AI & SOCIETY
https://doi.org/10.1007/s00146-023-01723-z

OPEN FORUM

Taking AI risks seriously: a new assessment model for the AI Act

Claudio Novelli[1] · Federico Casolari[1] · Antonino Rotolo[1] · Mariarosaria Taddeo[2,3] · Luciano Floridi[2,1]

Received: 19 May 2023 / Accepted: 27 June 2023
© The Author(s) 2023

Abstract
The EU Artificial Intelligence Act (AIA) defines four risk categories: unacceptable, high, limited, and minimal. However, as these categories statically depend on broad fields of application of AI, the risk magnitude may be wrongly estimated, and the AIA may not be enforced effectively. This problem is particularly challenging when it comes to regulating general-purpose AI (GPAI), which has versatile and often unpredictable applications. Recent amendments to the compromise text, though introducing context-specific assessments, remain insufficient. To address this, we propose applying the risk categories to specific AI scenarios, rather than solely to fields of application, using a risk assessment model that integrates the AIA with the risk approach arising from the Intergovernmental Panel on Climate Change (IPCC) and related literature. This integrated model enables the estimation of AI risk magnitude by considering the interaction between (a) risk determinants, (b) individual drivers of determinants, and (c) multiple risk types. We illustrate this model using large language models (LLMs) as an example.

Keywords Risk assessment · Artificial intelligence · AI Act · Climate change · EU · IPCC

1 Overview

The EU Artificial Intelligence Act (AIA) categorizes AI systems (AIs) into four risk categories—unacceptable, high, limited, and minimal—assigning corresponding regulatory burdens to their providers. Unfortunately, the four risk categories are statically dependent on the fields of application of AI systems. For instance, AIs for facial recognition or social scoring are deemed unacceptably risky and prohibited (Article 5). Likewise, AIs used in fields such as education, employment, migration, justice, and law enforcement are considered high risk and, therefore, undergo conformity assessment procedures (hidden reference) and require additional safeguards (Article 8 ff.). The AI risk is conceived as legal in nature, expressing the potential detriment that comes from the violation of a legal value by an AIs (Mahler 2007), but the AIA treats these values as technical standards, which are either met or not (Smuha et al. 2021). Thus, the AIA predetermines the outcome of the balancing test between the values and interests of the exposed community, with no option for revision of risk management measures based on further circumstances. This causes a mistaken evaluation of the risk magnitude of AI— i.e., the likelihood of detriment and severity of consequences—which leads to ineffective legal rules, too strict or lenient. As legal compliance always comes at a cost (Khanna 2021)—and regulatory burdens cannot be eased by a proportionality judgment—the AIA may become unsustainable for AIs providers or deployers. The EU strategy on AI may be jeopardized, discouraging innovation, and forfeiting AI's potential benefits for the values the AIA aims to protect. Thus, the AIA needs a clear model of risk assessment (see below).

The AIA risk categorization is particularly inadequate for regulating general-purpose AI (GPAI), such as large language models (LLMs), or foundation models, which have versatile and unpredictable applications, even for their creators. The lack of intended purposes of GPAIs makes it even more arbitrary to predetermine their risk level based on AI scopes and the abstract weight of the values involved.

Moreover, it is important to point out that the compromise text, approved on 14 June 2023 by the European Parliament, contains two critical changes to the first draft, introducing (a) an additional assessment stage that makes

✉ Claudio Novelli
claudio.novelli@unibo.it

[1] Department of Legal Studies, University of Bologna, Via Zamboni, 27/29, 40126 Bologna, Italy

[2] Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

[3] Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK

Published online: 12 July 2023

high-risk categorization less automatic and (b) a fundamental rights impact assessment. As for the first change, AI systems to be classified as high-risk must also pose what is called a 'significant risk', requiring evaluation of the risk's severity, intensity, likelihood, duration, and potential targets, whether an individual, multiple people, or a specific group (AIA, Recital 32). The second update mandates deployers of high-risk systems to conduct a fundamental rights impact assessment and develop a risk mitigation plan in coordination with the national supervisory authority and relevant stakeholders before market entry (AIA, Recital 58a). These changes are welcome and mark substantial advancements. However, it remains unclear what methods will be used for these evaluations and why their application is exclusively confined to high-risk systems.

To effectively implement AIA, especially when evaluating the significant risk and the impact on fundamental rights, we propose a risk assessment model that provides the risk magnitude of AIs in specific scenarios based on multiple interacting factors. To identify and combine these risk factors, we adapt the framework developed by the Intergovernmental Panel on Climate Change (IPCC), further refined by the related literature (Simpson et al. 2021), which assesses the risk magnitude of a phenomenon based on the interaction among (1) four determinants of risk, (2) individual drivers of determinants, and (3) extrinsic types of risk. This approach offers a more structured approach to the last modifications introduced by the EU legislator. We suggest extending this assessment so that, based on the risk magnitude stemming from the specific scenario, an AI system will be treated as unacceptable, high-risk, limited-risk or minimal-risk. We shall see how this risk assessment model applies to an LLM, as a prototype of GPAI.

2 Risk assessment in climate change: the expanded IPCC model

Climate change risk and AI risk share some similarities. Both exhibit highly unpredictable risk magnitudes and escalating complexity due to the interplay of multiple factors. Moreover, they display a substantialdependence on the specific context and the impacted parties. For these reasons, both necessitate a continuous evaluation of trade-offs in risk mitigation efforts. Given these similarities and the advanced nature of climate risk assessment models in the literature and policy reports, we use the IPCC model as a starting point, while considering relevant literature for further refinement.

The IPCC views climate change risks as the consequence of hazard (H), exposure (E), and vulnerability (V). Hazard refers to potential sources of harm. Exposure refers to what might be affected by the hazard source. Vulnerability refers to attributes or circumstances that make exposed elements susceptible to harm (Cardona et al. 2012). Simpson et al. 2021 expanded the IPCC framework by introducing a fourth determinant, the response (R), which refers to measures that counteract or mitigate risk. They also included interactional risk types with their determinants and the individual components of the determinants—i.e., the drivers—in the risk assessment model. Thus, the overall risk results from the interaction among (1) determinants, (2) drivers, and (3) risk types.

The weight of each determinant is given by the drivers and their interactions, both within and across determinants. Interactions among drivers may be aggregate, compounding, or cascading. The same applies to interactions between multiple risk types (Simpson et al. 2021). Shows the three sets of interactions, occurring at stages of increasing complexity (Fig. 1).

The shortcoming of the AIA is that it considers only the lowest stage, taking risk determinants without the interactions among their drivers (or with cross-sectorial risks).

Adapting the IPCC model to AI, hazard drivers (H) may be purely technological, socio-technical or caused by human–machine interactions: e.g., the opacity of the model, data biases, interaction with other devices, and mistakes in coding or supervision. The last three hazard drivers generally interact in an aggregate way. The interaction is compounded when, e.g., low data representativeness compounds with overfitted machine learning models or biased data. It is cascading when, e.g., model opacity triggers cascading hazards of unpredictability, unmanageability, or threats to security and privacy. An accurate reconstruction of these interactions can provide evidence about the simplicity or complexity of the causal chain between hazard and harm, as well as its likelihood and distribution (Black & Baldwin 2012).

Exposure drivers (E) for AI risk may be tangible assets, like goods or environment, or intangible assets, like values. The exposed asset of the AIA mainly consists of EU fundamental values, e.g., health, safety, justice, and equality. Interactions between exposure drivers are aggregated if, e.g., an AI's adverse effects on the right to asylum and the privacy of asylum seekers. It is compounded if, e.g., an AI's adverse effect on the environment compounds with those on health. It is cascading if, e.g., an AI's adverse effect threatens access to education, and, thus, equality and democratic legitimacy.

The interaction between the exposed values of the AIA often requires balancing them through a proportionality judgment (Alexy 2002). This type of judgment helps determine whether risk mitigation measures for a specific risk category are disproportionate to the specific scenario through quantitative analysis. Risk categories are evaluated by weighing the positive impact of an AIs on values served by its intended scope against those of the exposed asset, using a proportionality test based on three principles:

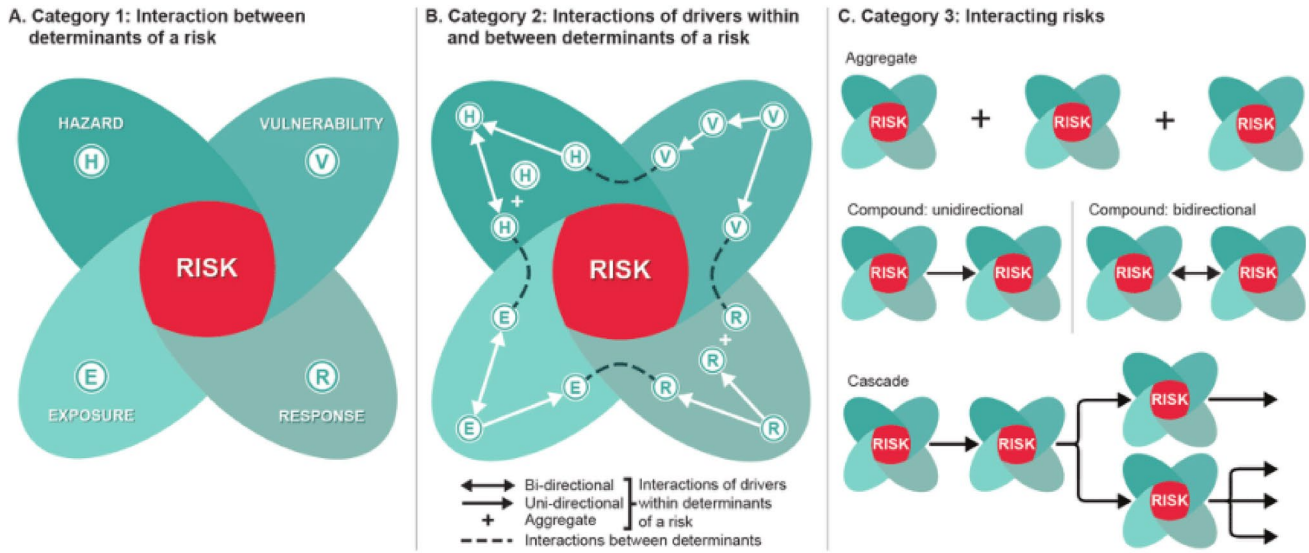

Fig. 1 Three stages of increasingly complex climate change risk by (Simpson et al. 2021)

suitability (a risk category that harms one value is suitable only if it has a positive impact on another value), necessity (when two means for promoting a value are equally suitable, the one that minimizes interference with other values ought to be chosen), and proportionality in the narrow sense (the greater the detriment to one value, the greater the importance of fulfilling the other). The test evaluates whether the benefits of a risk category to one value outweigh the harm it causes to another value.

Vulnerability drivers (V) in AI risk include income, education, gender, ethnicity, and health, as well as the lack of control bodies, procedures, or policies. Interactions among vulnerability drivers are aggregated if, e.g., deployment of an AI system in a vulnerable environment lacks surveillance or feedback mechanisms. Interactions between vulnerabilities can be compounded, as they intersect and influence each other. Interactions between vulnerability drivers are cascading if, e.g., the absence of AIs liability rules triggers other vulnerabilities for those under the adverse effect of AIs use.

Response determinant (R) indicates the environment's resilience to a specific risk. Response drivers in AI can be institutional safeguards on the development, design, and deployment of AIs. Consequently, risk assessment and categorization within the AIA should consider existing legal measures that mitigate the adverse effects of AI technologies, e.g., those contained in the GDPR. Adaptation and mitigation responses can affect risk levels, allowing discrimination of intrinsic vs net risk.

The third stage focuses on the interplay between AI risk and (interactional) risk types, which can be extrinsic—such as market, liability, and infrastructure risks—or ancillary. An aggregate interaction occurs between AI risk and policy risk: adverse effects of ineffective policies or regulations—perhaps external to AI—cumulate with the adverse effects of AIs' deployment. AI risk can then compound with the risk of the digital infrastructure in which an AIs operates. Finally, AI risk may cascade into other types of risk: the risk to innovation, digital sovereignty, economic sustainability, power concentration, and so forth. Ancillary risks are those posed or increased by the risk regulation itself: for example, banning AIs should be justified also against the loss of opportunity benefit of their use, the barriers to technological innovation that the ban raises, and the threats posed by the systems replacing the banned ones (Sunstein 2004). The AIA's regulatory choices cannot be justified just by their positive impact on the intended scope—i.e., the protection of fundamental rights—but also by the (difference between) the marginal gains and harms they generate for other values at stake (Karliuk 2022).

Assessing AI risk through hazard chains, trade-offs among exposed values, vulnerability profiles, and cross-sectorial risks provide a more accurate analysis of its risk. This approach turns the AIA risk categories into dynamic risk scenarios, changing with the interactions among factors, and ensures more proportionate regulatory measures.

Coherent governance of such an assessment must be ensured. Institutional bodies, such as national supervisory authorities (AIA, Title VI), should construct risk scenarios while following the EU legislator's political direction. The latter should identify and evaluate the key drivers of the four risk determinants and the main interactional risk types. Key risk drivers might be identified within the same AI scopes of the AIA, perhaps through its implementing acts, thus limiting Member States' discretion. In the next section, we

illustrate, with the example of LLMs, how some risk drivers can already be derived from the AIA.

One more aspect warrants consideration. Although the categorization of risk in the AIA is coarse-grained, connecting risk management measures to broad scopes of AIs makes it procedurally easier to approve and monitor them for marketing. Therefore, one may object that a risk categorization based on scenarios, which combines multiple risk factors might be too demanding, as it complicates the AIA procedures. The objection is reasonable but, in the end, resolvable by distinguishing short-term from long-term aspects.

In the short term, a scenario-based risk assessment may indeed deter AI deployment and investment. To mitigate this, different strategies might be recommended to make our proposal more sustainable. First, European legislation might indicate, in the AIA's implementing acts, the key risk drivers for each broad AI scopes already outlined in the regulation (e.g., in the Annex III). This would ease the task of deployers and minimize arbitrariness in the AIA's enforcement. We shall illustrate this in the next section. Second, automating risk identification and management can streamline processes. Finally, a phased, iterative approach starting with a granular risk assessment only for a few deployers—maybe with lower risky systems and then with lower compliance costs—might enable procedural refinement and prepare others for a smoother implementation. This means that, in the long term, the benefits of decreased compliance costs will offset the costs, as contextually tailored risk assessments yield less over-inclusive risk categories and more effective risk prevention or mitigation measures.

3 Illustration: large language models

Let us apply this risk assessment model to a LLM specialized in dialogue, recently popularized by OpenAI's ChatGPT.[1] Differently from traditional AI models, LLMs display wider scope and autonomy. Their smooth scalability enables them to process input from diverse domains without extensive training. At the same time, their unpredictable outputs raise concerns. The risk drivers here identified for LLMs can be easily inferred from the AIA, e.g., from the new Article 4a, which contains 'General principles applicable to all AI systems'. Of course, applying our proposed assessment model during the AIA implementation stage would necessitate enhanced legislative transparency in setting the drivers and interactional risk types.

The hazard drivers (H) of LLMs would be the inner opacity of the model, the size of the dataset, and the poor quality or misuse of the training data (e.g., AIA, Art.10). When these hazard drivers compound, they can lead to the AIs perpetrating, for instance, discrimination biases.

The exposure drivers (E) consist of the values potentially damaged by the use of LLMs specialized for dialogue. This would include legal principles, such as violating the copyright of the training data (e.g., AIA, Art. 28b) or the privacy of data subjects (e.g., AIA, Article 4a). The overall weight of the determinant is established by balancing potentially damaged values with those that the LLMs aim to enhance, e.g., public safety.

The vulnerability drivers (V) include attributes that increase the susceptibility of individuals or groups to the adverse effects of automated processing of natural language, which may foster discrimination or misinformation: e.g., ethnicity, gender, wealth, age, and education (e.g., AIA, Art. 4a).

The response drivers (R) would be those measures that counter the hazards of LLMs. They might be governance measures, such as standards for data quality and collection, transparency, bias examination, and human oversight (e.g., AIA, Recital 60f and Artt. 16 and 29). A response measure for LLMs is differential privacy, which adds noise to the training data preventing personal information from being leaked by adversary attacks (Pan et al. 2020).

Finally, the risk of LLMs interacts with extrinsic risk types, e.g., compliance risk, liability risk, and economic risk. Inadequate rules for liability allocation may increase LLMs' risk and may, in turn, cause the risk of a breakdown of the AI market. The overall risk should also be balanced with ancillary risks: e.g., entry barriers for LLMs' providers, or strict rules on training data sources, which may weaken competition and technological innovation. Radical bans may become missed opportunities for the general social interest.

4 Conclusions

This risk assessment model offers two contributions. First, it enhances AIA enforcement by facilitating the development of more sustainable and effective risk management measures for national regulators and AI providers, while pursuing the AIA's objective of protecting the EU values. Second, it favors a granular regulation of GPAIs using scenario-based risk assessment to adapt to their versatile and uncertain applications.

[1] The issue generated a major debate, resulting in the proposal of a series of amendments to the draft AIA: https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends intechnology, arts, science and society, commenting on issues of concernto the research community and wider society. Whilst the drivefor super-human intelligence promotes potential benefits to widersociety, it also raises deep concerns of existential risk,

therebyhighlighting the need for an ongoing conversation between technologyand society. At the core of Curmudgeon concern is the question:What is it to be human in the age of the AI machine? -Editor.

Author contributions In the creation of this paper, the authors divided the sections according to their areas of expertise and interest. Section 1 was written by MT and LF. Section 2 was written by CN. Section 3 was written by FC and AR. Finally, all authors contributed to the conclusion of the paper, providing their unique perspectives and final thoughts. Throughout the process, all authors participated in revising the manuscript and approved the final version for submission.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. Fujitsu, rep. 95/2021, Claudio Novelli.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alexy R (2002) A theory of constitutional rights. Oxford University Press

Black J, Baldwin R (2012) When risk-based regulation aims low: approaches and challenges. Regulation & Governance 6(1):2–22. https://doi.org/10.1111/j.1748-5991.2011.01124.x

Cardona OD, Aalst MKV, Birkmann J, Fordham M, Gregor GM, Rosa P, Pulwarty RS, Schipper ELF, Sinh BT, Décamps H, Keim M, Davis I, Ebi KL, Lavell A, Mechler R, Murray V, Pelling M, Pohl J, Smith AO, Thomalla F (2012) Determinants of risk: exposure and vulnerability. Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 65–108. https://doi.org/10.1017/CBO9781139177245.005

Karliuk M (2022) Proportionality principle for the ethics of artificial intelligence. AI Ethics. https://doi.org/10.1007/s43681-022-00220-1

Khanna VS (2021) Compliance as costs and benefits. In: van Rooij B, Sokol DD (eds) The Cambridge handbook of compliance. Cambridge University Press, Cambridge, pp 13–26. https://doi.org/10.1017/9781108759458.002

Mahler T (2007) Defining legal risk (SSRN Scholarly Paper Fasc. 1014364). Accessed on 15 Sept 2022. https://papers.ssrn.com/abstract=1014364

Pan X, Zhang M, Ji S, Yang M (2020) Privacy risks of general-purpose language models. 2020 IEEE Symposium on Security and Privacy (SP). p. 1314–1331. https://doi.org/10.1109/SP40000.2020.00095

Simpson NP, Mach KJ, Constable A, Hess J, Hogarth R, Howden M, Lawrence J, Lempert RJ, Muccione V, Mackey B, New MG, O'Neill B, Otto F, Pörtner H-O, Reisinger A, Roberts D, Schmidt DN, Seneviratne S, Strongin S, Trisos CH (2021) A framework for complex climate change risk assessment. One Earth 4(4):489–501. https://doi.org/10.1016/j.oneear.2021.03.005

Smuha N, Ahmed-Rengers E, Harkens A, Li W, Maclaren J, Piselli R, et al. (2021) How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an Artificial Intelligence Act; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991

Sunstein CR (2004) Risk and reason. In Cambridge Books. Cambridge University Press. Accessed on 12 Sept 2022. https://ideas.repec.org/b/cup/cbooks/9780521016254.html

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.