in practice.[73] Specifically, fairness is achieved by "leveling down," or making all groups impacted by a system worse off, rather than helping disadvantaged groups.[74] This approach is problematic in the context of EU and U.K. non-discrimination law, as well as ethically troubling. Leveling down can be extremely harmful in practice. In healthcare, for example, enforcing group fairness could mean missing more cases of cancer than strictly necessary while also making a system less accurate overall.

As these examples suggest, well-intentioned public and private entities in the European Union, the United States, or anywhere in the world following the state-of-the-art in measuring and mitigating bias can unknowingly take actions that clash with EU law. It is therefore incredibly important to explain to the public and private sectors whether the bias tests are legally compliant depending on the sector, application, and jurisdiction of use.

A better but still ambiguous approach would have been to phrase Article 10 in a way that assumed datasets and models are biased or discriminatory unless proven otherwise. Neutral data is a fantasy. Historical bias is deeply embedded in sectors such as employment, finance, education, criminal justice, and healthcare, and the data collected to train and test AI reflect this bias.[75] To that end, I have argued in previous work for a reversal of the burden of proof under which providers would be tasked with demonstrating that their systems are in fact unbiased.[76]

---

[72] Wachter et al., *supra* note 71, at 761-67.

[73] Brent Mittelstadt et al., *The Unfairness of Fair Machine Learning: Levelling Down and Strict Egalitarianism by Default*, ARXIV 25-34 (Mar. 12, 2023), https://arxiv.org/pdf/2302.02404 [https://perma.cc/4CFJ-9L8X].

[74] *Id.* at 6-9, 15-18.

[75] Wachter et al., *supra* note 71, at 767-74.

[76] *Id.* at 762, 775, 780.

E. Harmonized Standards

Of course, the AIA is not the only relevant governance mechanism. So-called "harmonized standards" requested by the European Commission are also highly relevant and intended to fill in these types of gaps in Articles 9-27.[77] It is a common approach in European lawmaking to have harmonized standards created by the European Committee for Standardization (CEN) and European Committee for Electrotechnical Standardization (CENELEC), with the aid of working groups composed of relevant stakeholders. The harmonized standards accompanying the AIA will outline more detailed requirements and serve a quasi-legal function, as discussed below.

In this context, standards bodies have an important policy role to play. It is essential to note, therefore, that as private bodies, CEN and CENELEC do not have direct democratic legitimacy, something they have been highly criticized for in the past.[78]

---

[77] For the standardization request issued by the Commission to CEN/CENELEC in December 2022, see *Draft Standardisation Request to the European Standardisation Organisations in Support of Safe and Trustworthy Artificial Intelligence*, EUR. COMM'N, annex I (2022), https://ec.europa.eu/docsroom/documents/52376/attachments/1/translations/en/renditions/native [https://perma.cc/EC6P-VS79].

[78] Martin Ebers, *Standardizing AI: The Case of the European Commission's Proposal for an 'Artificial Intelligence Act'*, *in* THE CAMBRIDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE: GLOBAL PERSPECTIVES ON LAW AND ETHICS 321, 340-41 (Larry A. DiMatteo et al. eds., 2022); Veale & Borgesius, *supra* note 5, at 105; Johann Laux, Sandra Wachter & Brent Mittelstadt, *Three Pathways for Standardisation and Ethical Disclosure by Default Under the European Union Artificial Intelligence Act*, 53 COMPUT. L. & SEC. REV., July 2024, at 1, 8-9. Improvements to multi-stakeholder representation in the standardization process may improve in the future. Article 40(3) calls on standards setting bodies to "enhance multi-stakeholder governance ensuring a balanced representation of interests and the effective participation of all relevant stakeholders." *AI Act*, *supra* note 8, art. 40(3). The newly established AI Office is likewise called upon explicitly by the Commission to "engage with the scientific community, industry, civil society and other experts," *id.* recital 111, in fulfilling their duties, which include assisting the Commission in preparing standardization requests and evaluating standards, *id.* arts. 56, 66. For details of the forthcoming AI Office's remit, see Commission Decision of 24 January 2024 Establishing the European Artificial Intelligence Office, 2024 O.J. (C 1459) 1.

In the context of AI governance, this lack of democratic legitimacy is even more worrying due to the far-reaching legal, ethical, political, and economic consequences of the widespread deployment of AI. Standards bodies will be tasked with creating frameworks that interpret the AIA. Most stakeholders in the relevant working groups are industry representatives.[79] Civil society is heavily excluded, and those who are part of the working groups only have so-called "observer status," which means they lack voting rights and are not always allowed to speak or submit opinions that must be considered by the relevant working group.[80] Having a more diverse group of stakeholders developing standards would be preferable to ensure effective mitigation of AI's risks for individuals and society.[81]

Apart from the homogeneity of the working groups, the envisioned scope and remit of the AIA's harmonized standards are also worrying. As I and others have elaborated elsewhere, standards bodies are typically asked to set technical standards which involve minimal normative or ethical content.[82] Standards bodies tend not to have the expertise or democratic legitimacy to set explicitly normative or ethical standards, and yet they have been put in this position by the absence of such details in the AIA.

In short, it is likely that normative issues concerning bias or other areas under Articles 9 through 27 will not be addressed satisfactorily in the AIA's harmonized standards. For example, clear definitions of fairness, acceptable levels of disparity, as well as approaches to detect and mitigate bias are unlikely to

---

[79] *See* Ebers, *supra* note 78, at 340-41.

[80] *Id.* at 343-44.

[81] Similar calls have been made by the U.N. Office of the High Commissioner for Human Rights. *See Call for Inputs: "The Relationship Between Human Rights and Technical Standard-Setting Processes for New and Emerging Digital Technologies (2023)" - Report of the High Commissioner for Human Rights*, UNITED NATIONS: OFF. HIGH COMM'R HUM. RTS. (June 30, 2023), https://www.ohchr.org/en/calls-for-input/2023/call-inputs-relationship-between-human-rights-and-technical-standard-setting [https://perma.cc/NBL3-3ZP8]; *see also* BEUC News, *New Study on the Role of Standards in EU Digital Policy Legislation*, BEUC (July 14, 2023), https://www.beuc.eu/news/new-study-role-standards-eu-digital-policy-legislation [https://perma.cc/A9MJ-RKP4].

[82] Laux et al., *supra* note 78, at 3.

be addressed in a way that gives providers or deployers clear, practical guidance or requirements. Explainability and transparency are similarly unlikely to be sufficiently addressed, as setting normative standards would require answering questions such as, "What is a good explanation?" or, "What criteria should be disclosed?"

F. Conformity Assessment

Conformity assessments are similarly a weak point in the AIA. To certify compliance with the AIA and its associated harmonized standards, conformity assessments must be undertaken. Confirmation of compliance with the harmonized standards creates a presumption of compliance with the AIA, in effect giving the harmonized standards quasi-legal effect,[83] despite their lack of democratic legitimacy.

For AI systems that are intended to be used as a safety component of a product, or where the AI system is itself a product, conformity assessments will be done by a third party (e.g., for medical devices under the Medical Devices Regulation, toys, or lifts).[84] However, for high-risk AI listed in Annex III (e.g., education, workplace, financial services), conformity assessments will be undertaken by the providers themselves and are not made public.[85] After assessing conformity, the provider can then put a CE mark on the product or service, signaling compliance with European law. Biometric systems are the only exception to the rule where a third-party assessment can be undertaken by the notified bodies, even though this is no longer mandatory in the AIA's latest text.[86]

This enforcement method is concerning. Providers are not only heavily involved in writing the harmonized standards to which they must adhere but also tasked with assessing whether they comply with those standards. This approach creates a major legal loophole in which those who are supposed to be regulated can testify to compliance with rules they have written for themselves.

---

[83] *See AI Act*, *supra* note 8, art. 40(1).
[84] *Id.* art. 43(3).
[85] *Id.* art. 43(2).
[86] *Id.* art. 43(1).

Recital 125 acknowledges this issue by explaining that, in an ideal scenario, a third party would undertake conformity assessments. However, recitals are not legally binding,[87] and the legally binding text of the AIA does not create a pathway to mandate third-party assessments in the future. This topic did not resurface during negotiations over the AIA, meaning it is highly unlikely that mandatory third-party conformity assessments will be required in the near future.

G. Individual-Level Rights

While the AIA contains many problematic ambiguities, enforcement gaps, and loopholes, the final text also features some positive developments. During negotiations, the European Parliament's suggestion to create individual-level rights was adopted. This approach breaks with the predominant regulatory approach of the AIA, which was originally formulated as a product-safety framework. Individual rights are now guaranteed in Articles 85, 86 and 99(10).

Article 85 enables individuals, or groups of individuals, to launch a complaint with a market surveillance authority if an AI system relating to them infringes the regulation.[88] Article 99(10) grants effective judicial remedies and due process against the actions of a market surveillance authority.[89] And under Article 86, individuals also have the right to receive an explanation about the output of a high-risk AI system that produced legal or similarly significant effects to the health, safety, socio-economic, or any other fundamental rights.[90]

None of these individual rights were conceived of in the original text.[91] In its original form, the AIA would not have given individuals any complaint or recourse mechanisms

---

[87] Tadas Klimas & Jūratė Vaičiukaitė, *The Law of Recitals in European Community Legislation*, 15 ILSA J. INT'L & COMPAR. L. 61, 83-86 (2008).
[88] *AI Act*, *supra* note 8, art. 85.
[89] *Id.* art. 99(10).
[90] *Id.* art. 86.
[91] This was criticized by the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS). *See Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, EUR. DATA PROT. BD. 9 (June 18, 2021), https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf [https://perma.cc/Y9BD-8CUJ].

against unlawful behavior. The addition of these rights is a great step toward effective individual recourse mechanisms for AI and toward finally creating a legally binding right to explanation.[92]

II. GAI in the AIA

The original draft of the AIA from April 2021[93] did not include special provisions for GAI. The Council's draft from November 2022[94] and the European Parliament's draft from May 2023[95] introduced provisions addressing both general-purpose AI models (e.g., GPT-4 and Gemini) as well as general-purpose AI systems (e.g., ChatGPT, DALL·E).[96] Article 2(1) of the current version explains that the framework applies to providers of GPAI, while deployers face various provisions under Chapters IV and VIII.[97]

Regulation of GAI was another big sticking point during the negotiations that almost caused the AIA not to pass. Even though political agreement was reached in October 2023, Germany, Italy, and France started a coordinated effort in November and December 2023 to remove most provisions on

---

[92] *See* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 97 (2017).

[93] *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*, *supra* note 53.

[94] *Interinstitutional File: 2021/0106*, COUNCIL OF THE EUR. UNION (2022), https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf [https://perma.cc/Q67E-W7FB].

[95] *Draft 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*, EUR. PARLIAMENT (2023), https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf [https://perma.cc/CD6V-WBK9].

[96] For a historical overview and critique of previous drafts, see Philipp Hacker et al., *Regulating ChatGPT and Other Large Generative AI Models*, FACCT '23: PROC. 2023 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 1112, 1114-15 (2023).

[97] *See AI Act*, *supra* note 8, art. 2(1); *id.* chs. IV, VIII.

GAI.[98] The three nations even threatened to vote against the whole Act if these provisions were left unchanged. Their opposition was heavily motivated by the lobbying efforts of Mistral AI and Aleph Alpha, the French and German AI companies that were widely hoped to rival OpenAI and Microsoft and to offer an EU alternative for GAI services and products.[99] The nations justified their pushback against regulation of GAI by explaining that the AIA could stifle these newcomers, would further increase the EU's dependency on the United States, and would fortify U.S. market power. This rhetoric was successful and led to a weakening of the regulation.[100]

All of this makes Mistral AI's announcement to enter in a partnership with Microsoft in February 2024—days after the final agreement on the AIA—even more disappointing.[101] The partnership diminishes the hopes of having a strong EU alternative for GAI products, while still leaving the EU with weaker laws regulating GAI for EU citizens. The following sections will assess the weaker framework that is the result of these lobbying efforts.

---

[98] Andreas Rinke, *Exclusive: Germany, France and Italy Reach Agreement on Future AI Regulation*, REUTERS (Nov. 20, 2023, 4:18 PM EST), https://www.reuters.com/technology/germany-france-italy-reach-agreement-future-ai-regulation-2023-11-18 [https://perma.cc/R4QY-U4MN].

[99] Luca Bertuzzi, *EU's AI Act Negotiations Hit the Brakes over Foundation Models*, EURACTIV (Nov. 15, 2023), https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models [https://perma.cc/BH9P-5BW5].

[100] Kelvin Chan, *Europe's World-Leading Artificial Intelligence Rules Are Facing a Do-or-Die Moment*, QUARTZ (Dec. 4, 2023), https://qz.com/europes-world-leading-artificial-intelligence-rules-are-1851069721 [https://perma.cc/ZE6L-HHW4]. The resulting weaknesses in the Act include its focus on transparency obligations rather than on liability, preference for codes of conducts over hard regulations, and tiered approach for systematic risks. *See infra* Sections II.A-D.

[101] *See* Madhumita Murgia, *Microsoft Strikes Deal with Mistral in Push Beyond OpenAI*, FIN. TIMES (Feb. 26, 2024), https://www.ft.com/content/cd6eb51a-3276-450f-87fd-97e8410db9eb [https://perma.cc/6DAH-LVQ2].

A. GPAI Models: A Tiered Approach

Articles 51 and 52 distinguish between general-purpose AI models and general-purpose AI models with systemic risks.[102] A model has systemic risks when the cumulative amount of computation used for its training measured in floating point operations (FLOPS) is greater than 10^{25},[103] unless the provider can demonstrate that their system does not pose a systemic risk.[104] Recital 110 gives examples of a wide range of systemic risks, including "major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; [and] the dissemination of illegal, false, or discriminatory content."[105]

Providers of all GPAI models must draw up technical documentation, including the model's training and testing process and the results of its evaluation.[106] Providers must also disclose certain information to downstream providers, respect copyright law, provide a sufficiently detailed summary of the training data, and report on known or estimated energy consumption.[107] Providers of GPAI models with systemic risks, meaning those using more than 10^{25} FLOPS, are required to perform model evaluations—including adversarial testing (e.g., red teaming) that does not need to be external—assess and mitigate possible systemic risks, document and report

---

[102] *AI Act*, *supra* note 8, arts. 51-52. These distinctions are also explored in Annexes XI and XIII.
[103] *Id.* art. 51(2).
[104] *Id.* art. 52(2).
[105] *Id.* recital 110. Recital 110 also lists "harmful bias and discrimination with risks to individuals, communities or societies; the facilitation of disinformation or harming privacy with threats to democratic values and human rights; [and] risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community" as systemic risks. *Id.*
[106] This information is not public but can be requested by the AI Office and the national competent authorities. *See id.* art. 53.
[107] *Id.* art. 53(1); *id.* annex XI. Open-source providers are exempt from these obligations unless they provide GPAI models with systemic risks. *Id.* art. 53(2).

serious incidents and possible corrective measures, and ensure an adequate level of cybersecurity.[108]

As with PredAI, harmonized standards will also be developed for both types of GPAI. Self-assessed conformity will once again create the presumption of compliance with the AIA.[109] The same concerns with this type of self-assessed regulatory enforcement as discussed *supra* Section I.B apply here.

This regulatory approach is disappointing. The two-tier model is quite unconvincing because many "systemic risks" occur in all GPAI models, regardless of their size or computation. Misinformation, hallucinations, bias, work displacements, data protection issues, explainability problems, and harmful outcomes occur in smaller and less "capable" systems.[110]

The 10^{25} FLOPS threshold is likely at the time of writing to cover only OpenAI's GPT-4, Google DeepMind's Gemini, and Meta's Llama 3.1.[111] This would exclude models such as GPT-

---

[108] *Id.* art. 55(1). For further discussion on the AI Act and GAI, see Claudio Novelli et al., *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, ARXIV (Mar. 15, 2024), https://arxiv.org/pdf/2401.07348 [https://perma.cc/BT4A-8CYY].

[109] *See AI Act, supra* note 8, art. 40(1).

[110] For example, Aleph Alpha, which does not meet the FLOPS threshold, produced racist and sexist outputs. Jakob von Lindern, *Aleph Alpha: Braucht die Deutsche Vorzeige-KI Mehr Erziehung?*, DIE ZEIT (Sept. 11, 2023), https://www.zeit.de/digital/2023-09/aleph-alpha-luminous-jonas-andrulis-generative-ki-rassismus [https://perma.cc/Q57D-7SJY]; Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜, FACCT '21: PROC. 2021 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 610, 612-18 (2021) (warning about risks like bias, misinformation, environmental impact, and hallucinations before Gemini and GPT-4 were on the market); Weidinger et al., *supra* note 3, at 216-22 (same).

[111] *Artificial Intelligence – Questions and Answers*, EUR. COMM'N (Dec. 14, 2023), https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683 [https://perma.cc/P9MK-YGZT]; Natali Helberger & Nicholas Diakopoulos, *ChatGPT and the AI Act*, 12 INTERNET POL'Y REV., Feb. 16, 2023, at 1, 3, 6 (convincingly calling for the need to make generative AI its own risk category).

3.5,[112] the version that is freely available to the public. Also excluded are models such as Anthropic's Claude, Aleph Alpha's Luminous, Mistral AI's Mistral Nemo, Meta's Llama 3, Midjourney, Stability AI's Stable Diffusion 3 Medium, and Falcon LLM.[113]

Using FLOPS as a proxy for danger or capability is unworkable. The FLOPS threshold would be more sensible for assessing GPAI's environmental impact in training and production, although energy consumption of less capable models is similarly important. Even here, recent work to optimize models undermines the threshold's relevance. If current trends hold, it will be entirely feasible for much smaller models falling below the FLOPS threshold to rival GPT-4 and other GPAI with systemic risks.[114] At the same time, multimodal models producing sound, images, and video are likely to consume huge amounts of computing power without self-evidently being more or less dangerous than their text-based counterparts. The FLOPS threshold also incentivizes providers to optimize models to fall under the threshold without necessarily making those models less dangerous.

An important enforcement question remains unanswered: What happens if a provider creates a large model and distils it down to a smaller one? Does this also count as a model with systemic risks, or is the FLOPS threshold of the original training what counts?

B. Transparency Overflow

AIA provisions applicable to GAI predominantly focus on providers of models, rather than providers or deployers of

---

[112] Zosia Wanat et al., *EU to Put Extra Guardrails on AI Foundation Models Like GPT-4*, SIFTED (Dec. 7, 2023), https://sifted.eu/articles/foundation-model-eu-ai-act [https://perma.cc/NW5V-9U4F].

[113] *See* Robi Rahman et al., *Tracking Large-Scale AI Models*, EPOCH AI (Apr. 5, 2024), https://epochai.org/blog/tracking-large-scale-ai-models [https://perma.cc/CE37-ZEPR]; *Large-Scale AI Models*, EPOCH AI (Aug. 10, 2024), https://epochai.org/data/large-scale-ai-models [https://perma.cc/8M8Z-44LF].

[114] *See* Kyle Wiggers et al., *Anthropic's $5B, 4-Year Plan to Take on OpenAI*, TECHCRUNCH (Apr. 6, 2023, 5:25 PM EDT), https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai [https://perma.cc/W9R3-FM8X].

GPAI systems.[115] Very limited obligations apply to providers and deployers of GPAI systems, such as ChatGPT, DALL·E, or Midjourney. Providers must make people aware that they are interacting with an AI system and watermark the output of their systems.[116] Deployers of AI systems must make users aware that the content is a deepfake and that the output is artificially generated or manipulated.[117] This light-touch approach is problematic because it is not self-evident that GPAI systems pose fewer or less severe risks than GPAI models.

Governance of GPAI providers overwhelmingly and problematically relies on transparency mechanisms. While it is essential that providers of GPAI models and systems make certain information and documentation available, this is only the first step in adequate governance.

The main weakness of the current approach is the lack of normative thresholds. While it is good to require documentation about performance, this will only be useful if thresholds or standards for good or bad performance exist to which providers of GPAI models and systems must adhere.[118] The same holds true for deployers of GPAI, who currently have almost no obligations.[119] If normative thresholds were to exist, the regulatory burden would shift from individuals and supervisory authorities to providers and deployers to create and deploy only GPAI models and systems that provably meet relevant risk-based normative standards.

---

[115] "GPAI model" refers to foundation models such as GPT-4 or Gemini. GPAI systems are applications, such as ChatGPT, that are built on top of these foundation models.

[116] *AI Act*, *supra* note 8, art. 50(1)-(2).

[117] *See id.* art. 50(4).

[118] *See* Sebastian Hallensleben, *Trust in the European Digital Space in the Age of Automated Bots and Fakes*, EU OBSERVATORY FOR ICT STANDARDISATION 28-33 (Jan. 2022), https://zenodo.org/records/5926395 [https://perma.cc/BAC2-R6TK].

[119] For example, Article 50's transparency requirements state, "Deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated." *AI Act*, *supra* note 8, art. 50(4). Article 72 also specifies post-market monitoring requirements. *Id.* art. 72.

The current system pushes questions of normative acceptability to deployers who are only weakly bound by the AIA. To be democratically legitimate, normative decisions should be taken at a legislative level and not solely left to the technology community. Providers as well as deployers should have legal obligations to guarantee minimum normative standards. As discussed in Section I.E, the AIA's harmonized standards are also unlikely to provide normative guidance.

To see why this approach is so problematic, imagine a water well that supports a village with water that has high levels of lead. Using the AIA's approach to GPAI, rather than fixing the problem at its source, households in the villages will be informed about the high levels of toxicity. The households can then build in water filters if deemed necessary. Establishing clear standards of (un)acceptable toxicity levels and requiring measures to prevent high levels of toxicity at the source, would be more effective in preventing lead poisoning among the village's population. Transparency about harms is not the same as responsibility for harms.[120]

C. Environmental Risks

We cannot underestimate AI's carbon footprint. Some statistics suggest that information and communication technology (ICT) contributes more to climate change than aviation globally and that the energy needed for AI has increased 300,000 times between 2012 and 2018.[121] It takes 360,000 gallons of water per day to cool a medium-size data center.[122] Beyond direct resource costs, one also needs to consider deforestation, animal killings, and environmental

---

[120] This lesson holds true for the obligation to provide a "sufficiently detailed summary of the content used for training." *Id.* recital 107. Instead of acknowledging that current copyright law is insufficient and trying to implement new mechanisms to protect copyrighted works, the Act merely opts for transparency.

[121] Bran Knowles, *Computing and Climate Change*, ACM TECH. POL'Y COUNCIL 2 (Nov. 2021), https://dl.acm.org/doi/pdf/10.1145/3483410 [https://perma.cc/2W2R-WEJD].

[122] *See* Caroline Donnelly, *Why Water Usage Is the Datacentre Industry's Dirty Little Secret*, COMPUTERWEEKLY (Sept. 21, 2021, 7:46), https://www.computerweekly.com/blog/Ahead-in-the-Clouds/Why-water-usage-is-the-datacentre-industrys-dirty-little-secret [https://perma.cc/5HH3-9JHN].

racism—which leads to displacement of indigenous communities or destruction of their environments by flooding, communities that are already least likely to benefit from the technology.[123] Many studies show how the water and energy consumption necessary to develop AI system contributes significantly to climate change.[124]

Given this context, it is appropriate that environmental concerns feature in the AIA. Unfortunately, legislators have again opted for transparency rather than responsibility. Providers of GPAI models (but not providers of PredAI) must report on their energy consumption if known or else estimated,[125] but the AIA does not set standards on acceptable levels of energy consumption or oblige providers to reduce their future carbon footprint.[126]

Such requirements could still be enacted through harmonized standards. For example, Article 40(2) calls for standards on reporting and documentation processes to "improve AI systems' resource performance, such as reducing the high-risk AI system's consumption of energy and other resources consumption during its lifecycle, and on the energy-efficient development of general-purpose AI models."[127] The Commission retains the ability to assess progress and set additional requirements four years after the law comes into force and every three years thereafter.[128] Nonetheless, the concerns about harmonized standards discussed *supra* Section I.E apply equally here. Particularly concerning are the dominance of industry stakeholders in standards bodies and the fact that conformity assessments are done internally.

---

[123] Bender et al., *supra* note 110, at 612-13.
[124] *See, e.g.*, Philipp Hacker, *Sustainable AI Regulation*, 61 COMMON MKT. L. REV. 345, 350-52 (2024).
[125] *See AI Act*, *supra* note 8, annex IX, § 1(2)(e).
[126] For a proposal that the AIA can be interpreted in a way that providers of high-risk systems also need to reduce the carbon footprint, see Hacker, *supra* note 124, at 373.
[127] *AI Act*, *supra* note 8, art. 40(2).
[128] *Id.* art. 112(7).

D. Model Evaluation and Adversarial Testing for GPAIs with Systemic Risks

Providers of GPAI models with systemic risks are required to perform model evaluations, undertake and document adversarial testing, and assess and mitigate possible systemic risks.[129] While a step in the right direction, this provision alone will not be sufficient to mitigate systemic risks.

Risk evaluations do not need to be public or submitted to competent authorities. The AI Office and national competent authorities only need to be informed about the corrective measures taken if serious incidents occur.[130] In other words, public authorities are only informed after harms have occurred.[131] Prior to this, providers only need to undertake internal assessments; their results and method will not be public.[132]

Similar concerns arise in relation to adversarial testing and auditing. Here, again, adversarial testing (e.g., red teaming) is a step in the right direction but insufficient to mitigate systemic risks. Systemic risk can only be evaluated if full access is given to a model, preferably to external parties. The Digital Services Act (DSA),[133] for example, has provisions that allow so-called vetted researchers to gain access to very large online platforms and very large online search engines to investigate systemic risks.[134] The DSA also requires external auditors to evaluate the systemic risks of the very large platforms and search engines and assess if the chosen mitigation strategies are actually successful.[135] These are promising approaches that could be

---

[129] *Id.* art. 55(1)(a).
[130] *Id.* art. 55(1)(c).
[131] For an exploration of how content-moderation tools and risk assessments under the EU's Digital Services Act could be applied to GAI, see Helberger & Diakopoulos, *supra* note 111, at 4; and Hacker et al., *supra* note 96, at 1120.
[132] Recital 163 states that a provider's documentation can be requested by the AI Office. *See AI Act*, *supra* note 8, recital 163.
[133] *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)*, 2022 O.J. (L 277) 1 [hereinafter *DSA*].
[134] *Id.* art. 40.
[135] *Id.* arts. 34, 37.

applied to GPAI models with systemic risks, provided that corporate capture is prevented.[136]

III. AI and Software Liability Directives

The AIA is not the only regulatory framework proposed in the EU in recent years to govern AI. Two other interesting frameworks are currently making their way through the legislative process: the AI Liability Directive (AILD),[137] and an update of the Product Liability Directive (PLD).[138] As Directives, both frameworks will need to be implemented by member state laws, potentially leading to a fragmented standard across the EU.

These frameworks have thus far received significantly less public attention than the AIA despite being equally important. In practice, the two liability directives can be seen as complementary tools to the AIA. Both frameworks aim to increase individual-level rights protections for people who suffer AI-related harms. Limitations on liability for AI are well established.[139] Creating new frameworks in response to such limitations is a sensible legislative step to take. However, both frameworks are unlikely to effectively mitigate novel, AI-related harms.

A. Product Liability Directive

The updated PLD addresses both software- and AI-related harms. However, the definition of harm used in the framework is severely limited in scope, as it relates only to material harm. Material harms must relate to one of the following: (1) death

---

[136] Johann Laux et al., *Taming the Few: Platform Regulation, Independent Audits, and the Risks of Capture Created by the DMA and DSA*, 43 COMPUT. L. & SEC. REV., Nov. 2021, at 1, 8.
[137] *AILD*, *supra* note 6.
[138] *PLD*, *supra* note 6.
[139] *See* Herbert Zech, *Liability for AI: Public Policy Considerations*, 22 ERA F. 147, 150-54 (2021); Christoph Schmon, *Product Liability of Emerging Digital Technologies*, 3 ZEITSCHRIFT FÜR INTERNATIONALES WIRTSCHAFTSRECHT 254, 254-58 (2018); Martin Ebers, *Liability for Artificial Intelligence and EU Consumer Law*, 12 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 204, 216-19 (2021); Christiane Wendehorst, *Strict Liability for AI and Other Emerging Technologies*, 11 J. EUR. TORT L. 150, 153-60 (2020).

or personal injury, including medically recognized harms to psychological health, (2) harm or destruction to property unless exclusively used for professional purposes, or (3) loss or corruption of data that is not exclusively used for professional purposes.[140]

The PLD also does not fully cover immaterial harm. The final text of the PLD has now been formally approved by the European Parliament.[141] Breaking with the Commission's original proposal for a revised PLD,[142] the approved compromise text between the Parliament, Commission, and Council provides for compensation for certain "non-material" harms, which are currently covered by national law.[143] Recital 23 specifies that the PLD should provide "compensation for non-material losses resulting from damage covered by this Directive, such as pain and suffering, . . . in so far as such losses can be compensated for under national law."[144]

This means that only material harm (e.g., "pain and suffering") that is a side effect of, or is associated with, a harm covered by the PLD—death, personal injury, destruction to property, or data loss—will be covered. Plus, compensation will only be granted if the laws of member states recognize these harms in their legal system, leading to fragmented standards across the EU.

Importantly, the revised PLD explicitly does not trigger liability or create a right to compensation for an expanded range of immaterial harms, such as "pure economic loss, privacy infringements or discrimination."[145] Purely economic losses are not covered.[146] Faulty algorithmic decisions that lead to dismissal, losing a promotion, or not being invited to a job interview would not be covered. Longer term financial harms suffered due to being sent to prison or denied bail, not being

---

[140] *See PLD*, *supra* note 6, art. 6.
[141] *PLD*, *supra* note 6.
[142] *Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products*, COM (2022) 495 final (Sept. 28, 2022).
[143] *PLD*, *supra* note 6, art. 6(2).
[144] *Id.* recital 23.
[145] *Id.* recital 24.
[146] Wendehorst, *supra* note 139, at 162. *See* Philipp Hacker, *The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future*, 51 COMPUT. L. & SEC. REV., Nov. 2023, at 1, 28 (criticizing this exclusion).

admitted to university, or being denied loans, credit, or scholarships likewise would not be covered. The same counts true for immaterial harm caused by biased, discriminatory, or privacy-invasive AI systems.

These limitations are especially problematic in relation to AI given that so many of its harms will be immaterial, collective, and borne by society and the individual directly, rather than as a byproduct of a physical or material harm. These types of harms will not necessarily or provably result from damages covered by the PLD. Artificially created outputs that are used for misinformation campaigns,[147] incorrect information and subtle hallucinations,[148] biased information and facts,[149] facial recognition software that is less accurate for people of color,[150] discriminatory outcomes for groups not protected by non-discrimination law,[151] and inaccurate emotion detection software[152] are just some examples of AI-driven, immaterial harms that will not be redressed under the PLD. Other AI-related, immaterial harms are addressed insufficiently through sectoral laws, with these limitations inherited by the PLD; such harms include inaccurately generated credit scores,[153] sexually exploitative and reputation-

---

[147] *See* Anna Wilson et al., *Multimodal Analysis of Disinformation and Misinformation*, 10 ROYAL SOC'Y OPEN SCI., Dec. 20, 2023, at 1, 2-3.
[148] *See* Mittelstadt et al., *supra* note 40, at 1831.
[149] *See* Bender et al., *supra* note 110, at 617-18.
[150] *See* Buolamwini & Gebru, *supra* note 26, at 10-11; Inioluwa Deborah Raji et al., *Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing*, AIES '20: PROC. AAAI/ACM CONF. ON AI ETHICS & SOC'Y 145, 145-46 (2020).
[151] *See* Sandra Wachter, *The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law*, 97 TUL. L. REV. 149, 158-62 (2022); Janneke Gerards & Frederik Zuiderveen Borgesius, *Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence*, 20 COLO. TECH. L.J. 1, 11-15 (2022).
[152] *See* Luke Stark & Jesse Hoey, *The Ethics of Emotion in Artificial Intelligence Systems*, FACCT '21: PROC. 2021 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 782, 787-90 (2021).
[153] *See* Gillis, *supra* note 40, at 1203.

damaging outputs of generative models,[154] biased advertising,[155] price discrimination,[156] and privacy-invasive inferential analytics.[157]

The PLD also places unnecessary and unreasonable evidentiary burdens on victims of AI-related harms that will make it difficult to successfully raise a claim against AI providers and deployers. At first glance, the PLD appears to establish a strict liability regime. However, this is unfortunately not the case.[158] Article 9 explains that claimants are required to prove both the defectiveness of the products in question and a causal link between the damage suffered and the defectiveness of the product.[159]

Claimants must also overcome problems of expertise. The PLD establishes mandatory disclosures of certain evidence from the defendant to the claimant.[160] While it is good that claimants can request such information, the PLD does not set requirements for the level of detail or necessary expertise at which this information is provided. Disclosures thus may be highly technical, making them very difficult for a person lacking domain expertise to understand and use to prove a product's defectiveness.[161]

---

[154] *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1771-75 (2019).

[155] *See* Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMMC'NS ACM 44, 50-52 (2013).

[156] *See* FREDERIK ZUIDERVEEN BORGESIUS, COUNCIL EUR., DISCRIMINATION, ARTIFICIAL INTELLIGENCE, AND ALGORITHMIC DECISION-MAKING 28 (2018), https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 [https://perma.cc/6AJD-XP56].

[157] *See* Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2 COLUM. BUS. L. REV. 494, 505-14 (2019).

[158] In relation to the old PLD, see Christoph Schmon, *EU Product Liability Directive: Not Fit for New Technologies*, *in* KONSUMENTENPOLITISCHES JAHRBUCH 2019 61, 73 (Maria Reiffenstein & Beate Blaschek eds., 2019). For a critique on the new proposal, see Hacker, *supra* note 146, at 5-29.

[159] *See* Schmon, *supra* note 158, at 70-71 (explaining how this has always been an issue with the PLD).

[160] *PLD*, *supra* note 6, art. 9.

[161] Hacker, *supra* note 146, at 18-19.

To overcome this limitation, the PLD introduces a rebuttable presumption of defectiveness.[162] If the defendant fails to comply with the disclosure requirements, or if the claimant shows that the product, for example, did not comply with mandatory safety requirements or that the damage was caused by an obvious malfunction, then this will lead to a rebuttable presumption of defectiveness.[163] This is not the same as a full reversal of the burden of proof. In practice, meeting these requirements poses challenges similar to those discussed above about technical and legal expertise.

The PLD's predominant focus on monetary damages is also extremely limiting. Important AI-driven harms cannot be measured only in monetary terms. For example, how do you measure the humiliation of a person of color when facial recognition software at airports does not work because of their darker skin? How do you measure the indignation of facial recognition software mislabeling people as criminals, or the societal cost of sexist online advertisements, or the harm caused when AI misreads a person's emotions? How much monetary damage should be awarded, and to whom, if large language models slowly erode common knowledge and disrupt scientific integrity?[164]

Other measures, such as mandatory redesign, (temporary) bans, mandatory external audits, or product recalls, would be more useful to address such harms. Monetary damages alone always run the risk of being considered a "cost of doing business," accountable for in annual budgets. Societal and collective harms cannot be remedied alone through monetary means; these harms must be addressed at their source.[165]

---

[162] *PLD*, *supra* note 6, art. 10.

[163] *Id.* art. 10(2).

[164] *See* Mittelstadt et al., *supra* note 148, at 1831; M. J. Crockett et al., *The Limitations of Machine Learning Models for Predicting Scientific Replicability*, 120 PSCYH. & COGNITIVE SCIS. art. no. e2307596120, Aug. 7, 2023, at 1, 1-2.

[165] Inspiration can be drawn from non-discrimination law, where scholars have argued that that individual-level damages are not sufficient and that proactive measures are needed to prevent discrimination. *See* SANDRA FREDMAN, EUR. COMM'N, MAKING EQUALITY EFFECTIVE: THE ROLE OF PROACTIVE MEASURES 8 (2009), https://ec.europa.eu/social/BlobServlet?docId=4551&langId=en [https://perma.cc/ZLZ4-Z98V].

B. AI Liability Directive

The AILD introduces similar challenges to the PLD. In terms of scope, the AILD only covers AI and not software-related harms caused by a provider, deployer, or user of such an AI system—both high-risk and non-high-risk. According to Article 2(9) of the AILD, harms are defined as damage to life, physical integrity, property, and the protection of fundamental rights.[166]

Compared to the PLD's focus solely on material harms, the inclusion of fundamental rights protection, and, by extension, immaterial harm, looks promising at first glance.[167] However, the AILD explains that fundamental rights are only relevant when allowed by member states' traditions.[168] In practice, this will severely limit the Directive's scope of application to public sector AI.

Traditional fundamental- and human-rights doctrine suggests that these rights only bind state actors.[169] This is true for the European Convention of Human Rights and the EU Charter of Fundamental Rights. The Charter only applies to EU institutions and to public institutions of member states when implementing European law.[170] Only in a handful of cases has the European Court of Justice declared that a few fundamental-rights provisions, mainly those addressing workplace discrimination, also apply to private actors.[171]

---

[166] *AILD*, *supra* note 6, art. 2(9). *See also* CHRISTIANE WENDEHORST, ADA LOVELACE INST., AI LIABILITY IN EUROPE: ANTICIPATING THE EU AI LIABILITY DIRECTIVE 14-17 (2022), https://www.adalovelaceinstitute.org/wp-content/uploads/2022/09/Ada-Lovelace-Institute-Expert-Explainer-AI-liability-in-Europe.pdf [https://perma.cc/8G3Y-YKUB] (providing an overview of the scope and remit of the frameworks and its implications).

[167] *But see* Wendehorst, *supra* note 139, at 166-67 (arguing that the exclusion of fundamental rights would be preferable).

[168] *AILD*, *supra* note 6, at 9-10.

[169] For a more detailed discussion of the scope and limitation of fundamental rights, the AIA, and GAI, see Sandra Wachter, et al., Do Large Language Models Have a Legal Duty to Tell the Truth? 22-23 (May 16, 2024) (unpublished manuscript) (available at https://ssrn.com/abstract=4771884 [https://perma.cc/T9RB-8QC4]).

[170] Spaventa, *supra* note 66, at 261-62.

[171] This is the case when a provision constitutes a fundamental principle of the EU, such as the prohibitions against discrimination based on nationality and against unequal pay between men and women. *Id.* at 266-67.

Member states have also historically been very reluctant to extend human rights obligations to private actors. In the context of AI development, which is predominantly driven by the private sector, this trend is problematic.

Beyond human rights, claimants can also appeal to relevant sectoral laws governing particular AI systems and services. However, given the limited scope of human rights and the variability of sectoral laws across member states, the AILD again leaves claimants in an uncertain position to seek legal recourse and facing a potentially fragmented liability standard across Europe. At this stage, it is therefore unclear how, if, and under what conditions fundamental and human rights and immaterial harm will be covered by the AILD.

As with the PLD, claimants will also face a herculean task to successfully bring a claim under the AILD.[172] The Directive stipulates that claimants must prove the defendant's fault, establish a causal link between the fault and the produced output, and prove that the outputs caused the damage.[173]

Here, again, legal disclosure mechanisms are intended to help claimants with proving fault. However, claimants' right to access can be limited according to the risk-based classification and documentation requirements defined in the AIA.[174] Access requests will only be successful if the claimant has demonstrated facts and evidence that support the plausibility for the damages. Trade secrets also need to be protected.[175]

Non-high-risk systems are not covered by the AILD's evidence-disclosure requirements, and the AIA does not set out binding documentation duties for their providers or deployers. The weaknesses of the AIA's risk-based system, discussed *supra* Section I.A, return here to introduce similar limitations within the regime of the AILD. Potentially harmful AI in media, science and academia, most of finance and

---

[172] *See* Hacker, *supra* note 146, at 25.

[173] *AILD*, *supra* note 6, art. 4. On the burden of proof, see Mindy Nunez Duffourc & Sara Gerke, *The Proposed EU Directives for AI Liability Leave Worrying Gaps Likely to Impact Medical AI*, 6 NPJ DIGIT. MED., no. 77, Apr. 26, 2023, at 1, 3.

[174] *AILD*, *supra* note 6, art. 3(1).

[175] *Id.* art. 3(4).

financial trading, and most types of insurance,[176] as well as recommender systems, chatbots, and pricing algorithms, are not included due to the AIA's classification approach. It is likewise unclear whether GPAI models and systems will be classified as high-risk systems and thus fall within the scope of the AILD.

The potential alleviation of the burden of proof is also insufficient. A rebuttable presumption of fault or of a breach of a duty of care is assumed under the AILD if the defendant fails to comply with a court's evidence disclosure order.[177] The information will be, as with the PLD, highly technical, hard to understand without expert knowledge, and limited to high-risk AI systems as defined in the AIA. Article 4(2) specifies other ways for the claimant to prove a failure to comply with specific provisions for high-risk systems which, if successful, would also result in a rebuttable presumption of a causal link between fault and output.[178] However, not all non-compliance automatically leads to a presumption of a causal link.[179] Further, the AIA also limits applicability of this rebuttable presumption when the defendant shows that sufficient evidence exists for the claimant to prove the causal link.[180] A similar limitation applies for non-high-risk AI systems.[181] The rebuttable presumption is only granted when it would be excessively difficult to prove a causal link.[182]

Proving fault in relation to a non-high-risk system is especially difficult. Fault is always coupled to an obligation, but the AIA only establishes obligations for high-risk systems. Thus, claimants will only succeed in cases where other sectoral or member-state laws create obligations. And in addition to proving a link between fault and output, claimants need to show a causal link between that output and damage—a showing for which the burden of proof is not alleviated.[183]

---

[176] Life and health insurance would be covered, but other types of insurance would not. *See id.* annex III.
[177] *Id.* art. 3(5).
[178] *Id.* art. 4(2).
[179] *Id.* recital 25.
[180] *Id.* art. 4(4).
[181] *Id.* art. 4(5).
[182] *Id.* This would be the case if an opaque AI system is in use. *See id.* recital 3.
[183] *Id.* art. 4(1)(c).

Relaxed conditions for the rebuttable presumption also exist for non-professional users[184] which can be problematic, as misinformation campaigns, revenge porn, or reputationally damaging content is often spread by non-professional users.

Finally, the AILD only covers harm caused by a fully automated system. Cases in which a human played a role in a decision-making process fall outside its scope.[185] In practice, it will be quite common for a human to be part of the decision chain, which will further limit the AILD's reach. This exception also creates a loophole through which the inclusion of a token "human in the loop" will avoid liability.[186]

C. Common Weaknesses

The limitations of the updated PLD and the AILD can be summarized in three points: First, claimants must know about the harm caused by AI to raise a complaint, but immaterial harms such as discrimination will often happen without an individual's awareness. AI can exclude people from seeing advertisements for jobs, housing, or financial services, or offer more expensive products without claimants being aware of it.[187] Complaint-based systems alone do not offer enough protection.

Second, much of the harm caused by AI is intangible. Both directives lack clear protections for immaterial harms. Moreover, it is hard to measure these harms solely in monetary terms—or indeed to measure them at all.

What is more, in the EU, compensation for immaterial and material harm has traditionally been relatively low. "[P]ain and suffering" and punitive damages are usually excluded, and the amount of damages that can be awarded is often capped by the member states,[188] though a limitation could not exceed 70 million euros.[189]

---

[184] *Id.* art. 4(6).
[185] *Id.* recital 15.
[186] Wachter et al., *supra* note 92, at 88; Hacker, *supra* note 146, at 13.
[187] *See* Wachter et al., *supra* note 70, at 6, 12; Sandra Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioral Advertising*, 35 BERKELEY TECH. L.J. 367, 377-80 (2020).
[188] Alberto Cavaliere, *Product Liability in the European Union: Compensation and Deterrence Issues*, 18 EUR. J.L. & ECON. 299, 307-08 (2004); *see also* Christopher J. S. Hodges, *Product Liability in Europe: Evaluating the Case for Reform*, 2000 BUS. L. INT'L 171, 175-79.
[189] Council Directive 85/374/EEC, art. 16, 1985 O.J. (L 210).

EU PLD case law is in general very scarce, with a range between 209 and 452 cases annually in the whole of the EU.[190] Furthermore, awarding damages to a comparable level—like in the United States, where claimants can recover millions of dollars—does not really exist in the EU.[191] Statistics from 2022 suggests a range between 20,000 and 1,500,000 euros for compensation paid for deaths, 1,500 to 700,000 euros for personal injuries, and 5,000-25,000 euros for property damage.[192]

The EU's hesitancy to award damages for defective products can also be seen in the old version of the PLD. Article 9 of the old PLD had a monetary cap of 500 euros for "damage to, or destruction of, any item of property other than the defective product itself," as well as an optional overall cap of 70 million euros for all damages.[193] As a result, only property damages above and below this threshold were compensable, which functions as a deterrence to bringing legal action. Although both these caps are lifted in the revised PLD,[194] it is questionable if awarded compensation will increase given that courts were always hesitant to use the full available scale of potential damages.

This means that not only does the monetary approach not fully fit the harm, but also that the compensation is likely to be inadequate. Therefore, additional tools, such as (temporary) bans or mandatory redesign, external audits, or product recalls, will be more useful to reduce harms permanently.

---

[190] *Commission Staff Working Document Impact Assessment Report: Accompanying the Document Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products*, EUR. COMM'N 83 (Sept. 28, 2022), https://single-market-economy.ec.europa.eu/document/download/348b3e35-7d1a-43df-8e9d-296fc09e2c3c_en?filename=SWD_2022_316_1_EN_impact_assessment_part1_v2.pdf [https://perma.cc/J849-DVWA] [hereinafter *Impact Assessment Report*]. I would like to thank Dr. Daria Onitiu for finding these statistics.

[191] For an overview of cases brought in the EU and United States and damages awarded see, Hodges, *supra* note 188, at 177-82.

[192] *Impact Assessment Report*, *supra* note 190, at 85-86.

[193] Council Directive 85/374/EEC arts. 9, 16.

[194] *Questions and Answers on the Revision of the Product Liability Directive*, EUR. COMM'N (Sept. 28, 2022), https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_5791 [https://perma.cc/WL5H-XCLW].

Third, the evidentiary standards required by the PLD and AILD are almost insurmountable for claimants, especially those without domain expertise or legal counsel. Liability is a civil-court procedure, meaning the risks and costs of litigation rest solely on the claimant. Some AI-driven harms may wrongly be seen as "trivial" (e.g., slightly higher prices), which means people are less likely to raise a claim. It is therefore questionable whether individuals will have the time, resources, and expertise to use the legal remedies provided by these directives.

IV. Solutions

While the EU's recent efforts to regulate AI through the AIA, the updated PLD, and the AILD leave much to be desired, there are various avenues available to improve legal protections afforded to individuals and groups impacted by AI systems. In the first instance, the AIA requires the Commission to regularly assess the framework's effectiveness and to adjust it if needed.[195] In addition, the Commission can use "delegated acts" to adjust certain provisions of the framework.[196] The harmonized standards currently being developed by CEN and CENELEC will carry significant quasi-legal weight and provide practical guidelines for AI providers, deployers, and users. As these standards are yet to be finalized, there remains an opportunity to push for stronger "on the ground" protections. Likewise, the liability directives are still being negotiated, with changes possible.

Based on the preceding analysis, I offer several concrete recommendations to improve regulation of AI in the EU.

A. Third-Party Conformity Assessment and External Audits

As discussed *supra* Section I.F, much of the AIA rests on conformity assessments that are conducted by providers of high-risk systems and GPAI models. To increase accountability and oversight, it would be better to have these assessments conducted by independent third parties. This change to enforcement of the AIA would be in line with the nonbinding part of the AIA, which reflects the legislators' awareness of the

[195] *AI Act*, *supra* note 8, art. 112.
[196] *Id.* art. 97.

loophole created by mere internal assessments.[197] A switch to external assessments could be accomplished as the Commission has the power to adopt delegated acts to update internal and third-party conformity assessments.[198]

In relation to high-risk AI systems and all types of GPAI models and systems, external audits could help to detect and mitigate systemic risks. Inspiration can be drawn from the DSA, which grants access to vetted researchers and requires external audits to investigate systemic risks of very large platforms and search engines to assess the effectiveness of mitigation strategies for systemic risks. Internal checks, such as red teaming, will not be sufficient on their own.

B. Clarify Responsibility Along the AI Value Chain for GPAI

One of the main issues facing current GPAI regulation efforts is the predominant focus on providers of GPAI models, and, to a much lesser extent, on providers of GPAI systems and deployers, even though GPAI systems are just as likely to cause (im)material harms. The same is true for narrow AI systems where deployers currently have very limited obligations.

The second issue is the overreliance on transparency as an accountability mechanism. Transparency about harms is not the same as responsibility for harms. The AIA's harmonized standards can be used to create clear, normative, and technical thresholds that must be met by providers, deployers, and users, including concrete guidance for questions around bias, explainability, and performance.

C. Ethical Disclosures by Default

Individuals will often be unaware of harms driven by AI because providers and deployers of AI systems, which are not user- or customer-facing, do not publicly report their impact. Where harms are hidden due to the deployment model or lack of public reporting, complaint-based mechanisms will be insufficient to protect individual rights.[199]

This gap can be closed through the harmonized standards. Provisions should be introduced to require consistent testing

---

[197] *Id.* recital 125.
[198] *Id.* art. 43(5)-(6). *See also id.* annexes VI, VII.
[199] *See supra* Section III.C.

and publication of a summary of the results for affected parties. In particular, with respect to bias and AI, I propose a reversal of the burden of proof.[200] There is no such thing as unbiased data. AI models, systems, and their training and testing data should be assumed to be biased unless proven otherwise. This reversal can be accomplished by publishing the aforementioned testing results and actions undertaken to mitigate and prevent biases.

D. Change the FLOPS Threshold for GPAI Models with Systemic Risks

The current FLOPS threshold of 10^{25} only covers GPAI models such as GPT-4, Gemini, and Llama 3.1. This would exclude models such as GPT-3.5, the version that is freely available to the public. Also excluded are models such as Anthropic's Claude, Aleph Alpha's Luminous, Mistral AI's Mistral Nemo, Meta's Llama 3, Midjourney, Stability AI's Stable Diffusion 3 Medium, and Falcon LLM.[201]

The threshold should be lowered to include computationally smaller models that have similar systemic risks. Other criteria to qualify as having systemic risks, such as a concrete number of end users, should be introduced.[202] This change would also better align the AIA's notion of systemic risk with the DSA, which regulates online platforms and services according to size measured by number of users.[203] External assessments can also help determine whether a GPAI system poses systemic risks independent of FLOPS and user base. In practice, changes to the FLOPS threshold can be made by the Commission through their power to adopt delegated acts[204] to

---

[200] Wachter et al., *supra* note 71, at 762, 775, 780.

[201] *See* sources cited *supra* note 113.

[202] Annex XIII(f) assumes systemic risk when the model is made available to 10,000 registered business users. *AI Act*, *supra* note 8, at annex XIII(f). While the number of end users is also relevant, no concrete number is established. *See id.* annex XIII(g).

[203] Very large online platforms and very large online search engines are those that reach at least 45 million monthly active users in the EU. *See DSA*, *supra* note 133, art. 33(1).

[204] *AI Act*, *supra* note 8, arts. 52(4), 97.

amend the thresholds[205] listed in the AIA and to classify new GPAI models as having systemic risks.[206]

E. Expand Bans and Add Additional High-Risk Categories

As discussed *supra* Section I.A, the risk categories in Annex III are insufficient. AI in media, science and academia, most financial services and trading, and most types of insurance, as well as recommender systems, chatbots, pricing algorithms, and GAI, exhibit well-known and systemic risks for both individuals and society. These types of systems should be included in Annex III.

Emotion detection AI should also be widely banned, and, as a priority, should be banned in immigration and criminal justice. Similarly, there should be a full ban on predictive policing, as well as on facial recognition software used in criminal justice. These changes are necessary due to the high levels of inaccuracy and lack of scientific evidence establishing the reliability of these techniques,[207] as well as to the significant human-rights infringements that their usage in these areas could cause.

Additions to Annex III can be accomplished through delegated acts.[208] Looking to the longer term, the AIA states that the "Commission shall assess the need for amendment of the list in Annex III, the list of prohibited AI practices in Article 5, once a year following the entry into force of this Regulation, and until the end of the period of the delegation of power."[209] The Commission should regularly exercise this power going forward to ensure that the list of prohibited systems remains consistent with technological advances and societal expectations.

F. Reduce AI's Carbon Footprint

Reducing AI's carbon footprint should be a priority for all AI providers. In practice, the AIA's harmonized standards should create clear expectations and rules for the measurement

---

[205] *Id.* art. 52(4).
[206] *Id.*
[207] *See* O'NEIL, *supra* note 27, at 84-105; Big Brother Watch Team, *supra* note 26; Buolamwini & Gebru, *supra* note 26, at 11.
[208] *AI Act*, *supra* note 8, arts. 6(6), 7(1), 97.
[209] *Id.* art. 112(1).

and reduction of environmental impact to ensure that providers work toward this common goal. Providers and deployers should likewise be encouraged, as part of their internal assessments, seriously to ask whether the environmental impact of AI is justified for specific use cases under consideration prior to undertaking development or procurement.

G. Reforms of Liability Directives

Finally, both liability directives make it difficult for individuals to prove fault and defectiveness and to establish a causal link between harms caused by AI-specific faults and defectiveness. I propose the introduction of a lower evidentiary burden for claimants through a strict liability (e.g., no-fault) regime. A robust reversal of the burden-of-proof mechanism could also be introduced to replace the current rebuttable presumption.

The scope of harms covered by these directives should also be expanded. Harm should be redefined to include immaterial harms beyond those directly caused by harms covered by the PLD and national law. Many of AI's known harms are immaterial (e.g., purely economic losses, faulty emotion detection, facial recognition software, privacy-invasive algorithms, racist and sexist advertisements), and yet these explicitly are not covered under the PLD and are unlikely to be covered under the AILD. As they currently stand, both directives will miss a huge range of severe, AI-driven harms.

AI's harms are not only immaterial but also societal. A punitive system that only focuses on individual cases and monetary compensation is not good enough. The harms caused by mislabeling people as criminals, eroding scientific integrity, and misinformation campaigns are felt by society, not just individuals. To combat faulty or inaccurate AI systems, other legal tools should be employed. This could include mandatory redesign, (temporary) bans, and mandatory external audits.

The proposed changes outlined here are only a first step toward truly effective AI regulation in Europe. However, they provide a roadmap for building on existing regulatory and standardization efforts to ensure that the law effectively protects groups as well as individuals, and that it holds the providers, deployers, and users of harmful AI accountable for the long-term societal impact of their systems.

Conclusion

AI has long become an integral part of our everyday life. Although many praise the resulting increases in productivity and resourcefulness, these efficiency gains come with many ethical, legal, and societal challenges. The AIA, PLD, and ALD are the EU's first comprehensive attempt at tackling some of these challenges. In this Essay, I have shown that these frameworks—though laudable attempts—leave much to be desired. I have pointed out substantive loopholes and accountability gaps in regulating PredAI and GAI. However, there is cause for hope. Though these frameworks are imperfect, the regulatory mechanisms I have outlined in this Essay are available to close the identified loopholes, to create a system that prevents harmful technology, and to foster ethically and societally beneficial innovation.