

A Crack in the Framework: Risk Management and the Governance Of General-Purpose AI with Systemic Risks

Toni Lorente

The Future Society
toni.lorente@thefuturesociety.org

Abstract

Risk-based regulation has become the default European response to general-purpose AI (GPAI). Systemic risk, however, has been translated into different duties of identification, assessment, and mitigation across the Digital Services Act, the EU AI Act, and its operational interpretation in the safety and security chapter of the GPAI Code of Practice. This paper argues that the risk-management frameworks (RMFs) that underwrite that translation cannot, on their own, accommodate systemic risks from frontier AI. To show this, I situate the rise of risk-based regulation in the late twentieth-century deregulatory turn and recover the industrial provenance of the risk equation. I then conduct a comparative analysis of the three principal EU definitions of systemic risk (DSA Article 34; EU AI Act Article 3(65) and Recital 110; and Appendix 1 of the Safety and Security Code of Practice) and show that they make a divergent treatment of fundamental rights, individual harm, and intentional misuse. Finally, I identify four structural reasons why RMFs strain when applied to systemic AI risk: the categorical heterogeneity of systemic risks; AI's inverted dual-use trajectory (civilian-to-military rather than military-to-civilian); the complex regulatory geometry that produces multiple incommensurable thresholds; and the asymmetry between actors who accept risk and actors who bear its consequences.

1 Introduction

The European Union's three principal digital regulations (the General Data Protection Regulation (Regulation (EU) 2016/679), the Digital Services Act (Regulation (EU) 2022/2065, hereafter DSA), and the EU AI Act (Regulation (EU) 2024/1689)) share a common analytical commitment: the governance of digital harm proceeds through the identification, assessment, and mitigation of risk. This commitment is, as I have previously argued, the inheritance of a broader regulatory turn that, over the last quarter of the twentieth century, replaced command-and-control regulation with risk-based and industry-informed frameworks

across most Western jurisdictions (Hutter, 2005; Lorente, 2024). Risk-based regulation has clear strengths: it is calibrated and tractable. It has, however, a less-discussed pre-condition: that the events it governs are amenable to probabilistic decomposition into likelihood and severity, and that the agents who bear residual risk are roughly the same as those who decide what residual risk is acceptable.

Neither pre-condition holds cleanly for the systemic risks that motivate the governance of general-purpose AI under the EU AI Act. The events at stake, such as loss of control over highly capable systems, chemical, biological, radiological and nuclear (CBRN) uplift, large-scale manipulation, or harms to democratic processes, sit uncomfortably inside the risk equation. The paths to harm, including the computation of probabilities and the estimation of severity from risks materialising are not commensurable across the categories the regulation lists side by side. The asymmetry between who accepts the risk (providers and regulators) and who bears the consequences (third parties, including population groups distant from the AI company), when it comes to systemic risk, cracks the foundations of risk management. This paper argues that these are structural features of systemic risk that distinguish it from the operational risks that the risk-management apparatus was built to handle.

In this paper I first provide a comparative reconstruction of how systemic risk is operationalised across the EU's most consequential digital regulations. In particular, the analysis covers three instruments drawn from two regulatory frameworks, namely Article 34 DSA, Article 3(65) and recital 110 EU AI Act, and the Safety and Security Chapter of the GPAI Code of Practice. This allows pinpointing the tension between the definitional anchor in the DSA, introduced in Art. 34, and the EU AI Act, introduced in Art. 3(65) and recital 110 EU AI Act. These anchors served as the conceptual basis for each regulatory framework. Second, I identify four structural reasons why risk-management frameworks

(RMFs) strain when applied to systemic AI risk: (i) the categorical heterogeneity of systemic risks; (ii) AI's inverted dual-use trajectory; (iii) the complex regulatory geometry of overlapping thresholds; and (iv) the asymmetry between risk acceptors and risk bearers.

I thus argue that the inherited RMF vocabulary distorts the systemic risk problem when it is applied without modification. Naming that distortion is a precondition for the operationalisation work that the AI Office, the Scientific Panel, and signatories to the GPAI Code of Practice will undertake over the coming years.

2 The Regulatory Turn and the Risk Equation

2.1 Deregulatory Inheritance

Risk-based regulation is an inheritance of the deregulatory turn of the late twentieth century. From roughly the 1980s onward, several Western jurisdictions replaced direct command-and-control regulation with frameworks calibrated through cost-benefit analysis and probabilistic risk assessment, and modeled, in important respects, on the practices of the private sector (Hutter, 2005; Lorente, 2024). Consequently, risk became the lens through which most events to be governed were re-described: an event was made tractable by being decomposed into a probability of occurrence and a severity of consequence. More importantly, the methodologies for assessing those quantities were primarily drawn from industrial settings where probabilistic decomposition had a strong empirical base.

Through this historical note, it becomes apparent how risk, finding its roots in industrial activity, is inherently better positioned to serve as a proxy for certain types of events. Events tied to operational, financial, or reputational consequences map more cleanly onto the lens of risk than those related to the infringement of fundamental rights or to the lived experiences of different individuals. This is because risk-based methodologies transfer unequally across the space of events that regulation in general, and the EU AI Act or the Digital Services Act in particular, seek to address.

This uneven transfer is reinforced by the increasingly central role that standards play. As Saari and Mügge (2026) show, the literature on risk-based regulation underscores how these standards are technologies that structure what counts as a governable problem. Documents like ISO 31000, the most widely referenced risk-management standard, were developed for industrial and managerial settings. These standards rest on the assumptions that (i) the risk owner is identifiable; (ii) the consequences of an event accrue principally to the owner or to a clearly bounded set of stakeholders; and (iii) the relevant uncertainty can be expressed as a

probability distribution informed by historical data. However, in the case of GPAI models that pose systemic risk, this is not necessarily true.

2.2 The Risk Equation

Risk management can be broadly conceptualised as a set of activities comprising identification, assessment, evaluation, and mitigation. The starting premise in risk-based approaches to governance and regulation is that every activity has inherent risk. Industrial activities, like manufacturing and commercialising any given product, but also non-industrial activities like going on a walk in the forest, smoking, or reading a book, have inherent risks.

The predicament is that by conducting several activities (commonly referred to as controls or mitigation measures), the likelihood of an event and/or the severity of its consequences can be reduced. The risk that is left is what is commonly referred to as residual risk. The risk equation therefore, can be conceptualised as: **(Inherent Risk) - (Mitigation) = (Residual Risk) < (Risk Tolerance)**.

Each step in the risk-management framework is subject to limitations: one can fail to identify a risk, misrepresent the consequences of an event materialising, or miscalibrate mitigations. But a critical breaking point of risk-management frameworks when it comes to risk-based regulation can be found in the risk-acceptability step. The problem of determining the risk tolerance or threshold drives the most pressing normative questions in risk-based governance, from who should set the risk tolerance or how much risk is acceptable to how are the consequences distributed across population groups.

As a risk-based regulation, the EU AI Act is exposed to these issues. The text defines risk as 'the combination of the probability of an occurrence of harm and the severity of that harm' (Article 3(2), Regulation (EU) 2024/1689). This definition is a practical simplification of the definition in the risk-management gold standard ISO 31000, whereby risk is usually expressed in terms of risk sources, potential events, their consequences, and their likelihood.

The lack of exhaustive risk-identification processes (Schuett, 2024), as well as the questionable robustness of existing mitigation strategies, raises concerns about the maturity of risk management for frontier AI. This is further exacerbated by the necessity to address systemic risks, which are similar in form but different in kind from operational or particular risks. The probabilistic calculation of systemic risks, and the consequences of those risks materialising, rests on assumptions about societal preferences, technology, and distributional impacts that are far more brittle than the corresponding assumptions in the operational case. This raises significant questions about the appropriateness of a purely risk-based approach to governing the frontier of general-purpose AI (GPAI).

3 Systemic Risk in EU Law: A Comparative Analysis

This section reconstructs how systemic risk is operationalised across the three EU instruments most directly relevant to GPAI: the EU AI Act, the DSA, and the Safety and Security Code of Practice (CoP) drafted under Article 56 EU AI Act. These reconstructions surface a structural tension, since the two regulatory instruments under analysis rest on incompatible definitional anchors. This has direct consequences over what counts as a systemic risk and who is empowered to address it.

3.1 Systemic Risk in the EU AI Act

Systemic risks are defined in Article 3(65) of the EU AI Act as 'a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain'. High-impact capabilities are further defined in the previous paragraph via a dynamic reference to the capabilities of the 'most advanced GPAI models' (Article 3(64)–(65), Regulation (EU) 2024/1689).

Recital 110 of the EU AI Act provides a partial list of risk sources and events that contribute to the space of systemic risks. This list has been further expanded in the Safety and Security Code of Practice, a voluntary set of commitments aimed at facilitating compliance with GPAI obligations that was drafted by independent experts at the request of the European AI Office. Appendix 1 of the Code, entitled 'Systemic Risks and other considerations', provides a list of systemic-risk types, their different nature, some of their sources, and a list of specified systemic risks that must be taken into account during the risk-identification process.

Recital 110 follows, with some adaptation, the structure of ISO 31000. It distinguishes sources from events and consequences, and it lists factors that increase or modulate risk along the lifecycle of a model. The list is illustrative: it names major accidents, disruptions of critical sectors, harms to public health, safety, security, democratic processes, and the dissemination of illegal, false, or discriminatory content. It also identifies CBRN uplift, offensive cyber capabilities, interaction with physical systems, self-replication, harmful bias, and chain-reaction risks as concrete pathways. Appendix 1 of the CoP refines this taxonomy by distinguishing capabilities, propensities, and affordances as three families of sources, and by specifying four 'specified systemic risks' that signatories must consider as a baseline (that is, CBRN, loss of control, cyber offence, and harmful manipulation).

3.2 Systemic risk in the DSA

Article 34 DSA defines systemic risks along four dimensions: (a) the dissemination of illegal content; (b) "any actual or foreseeable negative effects for the exercise of fundamental rights"; (c) "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security"; and (d) any actual or foreseeable negative effects in relation to gender-based violence, public health, the protection of minors, and serious negative consequences to a person's physical and mental well-being (Article 34, Regulation (EU) 2022/2065).

The DSA's framing is structurally different from the EU AI Act's along three axes. First, the qualifying threshold is reach, not capability. The obligation to assess and mitigate systemic risks is linked to Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), defined as services with at least 45 million monthly active users in the Union. Second, fundamental rights are enumerated with reference to specific Charter articles (e.g., dignity (Art. 1), private and family life (Art. 7), data protection (Art. 8), freedom of expression (Art. 11), non-discrimination (Art. 21), the rights of the child (Art. 24), and consumer protection (Art. 38)) rather than appearing as a single residual category. Third, Article 34(2) treats intentional manipulation of the service (inauthentic use, automated exploitation, algorithmic amplification) as a structural analytical category.

3.3 Differences Across the Conceptual Anchors

Under the DSA, a service becomes subject to systemic-risk obligations by accumulating users, regardless of how sophisticated its underlying technology is. The EU AI Act, by contrast, relies on capability to trigger obligations, operationalised through a presumption at 10^{25} floating-point operations of training compute (Article 51(2)), with reach playing a contributing but not constitutive role. Moreover, the safety and security chapter of the Code of Practice partially bridges the two by virtue of an analytic softening in treating capability-dependence and reach-dependence as equal contributing factors (Appendix 1.2.2).

Among the three governance mechanisms, three additional divergences stand out. First, fundamental rights occupy categorically different positions: a primary, Charter-anchored category in the DSA, in contrast with a sub-category of impact in the EU AI Act and CoP. Second, individual wellbeing is an explicit systemic-risk dimension in DSA Article 34(1)(d) but has no equivalent in the EU AI Act or CoP, both of which frame risk at a population scale. Third, the DSA traces harm from service design through user behaviour to outcome, whereas the EU AI Act traces harm from model capability through deployment to outcome.

Dimension	DSA — Art. 34	EU AI Act — Art. 3(65) + Rec. 110	CoP — Appendix 1
Definitional anchor	Reach (≥ 45 M monthly active EU users); no capability threshold.	High-impact capabilities (presumed at $\geq 10^{25}$ FLOP); reach is contributing.	N/A (Tacitly inherited from the EU AI Act, not explicitly defined in the text of the Code)
Risk categories	Four enumerated: (a) illegal content; (b) fundamental rights; (c) civic discourse and public security; (d) individual well-being.	Open-ended, illustrative list (Recital 110): major accidents, critical sectors, public health, democratic processes, CBRN, content harms.	Four specified systemic risks (App. 1.4): CBRN, loss of control, cyber offence, harmful manipulation.
Fundamental rights	Primary category, with seven Charter articles enumerated.	One impact domain among others; Charter rights not enumerated.	Listed as one risk type (App. 1.1(4)); not mapped to specific Charter articles.
Individual harm	Explicit (Art. 34(1)(d)): physical and mental well-being.	Implicit; framing is societal.	Implicit; framing is population-scale.
Causal theory	Service design $\rightarrow$ user behaviour $\rightarrow$ harm.	Model capability $\rightarrow$ deployment $\rightarrow$ harm.	Capabilities + propensities + affordances $\rightarrow$ harm.
Misuse	Explicit and structural (Art. 34(2)).	Implicit (Recital 110).	Explicit (App. 1.3.3); offence-defence balance treated as a factor.
Threshold criterion	Reach (≥ 45 M EU MAU).	Capability (presumed $\geq 10^{25}$ FLOP).	N/A
Legal status	Binding; enforced by Digital Services Coordinators and the Commission.	Binding; enforced by the AI Office.	Voluntary; compliance creates a presumption of conformity (Art. 56(5) EU AI Act).

Table 1. Comparative analysis of systemic risk definitions across the DSA (Art. 34), the EU AI Act (Art. 3(65) and Recital 110), and the Safety and Security Code of Practice (Appendix 1).

The most analytically developed treatment of systemic-risk propagation, with the six-characteristic taxonomy in Appendix 1.2.2 of the CoP, appears in the legally weakest of the three instruments. That is because even though the Code of Practice can be used to signal compliance with the EU AI Act, it is not legally binding, and compliance can be demonstrated through alternative means (see point 9 on page 2 and point 91 on page 24 of the Guidelines on the scope of the obligations for general-purpose AI models established by AI Act, published on July 18, 2025). This inversion, where the soft-law text carries the conceptual load while the binding instruments carry the obligations, is itself part of the problem identified in Section 4.

The divergences explored across instruments are, on the one hand, functional and well-grounded, given that the DSA governs platforms whereas the relevant sections of the EU AI Act govern general-purpose AI models. Different triggers and frameworks for different technical objects are thus justified. However, and on the other hand, these divergences are also inconsistent when assessed from the perspective of the regulator, since differences in treatment of fundamental rights or individual wellbeing are hardly justified by differences in regulated objects. The integration of AI into platforms (and even the open question on concurrent applicability of different frameworks like the EU AI Act and the DSA to the same object, as I have previously explored in Lorente & Gardhouse, 2026), means that providers and regulators

must navigate multiple laws simultaneously. It is particularly in those cases that divergences among definitional anchors on fundamental rights, or individual harm, for example, may lead to conflicting compliance signals for the same deployment action.

3.4 Related Work

Work on risk thresholds by Koessler et al. (2024) argues that frontier AI requires a move beyond compute and capability thresholds towards risk thresholds (understood as direct estimates of harm categories deemed unacceptable) set by regulators rather than by companies. The argument in this paper is upstream of theirs: even a regulator-set threshold, applied through the standard probabilistic approach, runs into the four structural issues set out in Section 4.

On the other hand, work on frontier AI risk management frameworks, exemplified by Campos et al. (2025), explores how frontier AI developers could adopt the practices of established risk-management disciplines (such as ISO 31000, NIST AI RMF, the financial-services three-lines-of-defence model), and bridges the gap between current AI practice and that established literature. Campos et al. rightfully identify that the current practice of frontier developers falls short of the methodological standards of mature RMF disciplines. However, this paper holds that the established RMF disciplines themselves were calibrated for operational risks and that mapping them onto systemic risks inherits the four structural mismatches identified in Section 4. While the conceptual bridge is sound, there is a need to limit its applicability.

Finally, this work also engages with the literature developing a risk taxonomy such as the AI Risk Repository (Slattery et al., 2024), the systemic-risk taxonomy of Uuk et al. (2024), the language-model risk taxonomies of Weidinger et al. (2021, 2022), and the foundational-challenges synthesis of Anwar et al. (2024). Risk taxonomies classify the harm categories that any RMF for frontier AI must contend with.

Two further bodies of work bear on my argument but are addressed only briefly here. Chamberlain (2023) develops a tort-law critique of the EU AI Act's risk-based approach that converges with our argument on the asymmetry between risk acceptors and risk bearers. I return to this convergence in section 4.4., when discussing the asymmetry argument. Stilgoe et al. (2020) articulate a responsible-research-and-innovation framework that treats the institutional response to events as analytically prior to the estimation of probabilities. I draw on this directly in section 4.5.

4 Conceptual Incompatibility Between Systemic Risk and RMFs

The incompatibility among definitional anchors identified in section 3 is further amplified by the fact that the inherited risk-management vocabulary (that is, inherent risk, mitigation, residual risk, tolerance) was developed for a class of events that systemic AI risks do not belong to.

4.1 The Categorical Heterogeneity of Systemic Risks

Despite the variance across systemic risks, systemic risk as a class is categorically different from operational, reputational, financial, or compliance risks. The taxonomy of systemic risks proposed by Uuk et al. (2024) enumerates thirteen categories that span dimensions as different as the concentration of military power and the erosion of communication systems. The AI Risk Repository (Slattery et al., 2024) and the foundational-challenges literature (Anwar et al., 2024) reach similar conclusions through different routes: the events grouped under the systemic risk label do not share a common metric of severity, time horizon, or set of causal pathways.

The risk equation requires that severity and likelihood be combined into a comparable quantity so that residual risk can be tested against a tolerance threshold. When the events being aggregated are categorically different, the aggregation operation may be unfeasible. Following Chang (2002), I describe such risks as incommensurable, in that there is no common scale on which they can be ranked, summed, or traded against one another. The RMF, applied to incommensurable categories, produces a number that looks tractable but is not.

Heterogeneity is both between categories of harm and between scales of harm. Smuha (2021) develops the case that AI's societal harms are systematically under-recognised by regulatory frameworks oriented towards individual rights and individual remedies, and that the gap between individual and societal harm is a category mismatch that the EU AI Act inherits from its product-safety lineage (the earlier claim that DSA Article 34(1)(d) names individual well-being as a systemic-risk dimension where the EU AI Act does not is directly related to this point). The categorical heterogeneity of systemic risks is, in part, a heterogeneity between harms that have individual victims and harms whose victims are collectivities (a democratic public, a labour market, an information ecosystem).

4.2 AI's Inverted Dual-Use Trajectory

The governance frameworks for technologies that pose systemic risks were largely developed for technologies that originated in the military and then found civilian applications. Nuclear technology is the canonical case: the

weaponisation of fission preceded its commercial deployment by roughly a decade, and the institutional architecture (the IAEA, the NPT, the export-control regimes) was built when the dual-use boundary still ran from military to civilian (Duffey & D'Auria, 2020), with biological and chemical weapons following a similar pattern.

In contrast, GPAI model development has been industrially and commercially driven, with security and military applications emerging downstream, often after capabilities have been deployed at scale to the consumer market. The asymmetry matters because the standard governance instruments (including export controls or secrecy classification) operate on the assumption that the state can hold the dangerous capability away from civilian markets long enough to design civilian-grade safeguards. With GPAI, that assumption is reversed: civilian deployment outpaces governance, and the artificially excludable resources (training compute, frontier datasets, specialised talent) are concentrated in private hands. Risk management designed for the military-to-civilian path inherits affordances that the civilian-to-military path does not provide.

4.3 Complex Regulatory Geometries: The Problem of Thresholds

The EU AI Act includes three provisions under which a GPAI model is deemed to pose systemic risks: (i) if it has high-impact capabilities, presumed for models trained with a cumulative compute budget of over 1025 FLOP (Article 51(2)); (ii) on the basis of a Commission decision; or (iii) following a qualified alert by the Scientific Panel on the basis of the criteria in Annex XIII, which include the number of parameters, the data used to train the model, and the model's impact on the internal market.

In the literature, these provisions (broadly described as thresholds) are categorised in three clusters: compute thresholds, capability thresholds, and risk thresholds, the last of which are direct estimates of categories of risk deemed unacceptable (Koessler et al., 2024). Compute thresholds are easy to verify but are increasingly disconnected from capability as algorithmic and architectural efficiency improve. Conversely, capability thresholds are conceptually closer to the harms regulators care about but require evaluation methods that do not yet exist at sufficient maturity for systemic risks (Anwar et al., 2024). Finally, risk thresholds are intended to be more accurate but require defining acceptable levels of harm in domains where there is no established societal consensus and where the consequences are unevenly distributed across populations.

Beyond the EU AI Act's own three pathways, the regulatory geometry is further complicated by the DSA's reach threshold (45 million monthly active EU users for VLOP or VLOSE designation). A single deployment can sit under three different threshold regimes simultaneously, with the

result that systemic-risk-bearing is a layered classification that changes depending on which instrument is doing the classifying. The notion of a tolerance threshold (in the right-hand side of the risk equation) presupposes a single comparator, yet the regulatory geometry does not provide one.

4.4 Asymmetrical Risk Appetites

Risks and benefits of AI development and implementation vary substantially among stakeholders and population groups, and they can co-occur, adding an additional layer of complexity to drawing the line between acceptable and unacceptable practices. The notion of 'risk appetite' (borrowed from business practice as an equivalent term to risk thresholds) is problematic because appetites vary substantially among stakeholders, and those accepting certain risks and reaping the rewards are categorically different from those suffering potential downstream, negative consequences.

Koessler et al. (2024) argue that regulators are better positioned to define socially desired thresholds, especially when certain activities may cause externalities (GPAI with systemic risk being an instance of that). However, given the political economy of GPAI model development, company-established thresholds may excessively inform political and legal commitments (Walker, 2000), resulting in a lock-in or endorsement of industry-led risk appetites that are not representative of all the stakeholders exposed to AI risk.

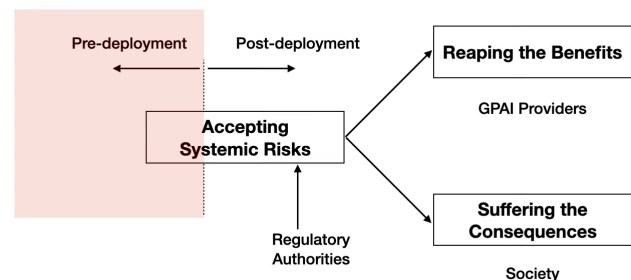


Figure 1. Visual representation of the imposition of systemic risks from GPAI providers (who may potentially reap the benefits) onto society (as a broad category of actors bearing the consequences of risk materialising). Moreover, the red area is intended to signify how regulatory authorities intervene on the acceptance of risks after GPAI models have been deployed.

AI development and implementation, as well as risk-benefit debates, are rooted in a stark asymmetry of information. The scientific understanding of the most advanced GPAI models is incomplete at best, with notions like 'capabilities' lacking an agreed-upon conceptualisation (Anwar et al., 2024). Combined with methodologically vague approaches to measuring and characterising the risk that GPAI models pose, and insufficient means for independent third parties to

investigate them, this is shifting the balance of power towards companies, which have exclusive control over the technical object being governed. Even where regulators are best positioned to define acceptable risk thresholds, the lack of visibility and the mediation of AI companies on the available information about GPAI models unevenly distributes the power to affect resulting risk thresholds. The information asymmetry, as well as a notorious push to control the language of frontier AI governance framework and regulation, are contributing to explicit industry efforts to shape what safety at the AI frontier looks like, including the proposal of governance bodies and standardised processes (Google, 2026). All of this, in a context of well-documented regulatory capture by Big AI companies (Birhane et al., 2026).

External effects (or externalities) exist when an actor's utility or production function depends on the behaviour of another actor who does not take that effect into account in their decision-making (Verhoef, 1999). As Dahlman (1979) put it in introducing the concept's operational form, the presence of an externality marks a divergence between private and social cost: when all voluntary contractual arrangements have been concluded, some interactions remain that ought to be internalised but that the market, left to itself, cannot internalise. Systemic risks in AI are, in this respect, externalities of an unusually broad kind. They are imposed by some actors on others without consent or market-mediated compensation, and often without the affected parties even being identifiable *ex ante*. Pigovian instruments (e.g., taxes equal in value to the negative externality, liability rules, mediation, information disclosure) are not theoretically incoherent for systemic risks in the case of AI. However, they may be operationally impossible because the consequences are global, hard to monetise, and temporally detached from the activity that generates them.

Operational risks are reflexive, in that the actor who accepts the risk and the one who bears the consequences if it materialises is the same. Systemic risks, however, are imposed, given that the actor who accepts the risk is not the actor who bears the consequences. This is the asymmetry Chamberlain (2023) identifies from the tort-law side, where it appears as a problem of who can bring a claim and against whom (but, in this case, here it is about who can set the residual risk threshold). The risk equation can therefore not be applied without modification to systemic risks.

In this context, the EU AI Act's risk-based architecture could be understood as a governance fix, under Ulnicane's framework (2025), given it proposes a settlement that narrows the scope of regulatory attention from the general impacts of AI on society to the specific liabilities of identifiable providers, in exchange for legal certainty and a path to market. Paul (2024) documents the parallel move by which the European Commission has framed risk-based regulation as a tool of competitiveness, leading, in her own terms, to a

competitive common AI market within which 'trustworthy' becomes an export commodity. Both accounts converge on the observation that the asymmetry I identify is a constitutive feature of the governance framework. The framework is useful because it shifts political responsibility for the systemic consequences of AI development away from the actors who shape it and onto the populations who experience it.

5 Exploring Alternatives

The structural mismatches identified above do not leave regulators without analytical resources. They do, however, suggest that those resources must be drawn from vocabularies other than the inherited RMF. In this section, I explore three families of resources to illustrate what some alternatives could look like, without presuming to establish an exhaustive list.

First, precautionary governance (in a strong, action-guiding sense) reframes the regulatory question from "what is the probability of harm?" to "what is the burden of proof required before proceeding?" (Sunstein, 2005; Kriebel et al., 2001). Kriebel et al. (2001) articulate the precautionary principle as having four constitutive components: (i) taking preventive action in the face of uncertainty; (ii) shifting the burden of proof to the proponents of an activity; (iii) exploring alternatives to possibly harmful actions; and (iv) increasing public participation in decision-making. The second component (i.e., shifting the burden of proof) directly addresses the incommensurability and potential irreversibility identified in sections 4.1. and 4.4., since it proposes a shift from the likelihood of harm towards the guarantee of safety. However, Sunstein (2005) raises the objection that a strong application of the precautionary principle can be paralysing. This is particularly important when addressing the issue of systemic risk, since both action and inaction, of any kind, may generate risk, leaving the principle with limited action-guiding content. However, when it comes to catastrophic or irreversible outcomes, Sunstein agrees that precautionary reasoning may be well-suited (that is, for the type of scenarios fleshed out in Chapter V of the EU AI Act and the Safety and Security Code of Practice).

Second, deliberative governance frameworks (Dryzek, 2012) shift the risk-acceptability question from a technical to a democratic register, treating the determination of tolerable systemic risk as a matter of collective choice rather than expert estimation. Dryzek (2012) develops the concept of a deliberative system, whereby labour is functionally divided across actors, including institutions decision-making spaces, fostering authentic deliberation free of coercive reason-giving and where technical premises can also be contested. Such a proposal could help bridge the asymmetry identified in Section 4.4, where those who accept systemic risk differ

categorically from those who bear its consequences. Ultimately, it is this asymmetry that structurally hampers the possibility of deliberation.

Third, Stilgoe et al.'s (2020) responsible research and innovation framework places institutional anticipation, reflexivity, and inclusion prior to probabilistic assessment, rather than as supplements to it. Stilgoe et al. (2020) identify four constitutive dimensions of responsible innovation: (i) anticipation; (ii) reflexivity; (iii) inclusion; and (iv) responsiveness. Anticipation requires systematic thinking about what futures a given technology makes possible or forecloses, without presupposing that those futures can be assigned probabilities. In this regard, it displaces the question of the likelihood of a given harm, and brings to the fore the issue of addressing the institutional capacities needed to respond to surprises under irreducible uncertainty. On the other hand, the notion of reflexivity invites governance institutions to examine the assumptions embedded in their own frameworks. This is especially pertinent in the context of the inherited RMF vocabulary, as it emerged and was originally calibrated to address operational risks, therefore distorting the problem of systemic risk, when applied without modification. Lastly, and while inclusion is important as a procedural guarantee, responsiveness implies that governance architectures must be designed for adaptation, and not as residual-risk thresholds that lock in current probability estimates against rapidly evolving capabilities.

In addition to the three sets explored in this section, governance frameworks mandating insurance, as well as licensing and pre-market authorisation regimes may contribute to developing auditing roles or inverting the direction of proof respectively. Insurance can help relocate judgement about acceptable risk to market mechanisms and underwriters, weakening the agency of developers to unilaterally set tolerance thresholds, whereas licensing may help invert the burden of proof and even help develop better evaluation methods so that the issue is not about the residual risk, but about the guarantee of absence of certain capabilities. Moreover, red lines or categorical prohibitions can help delineate the set of acceptable use cases for frontier AI, regardless of the result of computing the risk equation.

Conclusion

Risk-management frameworks, in the form they were inherited from late-twentieth-century industrial regulation, cannot on their own accommodate systemic risks from frontier AI. Their core operations (namely identifying, assessing, mitigating, computing residual risk, and comparing it against tolerance) presuppose conditions that systemic AI risks do not satisfy. Systemic risks do not have a single comparator, an identifiable risk-owner, a stable severity scale, or

a reflexive relationship between the actor accepting the risk and the actor bearing the cost.

In this article I have shown that the three EU instruments most directly relevant to GPAI rest on incompatible definitional anchors. Moreover, and beyond this definitional point, I have discussed four structural reasons why RMFs strain when applied to systemic AI risk, that is the categorical heterogeneity of the events being aggregated, AI's inverted dual-use trajectory, the complex regulatory geometry of overlapping thresholds, and the asymmetry between actors who accept risk and actors who bear its consequences.

With this, I do not argue that the AI Act, the DSA, or the Code of Practice are wrongly designed. However, I hold that the conceptual vocabulary they import distorts the systemic-risk problem. The language of risk is not neutral: it constitutes a singular vocabulary, with a particular history, that distributes analytical attention in concrete ways. This language leaves the key questions tied to systemic risks from general-purpose AI unaddressed.

The structural mismatches identified do not, however, leave regulators without analytical resources. Precautionary frameworks that invert the burden of proof, deliberative architectures that make risk-acceptability a matter of collective determination, or responsible-innovation approaches that place anticipation and reflexivity prior to probabilistic assessment are partial alternatives to a hegemonic risk-based approach. And while none of these alternatives can replace the risk-management governance frameworks on their own, they collectively delimit more precisely the domain within which the risk management framework is the right tool to address the challenges posed by frontier AI.

Acknowledgments

I thank Kathrin Gardhouse and Zlatko Grigorov for discussions on an earlier draft of this paper, as well as the other participants at the LawAI Conference in 2025 and IE Law-tomation days in 2024, where the original idea for this paper took shape.

References

- Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E.S., Jenner, E., Casper, S., Sourbut, O., & Edelman, B.L. 2024. Foundational challenges in assuring alignment and safety of large language models. arXiv preprint arXiv:2404.09932.
- Birhane, A., Angius, R., Agnew, W., Pandit, H. J., Mitra, B., Dobbe, R., & Talat, Z. 2026, June. Big AI's Regulatory Capture: Mapping Industry Interference and Government Complicity. In The 2026 ACM Conference on Fairness, Accountability, and Transparency (pp. 2586-2607).
- Campos, S., Papadatos, H., Roger, F., Touzet, C., Quarks, O., & Murray, M. 2025. A frontier AI risk management framework: Bridging the gap between current AI practices and established risk management. arXiv preprint arXiv:2502.06656.

- Chamberlain, J. 2023. 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*, 14(1), 1–13. doi:10.1017/err.2022.38.
- Chang, R. 2002. The possibility of parity. *Ethics*, 112(4), 659–688. <https://doi.org/10.1086/339673>.
- Dahlman, C. J. 1979. The problem of externality. *The journal of law and economics*, 22(1), 141–162.
- Duffey, R. B., & D'Auria, F. 2020. Nuclear energy and its history: past consequences, present inadequacies and a perspective for success. *Energy and Power Engineering*, 12(06), 193–236. <https://doi.org/10.4236/epe.2020.126014>.
- Dryzek, J. S. 2012. *Foundations and frontiers of deliberative governance*. Oxford University Press.
- European AI Office. 2025. General-Purpose AI Code of Practice. Safety and Security Section.
- Google. 2026, June. A pragmatic approach to AI governance in America. <https://static.googleusercontent.com/media/publicpolicy.google/en/resources/a-pragmatic-approach-to-ai-governance-in-america.pdf>
- Hutter, B. M. 2005. The attractions of risk-based regulation: accounting for the emergence of risk ideas in regulation (Vol. 33). London: CARR.
- Koessler, L., Schuett, J., & Anderljung, M. 2024. Risk thresholds for frontier AI. arXiv preprint arXiv:2406.14713.
- Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E. L., Quinn, M., Rudel, R., Schettler, T., & Stoto, M. 2001. The precautionary principle in environmental science. *Environmental health perspectives*, 109(9), 871.
- Lorente, A. 2024. Institutional review boards as soft governance mechanisms of R&D. *Journal of AI Law and Regulation*, 1(1), 40–49. <https://doi.org/10.48550/arXiv.2410.19574>.
- Lorente, T., & Gardhouse, K. 2026. Between search and platform: ChatGPT under the DSA. arXiv preprint arXiv:2601.17064.
- Paul, R. 2024. European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a competitive common AI market. *Regulation & Governance*, 18(4), 1065–1082. <https://doi.org/10.1111/rego.12563>.
- Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). (2022). *Official Journal of the European Union*, L 277. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.
- Saari, L., & Mügge, D. 2026. Forking paths of AI governance—how risk management frameworks shape the politics of AI. *Critical Policy Studies*, 1–21. <https://doi.org/10.1080/19460171.2026.2669063>
- Schuett, J. 2024. Risk management in the artificial intelligence act. *European Journal of Risk Regulation*, 15(2), 367–385. <https://doi.org/10.1017/err.2023.1>
- Slattery, P., Saeri, A. K., Grundy, E. A., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., Thompson, N. 2024. The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. arXiv preprint arXiv:2408.12622.
- Smuha, N. A. 2021. Beyond the individual: governing AI's societal harm. *Internet Policy Review*, 10(3). <https://doi.org/10.14763/2021.3.1574>.
- Stilgoe, J., Owen, R., & Macnaghten, P. 2020. Developing a framework for responsible innovation. In *The Ethics of Nanotechnology, Geoengineering, and Clean Energy* (pp. 347–359). Routledge.
- Sunstein, C. 2005. The precautionary principle as a basis for decision making.
- Ulnicane, I. 2025. Governance fix? Power and politics in controversies about governing generative AI. *Policy and Society*, 44(1), 70–84. <https://doi.org/10.1093/polsoc/puae022>.
- Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., & Prunkl, C. 2024. A taxonomy of systemic risks from general-purpose AI. arXiv preprint arXiv:2412.07780.
- Verhoef, E. T. 1999. Externalities. In *Handbook of environmental and resource economics*. Edward Elgar Publishing.
- Walker, W. 2000. Entrapment in large technology systems: institutional commitment and power relations. *Research policy*, 29(7–8), 833–846.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., et al. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). <https://doi.org/10.1145/3531146.3533088>.