How to evaluate the risks of Artificial Intelligence: a proportionality-based, risk model for the AI Act

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

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

*Email of correspondence author: claudio.novelli@unibo.it

Abstract. The EU proposal for the 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 systems (AIs), the risk magnitude may be wrongly estimated, and the AIA may not be enforced effectively. Our suggestion is to apply the four categories to the risk scenarios of each AIs, rather than solely to its field of application. We address this model flaw by integrating the AIA with the framework arising from the Intergovernmental Panel on Climate Change (IPCC) reports and related literature. This makes possible addressing AI risk considering the interaction between (a) risk determinants, (b) individual drivers of determinants, and (c) multiple risk types. Then we integrate the proposed model with a proportionality-based balance among values considered by the AIA's risk analysis. The resulting semi-quantitative approach identifies a more efficient way to implement the AIA and addresses the regulatory issue of general-purpose AI (GPAI).

Keywords: risk assessment, AI Act, IPCC, proportionality, Artificial Intelligence

1. Introduction: from broad scopes to risk scenarios

The proposal of the European Commission for the Artificial Intelligence Act (AIA) is the first comprehensive legal framework on AI by a major supranational regulator. One of the critical aspects of the AIA is the classification of AI systems (AIs) into four risk categories: unacceptable, high, limited, and minimal. The legislator allocates regulatory burdens to AIs' providers so that the greater the risk posed by AIs, the greater the legal safeguards to minimise it. This approach is mainly based on the EU product safety legislation.[1] Still, it is compatible with a risk management standard used in the UK and other legislations for safety-critical industries: the As Low As Reasonably Practicable (ALARP) principle.

The AIA's current risk approach undermines its effective implementation due the lack of a granular risk assessment model (model flaw). Indeed, the AIA rests on a static view of AI risk: it does not consider the interconnection among hazard sources, vulnerability profiles, and exposed values, but treats them as stand-alone technical standards. As a result, risks posed by AIs are assessed considering the impact they may have on European fundamental values[2], without taking into account other risk factors and the interactions between them. Moreover, the AIA lacks a proportionality judgement between the risk mitigation measures and the principles and rights involved. Although the risk outlined in the AIA is legal in nature, and thus requires primarily a qualitative analysis, the categorisation of risk would benefit from a quantitative analysis and the related increased certainty.

The flaws in the risk design of the AIA lead to a significant problem of scope. The AIA assigns AIs to the four risk categories on the basis of broad fields of application. This approach may misestimate the magnitude of AI risks – i.e., the likelihood of detriment and severity of consequences on values like health, safety, privacy, and others – and make the overall legal framework ineffective, that is, with rules that are either too stringent or too soft for the actual applications of specific AIs.

In this regard, it is important to note that the compromise text from 11 May 2023[3] contains two critical changes to the first draft, introducing (a) an additional assessment stage that makes 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 (e.g., 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 (e.g., AIA, Recital 58a and Article 29a). We believe these changes are welcome and mark substantial advancements. However, it remains unclear what standards or methods will be used for these evaluations and why their application is only to high-risk systems.

---

[1] Cf. European Commission, Explanatory Memorandum to AIA, para 1.3.

[2] As they are legally framed as principles and rights, we will use these expressions interchangeably.

[3] https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

To effectively implement AIA, especially when evaluating the significant risk and the impact on fundamental rights, we suggest shifting from a scope-oriented categorisation of AI risks to one based on risk scenarios involving interactions among multiple risk factors. The four risk categories should be applied horizontally to AIs so that, under varying scenarios, the same system can be estimated as unacceptable, high-risk, limited-risk or minimal-risk. Otherwise, the application of the AIA can be enhanced by developing risk subcategories or by facilitating confirmatory and rejection evaluations of the default categorization, solely based on scopes (as implied by the AIA's compromise text). In any case, this calls for a two-stage risk analysis, addressing the flaws in the AIA by combining a qualitative and a quantitative risk assessment methodology (i.e., semiquantitative approach).

To address the model flaw, we need to identify those factors[4] that affect risk scenarios and combine them. While some legal arguments have been presented, suggesting a reading of the AIA's risks approach in light of tort law[5], we draw from research and policy analysis on climate change risk. In particular, we refer to the framework developed by the Intergovernmental Panel on Climate Change (IPCC) working groups and refined by the subsequent literature.[6] In this framework, the risk of a phenomenon is assessed by the interaction between (1) determinants of risk (i.e., hazard, exposure, vulnerability, and responses), (2) individual drivers of determinants, and (3) other types of risk (i.e., extrinsic, and ancillary risks). Once applied to AIs, this framework provides the risk magnitude of AIs under a given scenario. This is a measure defined on the basis of hazard chains, the trade-off among impacted values, the aggregation of vulnerability profiles, and the contextualisation of AI risk with risks from other sectors.

We ground this qualitative analysis on a quantitative assessment. In fact, the risk magnitude should be assessed by weighing the fundamental values (positively and negatively) affected by AIs against the intensity of the interference of AIA's risk containment measures on the same values. This type of judgment for interference between constitutional principles is the object of the proportionality test by Robert Alexy[7], one of the few quantitative approaches to balancing legal principles. In our case, the outcome of the test would indicate whether a risk category is appropriate for an AI under a specific risk scenario or whether it introduces grossly disproportionate limitations and trade-offs for competing values. This could be a way for implementing the fundamental rights impact assessment recently introduced in the draft.

The rest of the article is structured as follows. Section 2 presents the risk-based regulation of the AIA, bridging the risk model within the EU proposal and the ALARP

[4] We shall use the expression 'risk factors' to refer in a general way to all variables potentially able to increase or decrease the risk of an event. We shall specify its meaning by referring to determinants and drivers.

[5] Chamberlain J. The Risk-Based Approach of the European Union's Proposed Artificial Intelligence Regulation: Some Comments from a Tort Law Perspective. *European Journal of Risk Regulation* 2022; 1–13.

[6] Simpson NP, Mach KJ, Constable A, Hess J, Hogarth R, Howden M, et al. A framework for complex climate change risk assessment. *One Earth* 2021; 4(4):489–501.

[7] Alexy R. A Theory of Constitutional Rights. Oxford, New York: Oxford University Press, 2002.

principle. Section 3 discusses the strengths and weaknesses of the AIA risk-based regulation. Section 4 shows how to overcome the AIA's model flaw by using the IPCC framework for climate change risk assessment updated by the relevant literature. Section 5 shows offers a quantitative support to the model through a proportionality test. Section 6 discusses the competence and division of labour between supranational and national bodies in risk scenario building and proportionality assessment. Section 7 outlines the advantages of modifying the AIA's strategy towards risk in its enforcement and regulation of GPAIs (GPAI). Section 8 concludes the article.

2. AIA's risk-based regulation

Generally, risk-based regulations consist of (at least) three phases: assessment, categorisation, and management.[8] In this article, we shall focus more on the first two phases and less on the AIA's risk management system, that is, legal safeguards and requirements.

The AIA relies on the traditional conception that risk is the likelihood of converting a source of hazard into actual loss, injury or damage.[9] Sources of danger are those uses of AI that are most likely to compromise safety, health, and other values. Being the likelihood of damage, risk can be expressed through the ratio between hazard and safeguards so that, as the safeguards increase, the risk quotient decreases:

$$risk = \frac{hazard}{safeguards}$$

The risk may become untenable if safeguards do not offset severe hazards. The regulatory intervention should be proportionate to the hazards net of safeguards. Risk tolerance thresholds – in the AIA, the risk categories – indicate which risks are accepted without (strong) precautions and which instead require (further) mitigation practices.

In the AIA, the benchmark to calculate the risk of AIs is their potential adverse impact on health, safety, and EU fundamental rights. As a result, the AIA classifies AIs according to four risk categories: unacceptable, high, limited, and minimal.[10, 11] Stricter requirements are prescribed for suppliers and users of riskier AIs. This is explicitly stated in Recital 14 of the draft:

---

[8] Millstone, E, et al. (2004). Science in trade disputes related to potential risks: comparative case studies. IPTS technical report series EUR 21301 EN, European Commission Joint Research Centre/ IPTS Institute for Prospective Technological Studies, Brussels/Luxembourg.

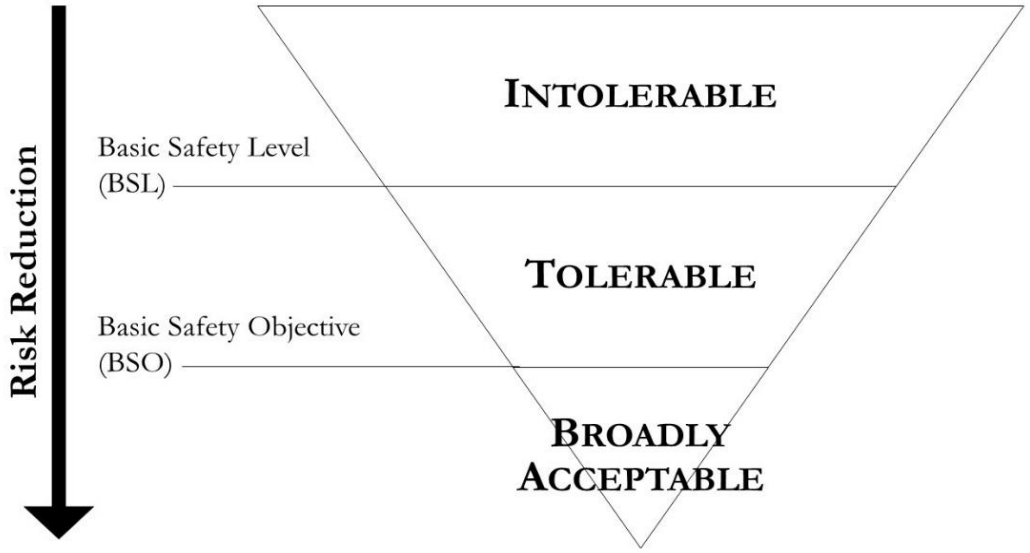
[9] This conception can easily be deduced from some sections of the draft, for instance, Recital 32 referring to high-risk systems: "[…] high-risk AI systems other than those that are safety components of products…it is appropriate to classify them as high-risk if, in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity of the possible harm and its probability of occurrence […]".

[10] Kaplan S, Garrick BJ. On The Quantitative Definition of Risk. *Risk Analysis* 1981; 1(1):11–27.

[11] The text refers to three categories, but a fourth sub-category of high-risk systems can be derived from the presence of lighter obligations.

> "In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate".[12]

This is why the AIA modulates the legal requirements to make the risk of deploying AIs at least tolerable. The tolerance thresholds that constitute the AIA's risk categorisation seem to be inspired by the ALARP principle. ALARP is a general principle in UK law for risk management systems in safety-critical industries[13], and in the UK health system.[14, 15] ALARP-inspired approaches involve a proportionality review of risk reduction measures so that they are not exorbitant to the improvement gained.[16] Typically, ALARP provides the following risk tolerance ranges[17]:



[12] 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, Com/2021/206 final, Recital 14.

[13] Abrahamsen EB, Abrahamsen HB, Milazzo MF, Selvik JT. Using the ALARP principle for safety management in the energy production sector of chemical industry. *Reliability Engineering & System Safety* 2018;169:160–5. Jones-Lee M, Aven T. ALARP—What does it really mean? *Reliability Engineering & System Safety* 2011;96(8):877– 82.

[14] UKHSE. Risk management: Expert guidance - ALARP at a glance. https://www.hse.gov.uk/managing/theory/alarpglance.htm

[15] The principle is also recognised in other legal systems, like the US, sometimes under the formula " as low as reasonably achievable" (ALARA).

[16] Bai Y, Jin WL. Risk Assessment Methodology. In: Bai Y, Jin WL, curatori. Marine Structural Design (Second Edition). Oxford: Butterworth-Heinemann 2016; 709–23.

[17] Hurst J, McIntyre J, Tamauchi Y, Kinuhata H, Kodama T. A summary of the 'ALARP' principle and associated thinking. *Journal of Nuclear Science and Technology* 2019;56(2):241–53.

Figure 1. The tolerance ranges of the ALARP principle.[18]

Although the transposition into EU law of ALARP is limited[19] and controversial[20], AIA's risk categories overlap with the tolerance ranges shown in Figure 1.[21] These risk categories can be summarised as follows.[22]

Unacceptable risk includes (AIA, Title II):

- AIs that may cause significant harm through (a) subliminal manipulation of individuals' consciousness that distorts their behaviour or (b) exploitation of vulnerabilities – age, physical or mental disability – of a specific group of people that distorts the behaviour of its members.
- AIs for social scoring that evaluate or classify natural persons or groups based on their social behaviour when social scoring leads to detrimental or unfavourable treatment (a) in social contexts that are unrelated to the contexts in which the data was originally generated or collected; (b) detrimental or unfavourable treatment are unjustified or disproportionate to the social behaviour of natural persons or groups.
- AIs for biometric categorisation that categorise natural persons according to sensitive or protected attributes or characteristics (e.g., gender, ethnicity, political orientation, religion, disability) or based on the inference of those attributes or characteristics.[23]
- AIs for risk assessments of natural persons or groups to assess the risk for offending or reoffending or for predicting the occurrence or reoccurrence of (an actual or potential) criminal or administrative offence based on assessing personality traits and characteristics, such as the person's location, past criminal behaviour of natural persons or groups of natural persons.

[18] This is a simplified version of the figure found in Hurst J. et al. (n 16). Note that the size of the three categories of the inverted pyramid is related to the severity and not to the numerousness of the relative risks.

[19] A specific reference may be found in the EU legislation on medical devices: cf. Annex I, Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, [2017] OJ L117/1.

[20] In particular, the debate developed when the European Commission argued that the use of ALARP (in the homologous version of SFAIRP) in the UK's Health & Safety at Work Act was not consistent with the European "Framework Directive" for occupational safety and health (Directive 89/391/EEC), asking thus the Court of Justice to declare that the Member State failed to fulfil its obligation to correctly transpose the Directive. However, the European Court of Justice, without taking a specific position on the compatibility of ALARP with the Directive's provision, dismissed the action brought by the European Commission, maintaining that the EU institution did not clearly identify the legal standard enshrined in the Directive that the UK failed to implement (Case C-127/05 *Commission v UK* EU:C:2007:338, para 58).

[21] Indeed, a textual reference to ALARP can be found in the section where the AIA describes the mandatory risk management system for high-risk systems: "In identifying the most appropriate risk management measures, the following shall be ensured: (a) elimination or reduction of risks as far as possible through adequate design and development; […]".

[22] This list is updated to the May 2023: Draft Compromise Amendments on the Draft Report (COM(2021)0206 - C9 0146/2021 - 2021/0106(COD)). However, the text is not yet conclusive.

[23] When the exonerating circumstances provided for in Articles 5(1)(d) and 5(2)(4) are not met.

- AIs for inferring emotions of a natural person in the areas of law enforcement, border management, in workplace and education institutions.

High-risk includes (AIA, Title III):

- AIs used as safety components of products covered by the European New Legislative Framework (NLF) and other harmonised European regulations (Annex II, Section A and Section B). Regulated areas include, e.g., automotive, fossil fuels and medical devices.
- AIs deployed in (a) biometric identification (when this is not forbidden) (b) management and operation of critical infrastructure, (c) education and vocational training, (d) employment, worker management and access to self-employment, (e) access to and enjoyment of essential private services and public services and benefits (e.g., healthcare), (f) law enforcement (g) migration, asylum and border control management, (h) administration of justice and democratic processes (Annex III).

Limited risk includes (AIA, Title IV):

- AIs that interact with natural persons, e.g., chatbots, when this is not obvious from the circumstances and the context of use or is not permitted by law to detect, prevent and investigate criminal offences.
- AIs that generate or manipulate images, audio, or video to simulate people, objects, places or other existing entities or events (i.e., deep fakes).

Minimal risk includes (AIA, Title IX):

- Residual AIs, some examples are AIs for video games or spam filters.

AIs posing unacceptable risks fall into the ALARP 'Intolerable' risk range, i.e., situations whose risk cannot be justified except in extraordinary circumstances. Under the AIA, specific exempt circumstances, like terrorist attacks, allow the time-limited use of AIs for remote biometric identification in publicly accessible spaces for law enforcement (Article 5(d)).

AIs posing high and limited risks fall into the 'Tolerable' risk range. That is where the ALARP principle comes fully into play: risk is tolerated only if all reasonably practicable mitigation measures are implemented. However, what counts as 'reasonably practicable' might be tricky to determine. A predominant interpretation is that: 'Efforts to reduce risk should be continued until the incremental sacrifice is grossly disproportionate to the value of the incremental risk reduction achieved. Incremental sacrifice is defined in terms of cost, time, effort, or other expenditures of resources'.[24]

This judgement should therefore consider the expected utility of risk containment. In the AIA, reasonable efforts consist of the legal requirements and guarantee mechanisms that providers (and deployers) must comply with to place high-risk AIs on the single market (Article 6 et seq.). We shall analyse the ALARP principle, seeking to improve its enforcement in the AIA, in greater detail in Section 5.

AIs posing minimal risks fall into the ALARP 'Broadly Accepted' risk range. In these cases, the risk is tolerable enough that no specific intervention is required, except

[24] Baybutt P. The ALARP principle in process safety. *Process Safety Progress* 2014;33 (1):1.

to ensure compliance with good practices. This is also what the AIA prescribes by encouraging the adoption of voluntary codes of conduct either by individual providers of AIs or by their representative organisations (Article 69).

Much of the legal framework concerns high-risk AIs, prescribing conformity assessment procedures, technical documentation, and certification duties to place them on the market (e.g., Article 43). Sometimes these safeguards involve post-market monitoring (e.g., Article 61). The other three risk categories produce fewer and simpler regulatory burdens: AIs that pose unacceptable risks are prohibited (Article 5), those that pose limited risk trigger a general transparency obligation (Article 52), while for those that pose minimal risks the AIA fosters voluntary codes of conduct (Article 69).[25] An exception to these rules is provided in the AIA insofar as it requires the Member States to introduce regulatory sandboxes: controlled environments in which AIs can be developed and tested for a limited time, before putting them on the market, prioritising small providers and start-ups (Article 53 seq.).

3. Strengths and weaknesses of the AIA's risk regulation

The supranational legislator expects the regulation of AI to increase legal certainty in this field and to promote a well-functioning internal market: reliable for consumers, attractive for investment, and technologically innovative.[26] This might trigger the Brussels effect, ensuring a competitive advantage over other international policy-makers while shaping their regulatory standards.[27] Nevertheless, should the AIA prove to be unsustainable or ineffective, the EU may lose its attractiveness for the production and commercialisation of AI technologies. To prevent this, the AIA must introduce norms that promote safety while not disincentivising the production or deployment of AIs.[28] In this regard, the AIA's risk-based approach has its strengths and weaknesses. Let us start with the strengths.

First, risk-based regulations rationalise governance interventions by setting their priorities and objectives. Well-delineated priorities and objectives facilitate accountability mechanisms towards the policy-maker.[29] In this respect, the AIA declares its priorities and objectives: the protection of the fundamental values and rights of the Union and the development of the AI market.

Second, risk-based regulations facilitate the fair distribution of resources (e.g., for supervision and certification) and costs. For example, costs are distributed according to the specific risks posed to a target community, and they are so transparently, as the

[25] Codes of conduct can be created by individual providers or their representative organisations.

[26] These are explicitly stated objectives of the AIA draft (p. 3).

[27] Bradford A. The Brussels Effect: How the European Union Rules the World. Faculty Books 2020.

[28] Of course, other factors will determine the success of the European AI strategy, like taxation and administrative efficiency. However, in this paper, we will only address the regulatory framework, namely the risk-categorisation of the AIA.

[29] Black J. The role of risk in regulatory processes. In: Baldwin R, Cave M, Lodge M, curatori. New York, USA: Oxford University Press 2010; 302–48.

criteria for distributing resources and costs are made evident in the regulation.[30] As the compliance cost is proportional to the risk, AIA introduces a kind of Pigouvian tax on the negative externalities of high-risk AIs.[31] To be acceptable, the AIA should allocate costs and resources efficiently among market players. However, the AIA does not consistently distribute resources in the best possible way, as we shall see when discussing its weaknesses.

Third, risk-based regulations cope with the uncertainty of phenomena – i.e., "when there is a lack of knowledge in qualitative or quantitative terms"[32, 33] – for example, by qualifying predictions about the occurrence of specific hazards probabilistically.[34] Moreover, risk-based regulations adapt to the political context or technological and market changes.[35, 36] In this regard, the AIA offers the possibility of updating its list of risky AIs at Articles 84-85. Unfortunately, the current version allows new AIs to be added only if they fall within the already established scopes. For this reason, some suggestions have been made to include reviewable risk categorisation criteria.[37]

The main limitation of the AIA is the lack of reviewable criteria for risk categorisation, which depends instead on the broad scopes of AIs. This threaten the effective enforcement of the AIA.[38] Providers may be discouraged from complying with EU norms due to the lack of reviewability of the AIA's requirements. In particular, the formalist approach of the draft precludes adapting risk categorisation to the interplay of hazard sources, vulnerability profiles of the exposed community, or values and interests at stake. No doubt, the model enshrined in the AIA heavily relies on a fundamental rights-based approach – as confirmed by the amendment introducing a fundamental rights impact assessment (AIA, Article 29a) – which characterizes the entire structure of the legislative proposal and, more broadly, the

---

[30] lack J. Risk-based Regulation: Choices, Practices and Lessons Being Learnt. Paris: OECD 2010; 185–224. https://www.oecd-ilibrary.org/governance/risk-and-regulatory-policy/risk-based-regulation_9789264082939-11-en

[31] Baumol WJ. On Taxation and the Control of Externalities. *The American Economic Review* 1972; 62(3):307–22.

[32] van der Heijden J. Risk as an Approach to Regulatory Governance: An Evidence Synthesis and Research Agenda. *SAGE Open* 2021;11(3):215.

[33] Sometimes, the concepts of risk and uncertainty are kept separate, the former being considered calculable and the latter not. For this purpose, the distinction between *epistemic* and *aleatory* uncertainty may be relevant, with only the latter being effectively addressable through risk assessment. On this, see Renn O. Risk Governance: Coping with Uncertainty in a Complex World. London: Routledge 2011; 368.

[34] Rothstein H, Borraz O, Huber M. Risk and the limits of governance: Exploring varied patterns of risk-based governance across Europe. *Regulation & Governance* 2013; 7(2):215–35.

[35] Black J, Baldwin R. Really Responsive Risk-Based Regulation. Law & Policy. 2010;32(2):181–213.

[36] At the same time, excessive uncertainty must be seen as a limitation of any risk model.

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

[38] The AIA is part of a complex body of European regulations on AI, including the Digital Service Act, the Data Governance Act, and the Digital Markets Act.

most recent pieces of legislation adopted at EU level in the digital context.[39] However, as legal compliance always comes at a cost,[40] if there is no possibility to ease regulatory burdens by a proportionality assessment, then the AIA might become unsustainable for AIs providers or deployers. This would be a severe loss for the EU AI strategy, disincentivising innovation and losing the benefits AI technologies can bring to those values the AIA aims to protect. The May 2023 amendment significantly advances the regulation by allowing revisions to high-risk system classifications based on an assessment of the risk's significance, i.e., its probability, severity, intensity, and potential population impact (AIA, Recital 32). Though this aligns with our argument, it may not integrate smoothly with the existing regulation, which was originally designed with a different risk approach and lacks a clear methodology for determining significant risk.

4. Removing the model flaw: the IPCC framework for risk assessment

The model flaw results from an insufficiently granular risk assessment model: the relevant factors of AI risk are not accurately identified and/or combined.

As argued in Section 2, the AIA's risk model is inspired by the ALARP principle and considers mainly two risk factors (a) the inherent risk of AI technology and (b) a value asset consisting of fundamental principles and rights of the Union. The EU legislator prescribes risk mitigation measures proportionate to the risk magnitude. As a result, risk management measures are allocated according to the four risk categories of the AIA.

Hence, the risk considered in the AIA is legal in nature, expressing the potential detriment that comes from the violation of a legal norm by an AIs (i.e., principles and rules).[41, 42] However, the AIA's risk assessment model does not fulfil the distinctive nature of the legal risk as it does not evaluate comparatively and proportionately the specific weight of legal norms. Quite the opposite, risk assessment in the AIA is modelled as a neutral tool that treats legal norms as technical standards which are either met or not.[43] Consequently, the risk is categorised through a formalist list of AI scopes potentially detrimental to fundamental principles and rights. However, risk assessment is not a neutral tool: it reflects the risk appetite of a specific community, weighing the costs and benefits of risk mitigation,[44] balancing the interests and values of that community, and all this dynamically and diachronically; while promoting a legal value,

---

[39] Ufert F. AI Regulation Through the Lens of Fundamental Rights: How Well Does the GDPR Address the Challenges Posed by AI? *European Papers - A Journal on Law and Integration* 2020; 5(2):1087–97.

[40] Khanna VS. Compliance as Costs and Benefits. In: van Rooij B, Sokol DD, curatori. The Cambridge Handbook of Compliance. Cambridge: Cambridge University Press 2021; 13–26.

[41] Mahler T. Defining Legal Risk. Rochester, NY; 2007. Available on: https://papers.ssrn.com/abstract=1014364.

[42] This is at least one of the meanings that the concept of legal risk can take, and it is the one associated with the Basel Committee on Banking Supervision's definition: "Legal risk includes, but is not limited to, exposure to fines, penalties, or punitive damages resulting from supervisory actions, as well as private settlements".

[43] Smuha N, Ahmed-Rengers E, Harkens A, Li W, Maclaren J, Piselli R, et al. (n 35).

[44] Krebs JR. Risk, uncertainty and regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2011;369(1956):4842–52.

it may be the case that the unexpected demotion occurs of other equally fundamental legal values. Accordingly, risk management measures should be modulated according to the outcome of such a balancing process. This is precisely what is missing from the AIA, which, despite claiming to be informed by the trade-off between economic development interest and the protection of fundamental rights,[45] seems to predetermine the proportionality judgment that settles the interference between values. Also significantly, not only the list of fundamental rights protected by the proposal is particularly rich, but it also includes interrelated rights,[46] making thus difficult a horizontal balance between competing fundamental rights.

The model flaw does not concern only the lack of granularity in the analysis of values and rights. The draft also lacks an accurate representation of the hazards' sources of AIs, of what makes people vulnerable to these hazards, and of whether hazards and vulnerabilities are mitigated by mechanisms, including legal ones, that already exist (i.e., the net risk).[47]

Against this background, the May 2023 compromise text's requirement for providers of high-risk AI systems to conduct a fundamental rights impact assessment before market introduction is a progressive move. The methodology we propose in this section aims to enhance the accuracy of the proposed assessment outlined by the EU policy-maker.

To improve the implementation of the AIA,[48] we propose a risk assessment model that includes multiple risk factors, and their interferences, and provides a proportionality judgement to review risk categories. This, however, without dismantling or multiplying the draft's tolerance ranges. On the contrary, we suggest applying the four risk categories horizontally to each of the AIs listed in the AIA, so that under varying conditions – e.g., a specific interference among fundamental rights involved – the same system can be treated as unacceptable, high-risk, limited-risk or minimal-risk. This implies that risk categories would not depend by default on AI scopes, but on the concrete risk scenarios associated with the application of AI systems due to the incidence and combination of multiple risk factors. Alternatively, to be consistent with the proposed frameworks of the AIA, we suggest that risk categories should be more granular, maybe providing subcategories, or allow confirmatory or reevaluative assessments (as the compromise text of the AIA implies).

To build risk scenarios, the Intergovernmental Panel on Climate Change (IPCC) provides a multifaceted risk assessment model, which has then be refined the subsequent literature[49] and which we can use to assess risks of AIs.

---

[45] This is clearly stated in the Explanatory Memorandum of the Proposal: "To achieve those objectives, this proposal presents a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market".

[46] Cf. European Commission, Explanatory Memorandum to AIA, para 3.5.

[47] Black J, Baldwin R. When risk-based regulation aims low: Approaches and challenges. *Regulation & Governance* 2012;6(1):2–22.

[48] Of course, this also presupposes changing the risk assessment and its metrics.

[49] Simpson NP, et al. (n 6).

The IPCC has often conceived the climate change risks – e.g., disaster risk – as the consequence of three determinants: hazard (H), exposure (E), and vulnerability (V).[50] Broadly speaking, hazard refers to the sources of potential adverse effects on exposed elements; exposure refers to the inventory of elements within the range of the hazard source; vulnerability refers to the set of attributes or circumstances that makes exposed elements susceptible to adverse effects when they impact the hazard source.[51, 52] The IPCC's approach can be developed further, as in the framework for climate change risk assessment proposed by Simpson et al., 2021, which evaluates risk at a lower level of abstraction by including the individual components of the risk determinants, i.e., the drivers. Simpson et al. expand the IPCC approach by incorporating a fourth risk determinant – the response (R) – and contextualise risk assessment by including multiple types of risk with their own determinants. Thus, according to their framework, the overall risk results from the interaction among (1) determinants, (2) drivers, and (3) risk types (

Figure 2). These three sets of relations occur at stages of increasing complexity. The AIA only considers the lowest complexity stage, where the relevant risk factors are the determinants taken statically, that is, overlooking interactions among their drivers (or with cross-sectorial 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 (i) aggregate, if drivers emerge independently of each other but jointly influence the overall risk assessment; (ii) compounding, if drivers produce a specific effect on risk assessment when combined, unidirectionally or bi-directionally; (iii) cascading, when drivers trigger others which themselves may produce further drivers in a cascading process. The same applies to interactions between multiple risk types.[53]

Figure 2 below shows the three sets of interactions.

[50] This conceptual approach is clearly set out in Cardona OD, Aalst MKV, Birkmann J, Fordham M, Gregor GM, Rosa P, et al. 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 2012; 65–108. This approach also emerges in special IPCC reports, e.g., Special Report on Climate Change and Land — IPCC site 2019 https://www.ipcc.ch/srccl/ and Special Report on the Ocean and Cryosphere in a Changing Climate 2018 https://www.ipcc.ch/srocc/.

[51] Cardona OD, Aalst MKV, Birkmann J, Fordham M, Gregor GM, Rosa P, et al. (n 48).

[52] That hazard, exposure, and vulnerability are relevant to risk assessment is also widely believed in the literature other than climate change, such as in Renn O. (n 16). In studies on global environmental change and sustainability, the same four determinants were considered as parts of a risk sequence chain. Turner BL, Kasperson RE, Matson PA, McCarthy JJ, Corell RW, Christensen L, et al. A framework for vulnerability analysis in sustainability science. *Proceedings of the National Academy of Sciences* 2003; 100(14):8074–9.

[53] Simpson NP, et al. (n 6).

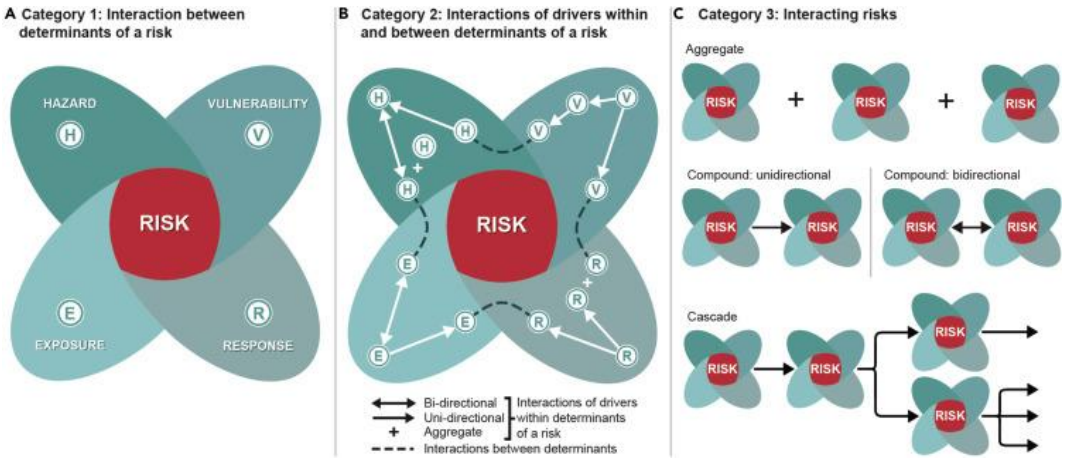


Figure 2. Three categories of increasingly complex climate change risk by (Simpson et al., 2021)

In climate change, the drivers of the hazard (H) can be natural or human-induced events. In AI, these drivers may be either purely technological 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 interact in an aggregate way. Interactions are compounded when, e.g., low data representativeness compounds with overfitted machine learning models or biased data. The interaction between drivers 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.[54]

Drivers of exposure (E) in climate change risk may be people, infrastructure, and other social or economic assets. For AI risk, exposure drivers may be tangible assets, like goods or environment, or intangible assets, like values and rights. As already stressed, the exposed asset of the AIA mainly consists of fundamental rights and values, such as health, safety, employment, asylum, education, justice, and equality. Interactions between drivers of exposure may be aggregated if, e.g., an AIs has adverse effects on the right to asylum and the privacy of asylum seekers. It is compounded when, e.g., an AI's adverse effect on the environment compounds with those on health. The interaction between drivers of exposure is cascading when, e.g., an AI's adverse effect threatens access to education, and thus equality and democratic legitimacy (and so on).

Vulnerability (V) drivers of climate change risk may concern the propensity to suffer adverse effects of communities – e.g., poverty – and infrastructure – e.g., lack of flood containment. Drivers of vulnerability in AI risks are multiple and overlapping, e.g., income, education, gender, ethnicity, health status, and age. The lack of appropriate control bodies, procedures, or policies should be included among the drivers of vulnerability for AI risk. The AIA shows two conceptions of vulnerability:

[54] Black J, Baldwin R. (n 27).

a generic one, whereby the mere entitlement to fundamental rights entails the propensity to suffer adverse effects of hazards; and a more specific one, whereby all those AIs that "[…] exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability", (AIA, Article 5) should be banned. In the latter case, the list of vulnerability drivers is rather poor.

The interaction between vulnerability drivers is aggregated when, e.g., an AIs is deployed in a vulnerable environment, and there are few surveillance or feedback mechanisms. The compounding interaction is perhaps the most interesting one, as an intersectional reading of vulnerabilities can also be advocated in AI risk: ethnicity, gender, health, age, education, economic status, and other characteristics are profiles of vulnerability that have to be considered in the way they intersect and influence each other. The vulnerability would be "the result of different and interdependent societal stratification processes that result in multiple dimensions of marginalisation".[55] In this sense, the intersectional approach to vulnerability is a risk management principle that enables policy-makers to identify the most appropriate measures to counter hazards to individuals and groups. These interactions make vulnerability a multi-layered condition.[56] The interaction between vulnerability drivers is cascading when, e.g., the absence of AIs liability rules triggers several other vulnerabilities for those under the adverse effect of AIs use.[57]

The analysis by Simpson et al. introduces a fourth determinant, i.e., response (R), which concerns existing measures that counteract or mitigate risk. The response indicates the environment's resilience to a specific risk and includes governance mechanisms. Regarding AI risk, the response drivers can be institutional safeguards on the development, design, and deployment of AIs or data quality rules. Consequently, risk assessment and categorisation within the AIA should consider already existing legal measures to avoid the adverse effects of AI technologies, e.g., those contained in the GDPR.[58]

Adaptation and mitigation responses may increase or decrease the risk level of specific AIs. As a result, the response determinant can be used to discriminate intrinsic from net risk, the latter adjusted to risk management measures:

> "[…] where the potential harm is higher than for the intrinsically lower risks, but the probability and/or impact is reduced by risk management and other control measures, or by systems of resilience – such as capital requirements in financial

[55] Kuran CHA, Morsut C, Kruke BI, Krüger M, Segnestam L, Orru K, et al. Vulnerability and vulnerable groups from an intersectionality perspective. *International Journal of Disaster Risk Reduction* 2020; 50:101826.

[56] Luna F. Identifying and evaluating layers of vulnerability - a way forward. *Dev World Bioeth* 2019; 19(2):86–95.

[57] In the proposed fundamental rights impact assessment, interest in vulnerability is emphasized: "This assessment should include […] (f) specific risks of harm likely to impact marginalised persons or vulnerable groups" (Article 29a, (f)).

[58] Consider, for example, art. 35 on data protection impact assessment: "1. Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data."

institutions, or engineered safety controls in power stations, or by the possibility of remediation"[59]

Simpson et al. also introduce a third stage of interaction, between climate change risk and other types of risk, which are extrinsic to it and have their own determinants. Risk types that interact with AI risk may be, e.g., market, liability, and infrastructure risks. Some of these risk types are created by the AI risk itself – i.e., cascading interactions – others are independent but may affect the overall assessment of AI risk – i.e., aggregate or compounded interactions (
Figure 2). For instance, an aggregate interaction occurs between AI risk and policy risk, in the sense that 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 can cascade into multiple other types of risk, the risk to innovation, to digital sovereignty, to economic sustainability, to power concentration, and so forth.

This aspect should be linked to that of ancillary risks, i.e., risks 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 potential barriers to technological innovation that the ban raises, and the risk posed by the systems replacing the banned ones.[60] 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.[61]

To sum up, the risk magnitude of each AIs listed in the AIA should be assessed in terms of the interactions among determinants, drivers, and other risk types. Although AIA considers some interactions among determinants – e.g., the scale and the likelihood of adverse effects on values – it does not account for the interaction among the individual drivers of those determinants, nor does it evaluate the risk of AIs in relation to other types of risk. Therefore, the AIA misestimates the AI risk magnitude and anchors risk categories to static, coarse-grained factors.

Once the determinants, drivers and external types of risk are identified, adaptation and mitigation become easier, i.e., to reduce the risk of AI by planning actions (including policies) that address the factors of hazards, exposure and vulnerability.[62]

The granular risk assessment we propose has a higher degree of variability. The risk categories of the AIA become risk scenarios[63], which change depending on the interactions among risk factors. This leads to a more accurate representation of the risk magnitude – i.e., the likelihood of detriment and severity of consequences on values – with connections among risk factors being made explicit. Even if the EU legislator intends to keep the current framework – where risk categories are pre-determined based on the AI's scopes – this model can aid in the proposed additional

[59] Black J, Baldwin R. (n 27), 5.
[60] Sunstein CR. Risk and Reason. Cambridge Books. Cambridge University Press 2004.
[61] Karliuk M. Proportionality principle for the ethics of artificial intelligence. *AI Ethics* 2022.
[62] Simpson NP, et al. (n 6).
[63] Renn O. (n31), 368.

assessments that could revise the risk categorization (i.e., risk significance). However, what we have presented in this section, is just a general framework. While risk magnitudes may correspond in the abstract to risk categories, as a preliminary evaluation, this assignment also must pass the proportionality test that we shall describe in the next section.

5. A quantitative basis for the model: the proportionality test

Though not directly mentioned in the AIA, an issue shared with the ALARP principle is setting risk management measures, without defining what qualifies as a "grossly disproportionate" containment measure.

A way to offer quantitative support for ALARP-based legislative choices is through the traditional cost-benefit analysis (CBA).[64] While the ALARP allows the costs of risk mitigation to exceed the benefits as long as they are not exorbitant, the CBA specifies that intervention is justified only if costs are less than or equal to the benefits. CBA does not account for uncertain costs and benefits.[65] Despite this drawback, CBA can support ALARP as a preliminary informational input: as far as possible, CBA quantifies known costs and benefits so that this information can be combined with a qualitative assessment of what is "reasonably practicable".[66] The risk assessment model presented in the previous section helps us to combine CBA, the ALARP principle, and the AIA to account for the likelihood and distribution of adverse effects, the causal chain between hazards and harms, the effects of AI risk regulation (i.e., ancillary risks), and alternative measures for risk mitigation. However, CBA remains an imperfect tool for the AIA, as the former expresses the value of things with a single numerical parameter, usually market prices, while the latter concerns a legal risk, whose exposed asset consists of fundamental rights and values, which, respectively, are intended to represent "principles of [EU] law of a constitutional nature"[67] and the "very identity" of the EU legal order.[68]

However, we suggest an alternative quantitative approach to ascertain when sacrifices to mitigate risk are "grossly disproportionate" (within the scope of the AIA). This quantitative assessment should be seen as complementary – a second step – to the risk assessment model of the previous section: to assign the appropriate risk category for a specific scenario, we need to compare the impact each risk category has on the assets served by the intended scope of the AIs (P_x) – e.g., law enforcement – against those of the exposed asset (P_y) – e.g., safety, health, and equality. Thus, if applying the high-risk category to AIs for law enforcement under a specific risk scenario has a sub-optimal impact on the joint realisation of principles and rights, it is

---

[64] French S, Bedford T, Atherton E. Supporting ALARP decision making by cost benefit analysis and multiattribute utility theory. *Journal of Risk Research* 2005;8(3):207–23.

[65] Jones-Lee M, Aven T. (n 12).

[66] Ale BJM, Hartford DND, Slater D. ALARP and CBA all in the same game. *Safety Science* 2015; 76:90–100; French S, Bedford T, Atherton E. (n 61).

[67] Joined cases C-402/05 P and C-415/05 P *Kadi and Al Barakaat International Foundation* EU:C:2008:461, para 276.

[68] Case C-156/21 *Hungary v European Parliament and Council of the European Union* EU:C:2022:97, para 127

desirable to opt for an alternative risk category. Whereas, if the marginal gains to law enforcement outweigh the marginal harms to other rights, then the risk category is justified.

Robert Alexy proposed a well-known method in legal theory to quantify this type of choices.[69] According to this approach, a legal norm that interferes with fundamental values[70] is legitimate when it meets a proportionality test characterised by the following optimisation principles:

- *Suitability*, which "excludes the adoption of means obstructing the realisation of at least one principle without promoting any principle or goal for which they were adopted".[71] In the AIA, the legislative choice of assigning a risk category R_1 to an AI that negatively impacts one principle P_2 is suitable if it impacts positively another principle P_1.
- *Necessity*, which "requires that of two means promoting P_1 that are, broadly speaking, equally suitable, the one that interferes less intensively in P_2 ought to be chosen".[72] In other words, R_1 with a negative impact on P_2 is necessary if it has a positive impact on P_1 and there is no alternative, R_2, having a higher positive impact on P_2 and non-inferior on P_1.[73] In the AIA, as in many other cases, Pareto-optimality equilibria are rather unstable: multiple values are involved, and a principle P_3 that is negatively interfered with by R_1 can easily occur. These unavoidable costs call, according to Alexy, for a third principle.
- *Proportionality in the narrow sense*, which states that "The greater the degree of non-satisfaction of, or detriment to, one principle, the greater the importance of satisfying the other".[74] This principle provides a basis for determining whether or not the importance of satisfying P_1 with R_1 justifies the impairment or failure to satisfy P_2. When multiple values are involved, as in the AIA, we will say that R_1 with a negative impact on P_2 is balanced if there is no alternative R_2 having a lower negative impact on P_2 and a higher overall utility on P_3, P_4...P_n.[75]

Such a proportionality test, which is (by and large) in line with the proportionality test the EU Court of Justice applies while balancing competing rights and values,[76, 77] may

---

[69] Alexy R. A Theory of Constitutional Rights. Oxford, New York: Oxford University Press 2002.

[70] In Alexy's theory, these fundamental values are typically constitutional principles. Alexy R. Constitutional Rights, Balancing, and Rationality. *Ratio Juris* 2003;16(2):131–40.

[71] Alexy R. (n 67), 135.

[72] Alexy R. (n 67), 135.

[73] Sartor G. A Quantitative Approach to Proportionality. In: Aitken C, Amaya A, Ashley KD, Bagnoli C, Bongiovanni G, Brozek B, et al., eds. Handbook of Legal Reasoning and Argumentation. Springer Verlag 2018; p. 613–36.

[74] Alexy R. (n 66), 102.

[75] Sartor G. (n 70).

[76] Tridimas T. The Principle of Proportionality. In: Schütze R, Tridimas T, curatori. Oxford Principles Of European Union Law: The European Union Legal Order: Volume I. Oxford University Press 2018; 243–264.

[77] See also art. 52, para 1, of the EU Charter of Fundamental Rights, stating that, "[s]ubject to the principle of proportionality, limitations may be made only if they are necessary and genuinely meet objectives of general interest recognised by the Union or the need to protect the rights and freedoms of others."

support legislative choices and trade-offs within the AIA, i.e., the exposed asset of AI risk. In particular, we suggest that it may serve to justify trade-offs between fundamental values/rights that (should) inform the risk categorisation of AIs. The outcome of the test may warrant the ascription of a risk category R_1 (e.g., high-risk) to specific AIs or shifting an AI to a new category R_2 (e.g., minimal risk). For this purpose, proportionality in the narrow sense should be broken down into three evaluations:

(1) the intensity of interference (I_x), the degree of non-satisfaction or detriment to a principle P_x to the benefit of a competing one P_y

(2) the concrete importance (C_y) of satisfying P_y

(3) the concrete weight of P_x ($W_{x,y}$), namely the ratio between I_x and C_y, which determines whether the importance of satisfying P_y justifies the non-satisfaction or detriment to P_x .[78]

Finally, the abstract weights of P_x (W_x) and P_y (W_y) also play a role in the overall balance.[79]

In some cases, P_x will prevail over P_y, e.g., when I_x is severe, and C_y is weak. In other cases, P_y will prevail over P_x. There may also be cases where there is no prevalence between P_x and P_y, $I_x = C_y$, which creates deadlocks, increasing discretion in balancing. The outcome of the ratio between the intensity of the interference on a specific principle and the concrete importance of the competing one is expressed by the following, simplified version, of the weight formula:[80]

$$W_{x,y} = \frac{I_x \cdot W_x}{C_y \cdot W_y}$$

Applying the weight formula to the AIA, I_x would correspond to the degree of interference a risk category, with its containment measures, has on a (set of) value(s) served by the intended scope of AIs: e.g., the interference to public safety (P_x) as served by biometric categorisation systems. C_y would correspond to the concrete importance of satisfying a competing (set of) value(s) explicitly protected by the AIA, which is part of risk-exposed asset in biometric categorisation systems: e.g., the right to privacy (P_y). The concrete importance expressed in C_y depends on qualitative assessments in relation to the risk scenario, i.e., what are the hazard factors, vulnerability profiles and response mechanisms that determine the magnitude of risk in the concrete scenario (as described in the framework shown in the previous section). Therefore, whether the EU legislator is authorised to restrict the use of AIs for biometric categorisation will depend on whether the magnitude of the privacy risk posed by these systems (C_y) justify the impairment of public safety caused by the measures of the relevant risk category (I_x).

---

[78] Alexy R. On Balancing and Subsumption. A Structural Comparison. Ratio Juris. 2003;16(4): 433–49.

[79] Alexy also includes another variable in his weight formula, namely the epistemic reliability of the balancing premises. For simplicity of exposition, we will not consider them here.

[80] Alexy R. (n 75).

Although the weight formula relies on non-numerical premises – like judgments about the degree of interference of a risk category or the abstract weight of principles[81]– numerical values can still be assigned to I_x and C_y. This can be done using a geometric sequence, like 2^0, 2^1, 2^2, 2^4, to assign numerical ranges to the AIA's four risk categories according to the degree of interference, or non-satisfaction, they cause to the intended scope of an AIs (I_x): unacceptable risk=16, high-risk=4, limited risk=2, minimal risk=1. The same numerical ranges may be assigned to the importance of satisfying the competing principle – (C_y): major = 16, severe = 4, moderate = 2, light = 1[82] – and to the abstract weights of principles (W_x and W_y). As shown below, where the asset served by the intended scope of an AIs prevails over the exposed asset, the concrete weight $W_{x,y}$ will be greater than 1. Conversely, $W_{x,y}$ will be less than 1.

A) $I_x \cdot W_x\,(16 \cdot 4)/C_y \cdot W_y\,(8 \cdot 2) = 4$
B) $I_x \cdot W_x\,(4 \cdot 4)/C_y \cdot W_y\,(8 \cdot 16) = 1/8$

This quotient describes the concrete weight of the asset served by the intended scope of an AIs given the interference of a risk category on it (I_x) and a competing asset protected for being partly exposed to the AIs (C_y). The inclusion of the vulnerability and response determinants' values in the ratio can make the proportionality test fully aligned with the risk assessment model outlined in Section 4.

To sum up, the quotient of the weight formula is a quantitative criterion to assess whether the risk control measures are "grossly disproportionate" in the AIA, given the balance of relevant values, and therefore whether a risk category is suitable for the risk scenario of an AIs or whether it should be changed. In particular, what is grossly disproportionate can be quantified over a range. In our example, according to the numerical parameters we employed, it is reasonable to argue that the quotient of the weight formula should not be less than 1 or greater than 4. If it falls outside this range, then the balancing between principles is disproportionate and it is advisable to alter the risk category. Indeed, out of the range, a specific risk category may be inadequate for the risk scenario, with measures too stringent or too soft to balance competing EU values, like privacy and technological innovation. In this way, the AIA fails to achieve one of its main objectives: a uniform protection of EU fundamental rights.

We are aware that compulsory numerical values of EU principles and fundamental rights cannot be pre-assigned. Also, attempts to establish a strict hierarchy among EU fundamental values and rights have so far failed. While acknowledging the importance of these circumstances, we believe that a quantitative method for assessing risk containment measures could help relevant actors make policy decisions and avoid significant imbalances when implementing the AIA. Numerical values have been assigned to the coefficients in the proportionality test to enhance clarity but these

[81] Alexy R. (n 75).

[82] This requires assuming that the abstract weights have the same impact on the concrete weight as the intensity of interference.

coefficients can also be compared through non-numerical preferences or magnitudes, such as the Paretian superiority illustrated in.[83]

On a different note, we cannot ignore the role that EU institutions – and, in particular, the role that the EU Court of Justice – shall play in preserving the constitutional framework of the Union and the untouchable core of the EU legal order, which include its fundamental values and rights. This is why in the next section, we shall discuss the allocation of competences and roles in scenario building and proportionality assessments.

6. Supranational and national risk assessment

Although the categorisation of risk in the AIA is coarse-grained, connecting risk management measures to broad scopes of AIs makes it easier to approve and monitor them for marketing. In contrast, a legal framework with risk scenarios built on multiple factors and tested by proportionality-based balancing, as the semiquantitative approach we are recommending, might complicate the procedures laid down in the AIA. Therefore, the construction of risk scenarios should be carried out by institutional bodies enforcing the regulation: under the current AIA framework, they might be the national supervisory authorities (AIA, Title VI).

Whilst this solution would cause a shift in the AIA's governance – the regulatory approach to high-risk AIs being now mainly established at supranational level – national authorities should not build risk scenarios or assess the proportionality of risk management measures just on their own, depriving thus the EU institutions of their roles. This would contradict the AIA's objective to provide a supranational risk assessment shared by all Member States. For this reason, it is crucial to determine the competences, functions and interactions of supranational institutions and national bodies in the risk assessment of AIs. In the light of the foregoing, and considering the shared nature of the competences exercised by the EU legislator to adopt the AIA,[84] it remains undisputed that the EU legislator should retain a primary role in shaping the risk-assessment model at stake.[85] Meanwhile the European Commission should keep its role of guardian of the AIA enforcement and the EU Court of Justice's authority in judging whether the risk assessment is consistent with the essential core of EU fundamental values.[86] This is particularly important considering the systematic backsliding on fundamental values and rights taking place in some EU countries.

---

[83] Sartor G. (n 70).

[84] As it is well-known the AIA proposal is based in the first place on Article 114 TFEU, providing a EU shared competence in adopting measures to ensure the establishment and proper functioning of the internal market. In addition, the proposal is based on Article 16 TFEU, due to its connection to the processing of personal data.

[85] On the fundamental role the EU legislator should play in this respect, see Fontanelli F. The Court of Justice of the European Union and the illusion of balancing in internet-related disputes. In O. Pollicino, & G. Romeo (eds.), The Internet and Constitutional Law: The protection of fundamental rights and constitutional adjudication in Europe. Routledge Research in Constitutional Law 2016; 94–118.

[86] Lenaerts K. Limits on Limitations: The Essence of Fundamental Rights in the EU. *German Law Journal* 2019; 20(6): 779–93.

More to the point, under our semi-quantitative model, the EU legislator could determine (a) the key drivers of the four risk determinants, (b) the extrinsic types of risk to account for and (c) the (abstract) weight of the principles involved in the proportionality test. These factors could be linked precisely to the scopes already identified in the AIA through risk categorization (e.g., Annex III). In the next section, we offer a case study that illustrates how key drivers of the four determinants and extrinsic risk types may be identified in connection to the scope of an AIs, i.e., justice (Section 7).

In this way, the scopes of AIs would still play a primary role in risk regulation – which means that the text of the AIA would not require substantial changes – and EU institutions would limit the discretion of Member States. National authorities would be responsible for assessing risk in particular cases – thereby enhancing their powers over what is in the AIA – through scenarios and proportionality tests.[87]

Detractors could claim that the proposed solution may lead to a partially diversified enforcement of the AIA within the EU, something which contradicts the idea of a uniform respect of the core essence of EU values and rights. However, this position does not consider that risk assessment, to be accurate, must be context-sensitive. Moreover, our semiquantitative approach does not necessarily weaken the effectiveness of the EU's fundamental values and rights as it is based on the idea of introducing a robust rational procedure, under the strict supervision of the European Commission and the ultimate control exercised by the EU Court of Justice.

For those who find the diversification of AIA enforcement problematic, the model we propose is only suboptimal; they would prefer to keep the risk assessment all at the supranational level or to identify a EU harmonization body. From our perspective, the optimal solution would probably be to revise part of the regulation according to the proposed model.

At the same time, the compromise text of the AIA, which now includes the evaluation of risk significance and the impact on fundamental rights, goes in the direction we advocated. It mandates these evaluations to be conducted by deployers but with the obligation to inform national supervisory authorities, relevant stakeholders, and representative groups of individuals who may be impacted by the application of the (high-risk) AI system (e.g., Recital 32 and Article 29a).

7. Contributions: enforcement and general purpose AI (GPAI)

Our analysis offers two contributions to the enforcement of the AIA and the regulation of general purpose AI (GPAI).

First, the risk assessment model shown in Section 4, supported by the quantitative proportionality test shown in Section 5, improves the enforcement of the AIA. It would provide risk management measures that are more appropriate to estimate and contain the dangers of AI, more specific for national regulators (and judges), more sustainable for AI providers, and ultimately more likely to achieve the AIA's goals of protecting all fundamental EU values involved. Ideally, such granular risk management

[87] Or any other type of proportionality-based balancing.

measures can help avert, or more effectively handle, issues related to the under-inclusiveness or over-inclusiveness of risk categories.[88]

To show how risk scenario building might work in the AIA, let us consider a case study, that of AIs used to assess the recidivism rate of natural persons. The semiquantitative approach consists of two stages: risk-scenario building and the proportionality test. The risk drivers here identified can be easily inferred from the AIA. Of course, applying our proposed assessment model during the AIA implementation stage would necessitate enhanced legislative transparency in setting the drivers and extrinsic risks.

Starting from the risk-scenario building, the four determinants of AI risk, the interaction among their drivers and with other risk types may be thus combined:

(a) Hazards. These drivers of an AI for recidivism rate assessment would be the inner opacity of the system and the poor quality or misuse of the training data. When these hazard drivers compound, they can lead to the AIs perpetrating discrimination biases. The greater these hazard drivers are, and the more likely they combine to produce such wrongdoing, the "heavier" the hazard determinant will be in the specific risk scenario. What is more, the hazards must be related to the vulnerability drivers of a specific environment in which AIs are deployed, not least because these will be inclined to replicate the social discriminations of the environment.

(b) Exposure. These drivers would be the fundamental values potentially affected by the use of an AI to assess the recidivism rate. This would involve some substantive legal principles – e.g., the principle of criminal culpability and of equality – and some procedural ones – e.g., the principle of transparency and the right of/to defence.[89] These drivers also interact with each other, and where they interfere, it is necessary to balance them to assess the overall weight of the exposure determinant. This also requires balancing those values that the use of AIs is intended to enhance (consistent with the proportionality test in section 5), such as the principle of predictability, legal certainty, safety and efficiency.

(c) Vulnerability. These drivers would be attributes that make individuals or groups susceptible to the adverse effects of automatic recidivism rate assessment: e.g., ethnicity, economic conditions, and education. When these drivers interact with each other, perhaps compounding or cascading, vulnerability should be treated as a multi-layered condition:[90] e.g., the compound of ethnicity and socio-economic conditions often leads to a heightened sensitivity to the biases of prediction systems. As mentioned above, drivers of vulnerability compound with hazard drivers: e.g., biases in the recidivism rate assessment will be greater where social discriminations are already in place.

---

[88] Hacker P. The European AI Liability Directives - Critique of a Half-Hearted Approach and Lessons for the Future. arXiv 2023. Available on: http://arxiv.org/abs/2211.13960.

[89] Garrett B, Monahan J. Judging Risk. *California Law Review* 2020;108(2):439–93.

[90] Luna F. (n 54).

(d) Response. These drivers would be measures that counter the hazards of automatic recidivism rate assessment. They might be governance measures, like standards for data quality and data collection, transparency, bias examination, and human oversight. A concrete solution is to exclude specific indicators that, while predicting some degree of social dangerousness, are directly or indirectly linked to ethnic or social background, e.g., the postal code.[91]

(e) Extrinsic risks. The risk of AIs for recidivism rate assessment would finally interact with extrinsic risk types. Some extrinsic risks, in this case, would be compliance risk, liability risk, and economic risk. Indeed, AI risk may be influenced by the lack of effective rules for the allocation of liabilities for adverse effects and may, in turn, cause or amplify economic risks in the AIs market. The overall risk should also be balanced with ancillary risks. In this case, such risks would be those to innovation, loss of opportunities and digital sovereignty. This means that the introduction of regulatory burdens, or entry barriers, on AIs' providers may weaken technological innovation and, in the case of a radical ban, resulting in the loss of opportunity for the general social interest.

The interactions among these risk factors determine the two input variables of the overall risk magnitude of the specific scenario: (1) the likelihood of the event depend on the interaction between hazard drivers and response drivers (e.g., preventive measures); (2) likewise, the severity of the detriment can be higher or lower depending on the hazard sources, exposed asset, and vulnerability profiles.[92] As a result, risk magnitude is associated with the four risk categories of the AIA – i.e., unacceptable (U), high (H), limited (L) and minimal (M) risk – as illustrated in the risk matrix below[93]:

[91] van Dijck G. Predicting Recidivism Risk Meets AI Act. *Eur J Crim Policy Res* 2022; 28(3):407–23.

[92] These are the same input variables of the conception of risk magnitude embraced by the AIA (e.g., Title III, art. 7).

[93] The risk matrix approach is widespread in semi-quantitative risk assessments, such as the one we are suggesting. See, for example, Ni H, Chen A, Chen N. Some extensions on risk matrix approach. *Safety Science* 2010; 48(10):1269–78.

Severity	**Major**	L	H	U	U	U
	Serious	M	H	H	U	U
	Moderate	M	L	H	H	H
	Light	M	M	L	H	H
	Negligible	M	M	M	L	L
		0 – 0.20	**0.20 – 0.40**	**0.40 – 0.60**	**0.60 – 080**	**0.80 - 1**
	***Likelihood* (%)**					

Table 1. Risk matrix inspired by Ni H, Chen A, Chen N. (n 91)

The five levels of severity are described qualitatively, and those of likelihood in percentages in a range between 0 and 1 (where 0.20 – 1 is remote risk, while 0.80 – 1 the risk is almost certain). Under this matrix, the intersection of the input variables correlates with one of the four risk categories of the AIA.[94]

The second step is to evaluate the suitability of the resulting risk category in relation to the asset exposed to the use of an AIs, by means of the proportionality test. Let us assume that the risk magnitude for a specific recidivism rate assessment system matches its current categorization in the AIA, i.e., unacceptable risk (U). One of the principles served by AIs for assessing recidivism rates is safety (P_x) and, according to the geometric sequence seen in Section 5, its abstract weight can be quantified with a score of 4 (W_x). The degree of interference (I_x) of the AIA's high-risk category on legal certainty is 16. In the denominator of the Weight Formula, the abstract weight of a competing principle, e.g., criminal culpability (P_y), might be 4 (W_y) as well as the concrete importance of satisfying it (C_y). Applying all these values to the ratio – $W_{x,y} = (I_x \cdot W_x)/(C_y \cdot W_y)$ – the outcome would be 4, which falls within the proportionality range we have assumed. As a result, we might conclude that the risk category is appropriate as it correctly balances the values involved. Of course, if the competing principle was deemed to be less significant, for instance, it held a light value such as 2, then the outcome of the equation might not fall within the range and the risk category should be revised.

The second contribution of the risk assessment model presented here concerns one of the regulatory issues that emerged from the debate on the AIA: the governance of general purpose AI (GPAI). The issue was raised in an amendment proposing a definition of GPAI and classifying them as high-risk systems.[95] GPAIs are systems that

[94] For example, someone else might think it more correct that a moderate detriment with a probability between 0.20 and 0.40 percent should correspond to a high-risk category.

[95] GPAIs were excluded from the previous draft of the AIA. However, they are given more room in the compromise text, as implementations of foundation models, and are no longer equate with high-risk systems. They must still adhere to certain documentation and transparency rules. For instance,

can be deployed in multiple fields and with different tasks, some of which were unintended by the developers (e.g., foundation and generative models).[96] This definition would also include open-source AI models (e.g., open-source datasets).

Indeed, if the intended purposes are not foreseeable, neither are the fundamental values that AIs would affect and based on which their risk would be categorised. This implies that the application of the AIA would be even more static than for AIs with intended purposes. Therefore, the construction of risk scenarios based on determinants, drivers and types seem the only way to categorise and regulate GPAI in a granular manner and avoid treating them all the same. Given these AI technologies' success on the market, undifferentiated regulatory treatment might negatively impact AI industry innovation.

The semi-quantitative model outlined in this article would facilitate risk assessment and categorisation for all those situations that the AIA leaves uncovered, for example, where it recognizes the discretion of the European Commission in updating or modifying the list (and to remove use-cases) of high-risk AIs provided that:

> "[…] (b) the AI systems pose a significant risk of harm to health and safety, or an adverse impact on fundamental rights, to the environment, or to democracy and the rule of law, and that risk is, in respect of its severity and probability of occurrence, equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in Annex III." (AIA, Article7)

However, to determine whether the risk associated with new AI systems matches or exceeds those already classified as high-risk, thus justifying their addition to the high-risk list, a robust and transparent risk assessment methodology is necessary. This is something the AIA does not currently offer, but an attempt to provide such a methodology has been made in this paper.

To conclude, it should be mentioned that a granular risk assessment methodology can offer a valuable contribution to actuarial science. For example, performing AI risk assessment with a high degree of accuracy would improve underwriting and pricing of insurance policies, closing gaps in risk coverage, so that insurance products can better address the externalities of AI and distribute costs among social actors more efficiently. Unfortunately, we cannot dwell on this aspect in this paper.

8. Conclusions

The AIA defines several risk management measures related to the design, development and deployment of AIs. However, we have argued that the categorization of risk on which these measures rely is not sufficiently granular. In particular, we challenge the static association between risk categories and broad fields of application of AIs (model flaw). We think that this flaw may undermine the enforcement of the regulation.

---

generative foundation models must always disclose that the content was AI-generated (AIA, Recital 60g).

[96] Foundational models are AI systems trained on a large amount of unlabelled data and are very versatile for downstream functions.

We have offered a semi-quantitative approach to AI risk, articulated in two stages: (1) the construction of risk scenarios and (2) a proportionality-based quantitative assessment. For scenario construction, we have referred to the IPCC's theoretical framework and the literature on climate change risk. Accordingly, risk results from the interaction among four determinants, among individual drivers of determinants, and among extrinsic types of risk. For the second stage, we have referred to the quantitative approach developed by Alexy for balancing legal principles. Such a quantitative assessment aims to check whether the risk category assigned following the scenario construction is proportionate to the values involved in employing AIs.

The analysis has shown that the AIA's risk categories should be applied horizontally to the fields of applications of AIs so that, under varying scenarios, the same AIs can be estimated as unacceptable, high-risk, limited-risk or minimal-risk. We have pointed out that a semi-quantitative approach can improve the enforcement of the AIA and help address issues uncovered by the EU regulation, e.g., risk assessment for the GPAI, without undermining the protection of EU fundamental values and rights.

Future research should investigate further governance issues, including identifying which institutional bodies are called upon to apply the semi-quantitative risk analysis, with what specific faculties and with how much discretion in evaluating risk factors.