AI, Digital Platforms, and the New Systemic Risk

Version September 19, 2025

(Early draft, comments welcome)

Authors: Philipp Hacker, Atoosa Kasirzadeh, Lilian Edwards[1]

Abstract. As artificial intelligence (AI) becomes increasingly embedded in digital, social, and institutional infrastructures, and AI and platforms are merged into hybrid structures, *systemic risk* has emerged as a critical but undertheorized challenge. In this paper, we develop a rigorous framework for understanding systemic risk in AI, platform, and hybrid system governance, drawing on insights from finance, complex systems theory, climate change, and cybersecurity --- domains where systemic risk has already shaped regulatory responses. We argue that recent legislation, including the EU's AI Act and Digital Services Act (DSA), invokes systemic risk but relies on narrow or ambiguous characterizations of this notion, sometimes reducing this risk to specific capabilities present in frontier AI models, or to harms occurring in economic market settings. The DSA, we show, actually does a better job at identifying systemic risk than the more recent AI Act. Our framework highlights novel risk pathways, including the possibility of systemic failures arising from the interaction of multiple AI agents. We identify four levels of AI-related systemic risk and emphasize that discrimination at scale and systematic hallucinations, despite their capacity to destabilize institutions and fundamental rights, may not fall under current legal definitions, given the AI Act's focus on frontier model capabilities. We then test the DSA, the AI Act, and our own framework on five key examples --- misuse for terrorism, large-scale discrimination, hallucinations, cybersecurity threats, and environmental effects --- to illustrate their respective strengths and limitations. Against this background, this paper proposes reforms that broaden systemic risk assessments, strengthen coordination between regulatory regimes, and explicitly incorporate collective harms. As localized AI and platform failures may increasingly escalate into structural disruptions, we provide a conceptual foundation and policy-relevant diagnostic toolkit for governing AI in complex, interconnected societies.

Keywords: systemic risk, AI systemic risk, Catastrophic risk, AI catastrophic risk, AI governance, AI policy, platform governance, AI Act, DSA, hybrid systems

[1] Philipp Hacker and Atoosa Kasirzadeh contributed equally as joint lead authors (family names are listed alphabetically); Lilian Edwards contributed in relation to the AI Act and its Code of Practice.

Table of Contents

I. Introduction

Throughout history, societies have grappled with technologies that promise progress yet can destabilize the systems into which they are introduced. This tension made the 20th century a period of sharper differentiation in how risk is understood: *individual risk*, specific harmful effects of technologies limited to a small number of individuals, versus *systemic risk*, whose collective or structural character warrants distinct analysis and, at times, regulation.[2] The 2008 financial crisis crystallized decades of research in financial economics into a global event that destabilized economies and drew urgent attention to the concept of systemic risk.

Today, systemic risks to financial infrastructure have not disappeared. Yet the growing complexity of digital infrastructure, and its entanglement with core political and social processes, adds new dimensions of systemic risk. Historically, systemic risk is well defined in finance: cascading failures that destabilize the financial system as a whole. AI and large platforms extend this risk capacity, creating common exposures and points of failure that traditional formulations of systemic risk do not fully capture.

This shift is now reflected in digital and AI risk governance which increasingly invokes *systemic risks* as a critical concern.[3] The concept has gained prominence in recent legislative efforts, notably in the European AI Act and the Digital Services Act (DSA). The DSA contains four major categories of systemic risk that providers of very large platforms and search engines need to identify and mitigate, ranging from the dissemination of illegal content, to negative effects on fundamental rights, risks for civic discourse and electoral processes, and harms related to minors and well-being (Marsh, 2024). The AI Act, in turn, establishes obligations for providers of general-purpose AI (GPAI) models that bring attention to systemic risks. Its treatment of systemic risk is not enumerative as the DSA's, but provides a conceptual definition that focuses on high-impact capabilities, impact on the EU market, and propagation down the value chain (Art. 3(65) AI Act).

In practice, however, the AI Act attempts to designate AI models as possessing systemic risks based primarily on training compute thresholds --- presuming ``high-impact capabilities'' when training involves more than 10^25 floating-point operations (FLOPs). By this measure, many large-scale language models (LLMs) on the market would be considered systemic risks. However, as Bertuzzi (2025), Hacker & Holweg (2025) and Somala et al. (2025) note, rapid improvements in reinforcement learning and compute efficiency have challenged this designation method even before proper implementation. If systemic risk is defined mainly by compute thresholds, it may miss smaller AI systems deeply integrated into critical infrastructure while misclassifying computationally intensive models with limited societal integration as systemic risks. Hence, the Act's compute threshold is already under scrutiny for revision.

More irritatingly, Art. 3(64) AI Act defines high-impact capabilities as "capabilities that match or exceed the capabilities recorded in the *most advanced* general-purpose AI models [our emphasis]." Systemic risks, in the AI Act, need to be specific to those most advanced GPAI models. Can this be said of large-scale manipulation, discrimination, or hallucinations? Such ambiguities highlight the need for a structured framework beyond simple size-based thresholds. While the AI Act theoretically contains a richer set of designation parameters in

[2] Various different terminologies are used to refer to non-individual-based risks such as "catastrophic risk" in some Frontier AI safety frameworks (Kasirzadeh, 2024); individual risk, in turn, is also called "idiosyncratic risk" (De Bandt & Hartmann, 2000).

[3] See Uuk et al., 2024 and Kasirzadeh, 2025 for a detailed discussion.

Annex XIII, they are also connected to the definition of systemic risk in the Act, which exclusively focuses on “the most advanced” GPAI models, as we shall discuss.

Despite these regulatory mentions, the boundaries of systemic risk in AI-driven systems and platforms remain underdeveloped. Systemic risk is often regarded as a ``you-know-it-when-you-see-it'' concept (Benoit 2017; Marsh 2024). This stance, we argue, is problematic and can render the notion so loose and useless as to provide little practical guidance: clarifying systemic risk in AI-driven systems, platforms, and hybrid structures combining both, requires identifying, measuring, and monitoring the complex propagation mechanisms by which localized failures cascade across interconnected systems. If, for example, systemic risk is exclusively believed to be caused by advanced capabilities of single AI models, the definition contradicts decades of research on systemic risks in domains like finance, where the focus has been on interaction effects rather than merely the properties of individual algorithmic components.

This paper aims to develop an analytically rigorous framework for understanding systemic risk in AI governance by identifying conditions under which AI risks transition from localized failures to systemic risks. We do so in four steps. First, we draw on lessons from the historical study of systemic risk, financial risk regulation, and complex system sciences to identify minimal conditions common to core definitions of systemic risk in the literature. Second, and based on our initial historical findings, we formulate conceptually rich list of four assessment criteria for specifying AI-driven systemic risks:

(i) the **scale and scope** of expected damage;
(ii) the **emergence of collective harms**, which exceed the sum of individual impacts;
(iii) the **potential irreversibility** of certain harms, and
(iv) **complexity**, for example **interconnectedness**, which enables cascading and non-contained effects;

Together, they capture the *dimension* (i), the *nature* (ii and iii) and the *unpredictability* (iv) of harm from the manifestation of systemic risk.

Third, we identify four key levels at which systemic risks from AI-driven systems could manifest:

(1) Single-model systemic risks: failures from a single AI model, when widely adopted;
(2) multi-model systemic risks: systemic failures when multiple AI models fail in synchronized ways;
(3) model-platform integration systemic risks: AI systems embedded into large-scale platforms creating self-reinforcing feedback loops;
(4) model-institution integration systemic risks: AI systems embedded into governance, law enforcement, and finance creating structural risks.

Finally, we apply our investigation to AI and digital risk governance. In particular, we discuss the merits of addressing systemic risks in the DSA and the EU AI Act; we argue that the interpretation of systemic risk in the AI Act is too narrow; that additional governance and monitoring mechanisms are needed to address and contain potential systemic failures; and that a holistic understanding of systemic risk is needed to tackle hybrid structures combining AI and platforms.

The remainder of the paper is structured as follows: Section II offers a historical perspective on systemic risk in finance, climate change, and cybersecurity, highlighting lessons for AI governance. Section III distills common dimensions across twenty existing definitions of systemic risk from scholarship and policy documents, and formulates conceptual assessment

criteria tailored to AI-driven systems. Section IV identifies four levels at which systemic risks from AI can materialize, ranging from single-model failures to model-institution integration. Sections V and VI examine how systemic risk is treated in the Digital Services Act and the AI Act, respectively, contrasting their approaches and limitations. Section VII develops a comparative analysis and critiques, drawing out the implications for AI governance. Section VIII discusses practical applications, providing examples of systemic risks – such as misuse for terrorism, discrimination at scale, information pollution through hallucinations, cybersecurity, as well as climate and environmental risks – under existing frameworks and our proposed approach. Section IX concludes by reflecting on the need for a more comprehensive and coherent regulatory framework for AI and platform systemic risks.

II. A Historical Perspective: Systemic Risk in Financial Markets, Climate Change, and Cybersecurity

The concept of systemic risk gained prominence following the Great Depression, but was formally crystallized in financial literature during the 1980s and 90s, generally by reference to internal or external shocks which propagate through the market and affect a variety of financial or market institutions (Allen & Carletti, 2013; Brimmer, 1989; Cline, 1984; De Bandt & Hartmann, 2000; Galaz et al., 2021; Micova & Calef, 2023, p. 24 ff.; Summer, 2003).

1. Systemic Risk in Finance and Banking

The concept entered the policy arena most notably with the Lamfalussy Report (Bank for International Settlements, 1990). Produced by a committee chaired by the former President of the European Monetary Institute, Alexandre Lamfalussy, and under the auspices of the Bank for International Settlements, the report analyzed cross-border and multi-currency interbank netting schemes and established a set of minimum standards for their operation, with the aim to reduce systemic risk in international payment systems. It influenced subsequent regulatory frameworks for financial regulation, which in the 1990s addressed the mitigation of systemic risk that arose from interdependencies in the banking system. For example, the Settlement Finality Directive (98/26/EC) of 1998 aimed to reduce legal and systemic risks in payment and securities settlement systems. Recital 1 of that directive notes the "important systemic risk inherent in payment systems." The Banking Consolidation Directive of 2000 (2000/12/EC), in its Recital 52, emphasized the reduction of systemic risk through the role of clearing houses, which reduce counterparty risk in financial derivatives.

From these early regulatory efforts, the concept of systemic risk was catapulted into the spotlight by the financial crisis sparked by the collapse of Lehman Bank in 2007. It became clear that the concept not only had theoretical value, but actually helped understand one of the biggest economic crises of the after-war period. This was reflected in another report: the de Larosière Report of 2009 recommended the creation of a "European Systemic Risk Council" (recommendation 16). The main legal instruments that followed are Regulation (EU) No 1092/2010, which established the European Systemic Risk Board (ESRB), which operates until today; the Capital Requirements Regulation (CRR); and the Capital Requirements Directives (CRD). These instruments form the framework for macroprudential policy and the application of systemic risk buffers.

These acts also shed further light on the legal understanding of systemic risk in the banking sector. Article 2(c) of the ESRB Regulation defines systemic risk as "a risk of disruption in

the financial system with the potential to have serious negative consequences for the internal market and the real economy. All types of financial intermediaries, markets and infrastructure may be potentially systemically important to some degree." Systemic risk, in this understanding, is marked by the potential to threaten the stability of an entire system (in this case, the financial system), and the severity of harm for actors in that system in case of materialization. Importantly, systemic risk, therefore, has a collective and an individual dimension. It is realization would have a severe negative impact on entities in the system, but would also destabilize the system as such.

Recent developments in the financial sector show an increasing focus on the extension of macroprudential policy and systemic risk management to non-bank sectors such as investment funds and insurance. This reflects lessons from recent market volatility and the expanding role of these sectors in financial intermediation (Gutiérrez de Rozas, 2022).

These observations and more motivated (Kaufman, 2003) to provide one of the most cited definitions: "the risk or probability of breakdowns in an entire system, as opposed to breakdowns in individual parts or components." This definition emerged from studying the 1930s bank runs, where (debandt, 2000) documented how local bank failures triggered cascade effects across the entire financial system. The 2008 financial crisis led to a refinement by (Schwarcz, 2008), who emphasized that systemic risk represents "the probability that cumulative losses will occur from an event that ignites a series of successive losses along a chain of institutions or markets comprising a system."

The conceptualization of systemic risk, initially developed within the confines of financial regulation, has undergone a significant transformation in recent years. While the 2008 financial crisis crystallized the understanding of systemic risk in banking and financial markets, with a focus on interdependencies between entities and structured financial products, emerging global challenges have necessitated a broader application of this framework. Two particularly salient areas where systemic risk thinking has gained prominence are climate change and cybersecurity, both of which demonstrate the interconnected vulnerabilities that characterize modern economic and social systems – and are of direct relevance to advanced IT systems, such as platforms and AI, as well.

2. Climate Change as a Source of Systemic Risk

The Organization for Economic Co-operation and Development (OECD) identifies climate change as a source of systemic risk as follows:

"Every day, people face a variety of risks that may result in damage to what they value: their life, their health, the lives and health of others, their property, or the environment. Some of these risks affect individuals but have only an isolated impact on society – car accidents are an example. Others, however, may be on a much larger scale and their effects may spread much further. This report is concerned with the latter, more specifically, with those risks that affect the systems on which society depends – health, transport, environment, telecommunications, etc. Five categories of such risks are addressed: natural disasters, industrial accidents, infectious diseases, terrorism, and food safety. The report does not deal with systemic risks to markets, notably to financial markets, although some aspects of financial systems are considered in the analysis." (OECD, 2003, 9).

The Intergovernmental Panel on Climate Change (IPCC), in turn, does not use the word "systemic risk" but describes the system-level risk of accelerating climate change as follows: "Multiple climate hazards will occur simultaneously, and multiple climatic and non-climatic

risks will interact, resulting in compounding overall risk and risks cascading across sectors and regions" (IPCC 2022, B5).

In a piece bridging sectors, Monnin (2021) articulates how climate-related risks transcend traditional sectoral boundaries, creating cascading effects throughout the financial system and the broader economy. This perspective aligns with the evolving understanding that environmental degradation poses not merely localized threats but system-wide vulnerabilities – in multiple systems – that mirror the interconnected nature of financial systemic risks. Again, the risk brought about by climate change has a strong collective and individual dimension: it threatens to destabilize the planetary ecosystem, as well as a multitude of human-engineered subsystems; and it can bring about, significant harm to individuals via extreme weather events, degradation of soils and livelihood, vanishing territories, and other effects.

The transmission mechanisms of climate-related systemic risk operate through multiple channels. In the financial sector, physical risks from extreme weather events can simultaneously impact asset values across diverse geographical regions and economic sectors, and create correlated losses that undermine portfolio diversification strategies. Transition risks, arising from the shift toward a low-carbon economy, can trigger abrupt repricing of assets, particularly in carbon-intensive industries, potentially leading to stranded assets and widespread financial instability. These risks are further amplified by the non-linear nature of climate impacts and the potential for tipping points that could fundamentally alter economic relationships.

The financial regulatory response to climate-related systemic risk has been evolving in an accelerated way. Central banks and supervisory authorities have increasingly incorporated climate risk assessments into their macroprudential frameworks, recognizing that traditional risk management tools may be insufficient to address the long-term, uncertain, and potentially catastrophic nature of climate impacts (Carè et al., 2024; European Systemic Risk Board & European Central Bank, 2023; Financial Stability Board, 2022; Hiebert & Monnin, 2023; Hidalgo-Oñate et al., 2023). This integration reflects a growing consensus that climate change poses risks to financial stability comparable to, if not exceeding, those addressed in the wake of the 2008 crisis.

3. Cybersecurity as a Source of Systemic Risk

The digitalization of financial services and critical infrastructure has brought cybersecurity to the fore as another crucial dimension of systemic risk. The ESRB report of 2020 highlighted how cyber incidents can propagate through interconnected systems, creating contagion effects reminiscent of traditional financial crises. Unlike conventional financial risks, cyber threats are generally characterized by their intentional nature, and potential for simultaneous attacks across multiple institutions.

Gutiérrez de Rozas (2022) emphasizes how the increasing reliance on common technological infrastructures, shared service providers, and interconnected payment systems creates new vectors for systemic vulnerability. A successful cyberattack on a critical node in the financial system could disrupt payment systems, compromise data integrity, and erode market confidence, potentially triggering liquidity crises and operational failures across multiple institutions. The speed at which cyber incidents can unfold – measured in minutes or hours rather than days or weeks – poses particular challenges for traditional crisis management frameworks. As a consequence, Recommendation ESRB/2021/17 was issued in 2021 and

introduces an EU wide systemic cyber incident coordination framework for relevant authorities.

The regulatory response to cyber-related systemic risk has focused on developing operational resilience frameworks that go beyond traditional prudential measures. The ESRB's work on operational policy tools for cyber resilience represents an attempt to translate macroprudential thinking into the digital domain (European Systemic Risk Board, 2024). This includes stress testing for cyber scenarios, establishing information-sharing mechanisms, and developing coordinated response protocols. However, the cross-border and cross-sectoral nature of cyber threats presents continuing challenges for regulatory coordination and enforcement.

4. Lessons Learned

Historically, scholars have thus conceptualized systemic risk primarily in terms of financial institutions, particularly banks. From this field, regulatory scholarship and practice can learn that sources of risk (shocks) can be external and internal; and effects are either idiosyncratic (largely limited to one entity) or systemic (affecting multiple actors) (see also De Bandt & Hartmann, 2000; Micova & Calef, 2023, p. 42). The underlying reason for this initial tendency to focus on the financial sector, we think, is straightforward: banks and other financial institutions serve as a critical system of capital, making their failure --- especially in large numbers --- potentially devastating to capital availability and cost. As (Davis, 1992) noted, the most serious direct consequences of systemic risk involve "disrupt[ing] the payments mechanism and capacity of the system to allocate capital." Similarly, (Billio, 2012) characterized systemic risk as "any set of circumstances that threatens the stability of or public confidence in the financial system." Hence, a key lesson learned from the financial literature is that the system itself needs to be neatly defined; the risk sources (external versus internal) as well as the effects (idiosyncratic versus systemic) properly described; and the measurement of these different risks thoroughly operationalized.

What makes both climate and cyber risks particularly significant from a systemic perspective is their geostrategic relevance as well as their potential for interconnection and mutual amplification. Climate-related disruptions can increase vulnerability to cyber incidents, as stressed systems and crisis conditions create opportunities for malicious actors. Conversely, cyberattacks on critical infrastructure can impair climate adaptation and mitigation efforts. In the worst case, this may create feedback loops that amplify both types of risk.

These interconnections suggest that the traditional approach to systemic risk management, developed primarily for financial contagion, requires fundamental adaptation. The temporal scales differ markedly – while financial crises typically unfold over weeks, months or years, climate risks operate on decadal timescales with potential for sudden materialization, and cyber risks can manifest instantaneously. This temporal complexity challenges conventional risk assessment methodologies and regulatory response mechanisms.

The expansion of systemic risk beyond its financial origins necessitates a reconceptualization of regulatory approaches. Traditional macroprudential tools, such as capital buffers and leverage limits, have limited applicability to climate and cyber risks. Instead, regulators are developing new instruments that emphasize scenario analysis, operational resilience, and cross-sectoral coordination. This evolution reflects a growing recognition that systemic risks in the 21st century are characterized by complex interdependencies that transcend traditional regulatory boundaries. An example is the Digital Operational Resilience Act (DORA), which entered into force in January 2023. DORA represents a comprehensive attempt to translate systemic risk thinking into operational resilience requirements. Notably, the regulation

recognizes that systemic cyber risk often originates from concentration in third-party service providers, particularly cloud services and other critical ICT infrastructure. Hence, in the case of both climate change and cybersecurity – and unlike many traditional financial systemic risk triggers –, systemic risks originate outside of the target system (exogenous shocks), which makes the monitoring and regulatory mitigation of these risks all the more challenging. This is an issue we shall return to in the treatment of systemic risks in platforms and AI regulation below.

III. A Selected Definitions of Systemic Risk

Source	Definition
Renn et al., 2022	"Systemic risks are characterized by high complexity, multiple uncertainties, major ambiguities, and transgressive effects on other systems outside of the system of origin."
Helbing, 2013	"Systemic risk is the risk of having not just statistically independent failures, but interdependent, so-called 'cascading' failures in a network of N interconnected system components."
Schwarcz, 2008	"The probability that cumulative losses will occur from an event that ignites a series of successive losses along a chain of [financial] institutions or markets comprising... a system."
Kaufman, 2003	"The risk or probability of breakdowns in an entire system, as opposed to breakdowns in individual parts or components."
Davis, 1992	The most serious direct consequences of systemic risk involve "disrupt[ing] the payments mechanism and capacity of the system to allocate capital."
Billio, 2012	"Any set of circumstances that threatens the stability of or public confidence in the financial system."
Bloomfield and Wetherilt, 2012	"We consider a serious nuclear incident that has the potential for the release of radioactivity with associated plant damage as a "systemic event" and hence make the link to a financial market crash: an event that both damages the market and also potentially impacts the wider financial system and the broader economy."
Kaufman, 1995	"The probability that cumulative losses will occur from an event that ignites a series of successive losses along a chain of institutions or markets comprising a system."

Li et al., 2021	"Systemic risk induced by climate change is a holistic risk generated by the interconnection, interaction, and dynamic evolution of different types of single risks, and its fundamental, defining feature is cascading effects. The extent of risk propagation and its duration depend on the characteristics of the various discrete risks that are connected to make up the systemic risk."
Article 2(c) of the ESRB Regulation	"a risk of disruption in the financial system with the potential to have serious negative consequences for the internal market and the real economy."

IV. Conceptual Dimensions of Systemic Risk/our conceptual framework

1) Dimension (scale and scope) of harm. Definitions emphasize the \textbf{scale and scope} of systemic risks. Unlike individual-level risks, systemic risks affect large portions of an entire system (can be society, financial group, specific demographic, etc). They go beyond isolated incidents. This expansive reach is a defining characteristic of systemic risk.
2) Nature of harm/mechanism I: Simple aggregation vs. .... Collective and societal nature of systemic risks. The harm manifests at a societal level rather than merely being the sum of individual harms. This qualitative difference between individual and collective impacts is crucial for understanding why systemic risks require different assessment and governance approaches from those focused on protecting individual rights or welfare. → threatens one critical system or infrastructure
3) Nature of harm II: Potentially irreversible. Several definitions emphasize the long-term and potentially irreversible nature of systemic risks. Effects may persist long after the initial triggering events, and some changes to social structures, cultural norms, or environmental conditions may be difficult or impossible to reverse. This temporal dimension adds urgency to the need for anticipatory assessment and preventive measures, rather than relying solely on reactive responses after harm has occurred.
4) Unpredictable/not fully predictable harm: Complexity, for example Connectedness/interconnectedness. (to model via networks). Definitions highlight the propagation and cascading nature of systemic risks. Effects spread or cascade beyond the original system to affect other systems, creating chains of cause and effect that can be difficult to predict or control. This propagation can occur through various channels, including technological interconnections, market relationships, social networks, and institutional links. The ability of effects to cross traditional boundaries between systems is a key feature of systemic risks

   More generally: Complexity features prominently in most definitions. Complex interactions between components or systems create unpredictable outcomes that cannot be understood through simple linear models of cause and effect. The high degree of interconnection in modern socio-technical systems enables the rapid

transmission of effects across different domains, amplifying the potential impact of initially localized disruptions.

Additionally, there are four levels at which AI systemic risk can manifest:
(1) **Single-model systemic risks** – failures from one widely adopted AI model;
(2) **Multi-model systemic risks** – correlated failures when multiple models fail in synchronized ways;
(3) **Model-platform integration risks** – AI embedded into large-scale platforms generating feedback loops;
(4) **Model-institution integration risks** – AI systems embedded into governance, finance, or law enforcement, creating structural vulnerabilities.

Four parameters help characterize systemic risks from AI (See Figure 1).

Dimension	Definition	Categories
Source	Describes where the risk originates.	*- Single-model* *- Multi-AI* *- AI–human interaction* *- AI–institutional interaction*
Causal mechanism	Describes how system behavior leads to harm.	*- Direct* *- Indirect*
Impact profile	Describes the scale and distribution of harm.	*- Systemic* *- Localized but intolerable*
Catastrophic threshold	Defines when a risk becomes catastrophic, based on normative considerations.	*- Magnitude* *- Irreversibility* *- Tractability* *- Moral-political violation*

Figure 1. Formulating dimensions of systemic risk from AI

V. Systemic Risk in the DSA

The DSA introduces a comprehensive framework for platform regulation that centers, inter alia, on the concept of systemic risk, though it approaches this concept differently than AI Act or in financial regulation (see also Micova & Calef, 2023, Section 7). While Article 2(c) of the ESRB Regulation provides an explicit definition of systemic risk, just like the AI Act in its Article 3(65) (see below, next section), the DSA embeds the concept throughout its provisions without offering a singular definition, instead operationalizing it through a non-exhaustive list of systemic risks for VLOPs and VLOSEs, as well as explanations in the Recitals, which leave this definition of systemic risk particularly prone to stakeholder intervention and strategic maneuvering by interested parties (Griffin, 2025).

1. Conceptual Foundations

The DSA's approach to systemic risk implicitly draws on complex systems theory and network science. Platforms are understood as socio-technical systems where technical architecture, business incentives, and user behavior interact to produce emergent effects. Small changes in algorithm design or policy enforcement can cascade through millions of users, transforming individual actions into collective phenomena.

This complex-systems perspective explains why the DSA focuses on structural features rather than specific content categories. As documented by Gillespie (2018), platform affordances shape user behavior in ways that transcend conscious design choices, creating "the politics of platforms" that influence democratic discourse and social cohesion. The DSA attempts to make these politics explicit and subject to democratic oversight.

2. Statutory Framework and Risk Categories

Article 34(1) DSA establishes the core obligation for risk assessment and mitigation, requiring VLOPs and VLOSEs to "diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services." Article 34(1) mandates that this assessment be "specific to their services and proportionate to the systemic risks, taking into consideration their severity and probability," establishing a risk-based approach that calibrates obligations to actual threat levels – again, just like financial regulation and the AI Act.

The provision identifies four specific categories of systemic risk that platforms must assess, with Recital 80 providing additional interpretive guidance. First, VLOPs and VLOSEs need to address "the dissemination of illegal content through their services" (Article 34(1)(a) DSA). Recital 80 expands this category significantly, clarifying that it encompasses not only content dissemination but also illegal activities conducted through platforms. The recital provides concrete examples: dissemination of child sexual abuse material and illegal hate speech for dissemination, and the sale of prohibited goods and services as illegal activities. The systemic dimension emerges when such content or activities "spread rapidly and widely through accounts with a particularly wide reach or other means of amplification." Hence, not every localized illegal act on the platform constitutes an element of systemic risk; rather, the dissemination, amplification and propagation of content or activities as a necessary element of qualifying illegal items as part of systemic risk. Crucially, Recital 80 mandates that providers assess risks from illegal content "irrespective of whether or not the information is also incompatible with their terms and conditions." Hence, the law, not private rules, defines this risk category.

Second, Article 34(1)(b) requires assessment of "any actual or foreseeable negative effects for the exercise of fundamental rights." The provision enumerates specific Charter rights requiring particular attention: human dignity (Article 1), respect for private and family life (Article 7), protection of personal data (Article 8), freedom of expression and information including media freedom and pluralism (Article 11), non-discrimination (Article 21), rights of the child (Article 24), and consumer protection (Article 38). As an example, Recital 81 lists the exploitation of the weaknesses and inexperience of minors or addictive behavior caused by platform design. Interestingly, and in line with recent judgments from Member State constitutional courts (xxx) as well as certain judgments of the CJEU (Eigenberger etc. xxx), the DSA seems to assume that the listed fundamental rights do apply horizontally in the relationship between affected subjects and VLOPs or VLOSEs, irrespective of the latter’s status as private companies. This is a controversial position concerning primary EU law which cannot, obviously, be decided by secondary law, such as the DSA (see below, Overarching Lessons).

Third, Article 34(1)(c) DSA mandates the evaluation of "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security." This category recognizes platforms' role as critical infrastructure for democratic participation. The conjunction of civic discourse, electoral processes, and public security acknowledges how

information manipulation can simultaneously undermine democratic deliberation, electoral integrity, and social stability. However, notably, this is the only systemic risk category that is not elaborated upon in the Recitals, likely due to disagreement between Member States as to the role of platforms in suppressing legitimate or facilitating illegitimate speech in electoral and civic discourses (cf. Recital 82 DSA).

Fourth, Article 34(1)(d) DSA addresses "any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being." This category's specific mention of gender-based violence reflects growing awareness of how platforms can facilitate harassment and abuse. The inclusion of mental well-being alongside physical health recognizes emerging research on platform-induced psychological harms, from addiction patterns to anxiety and depression linked to platform design choices. As a consequence, Recital 83 DSA mentions behavioral addictions, and online disinformation campaigns as serious risks to public health, with a clear nod to the spread of medical disinformation during the COVID-19 pandemic (Roozenbeek et al., 2020).

Article 34(1)'s requirement that assessments be "specific to their services and proportionate to the systemic risks" establishes a tailored approach to risk evaluation which, however, also invites laxity and reduced scrutiny behind the shield of proportionality.

3. Application Thresholds

The DSA employs a quantitative approach based on reach to determine which platforms bear systemic risk obligations. Article 33 establishes that platforms or search engines with average monthly active recipients of 45 million or more in the Union qualify as VLOPs or VLOSEs. This threshold, representing approximately 10% of the EU population, reflects a legislative judgment that platforms reaching this scale possess the capacity to generate Union-wide impacts (European Commission, 2022).

This numerical threshold contrasts sharply with the AI Act's approach, which focuses on technical capabilities rather than user reach (see below). The DSA's metric recognizes that systemic effects in the platform context arise primarily from network effects and scale of human interaction rather than technical sophistication. A platform need not employ cutting-edge technology to create systemic risks if it commands sufficient user attention and engagement.

4. Risk Factors and Amplification Mechanisms

Article 34(2) identifies specific factors that platforms must consider in their risk assessments, with Recital 80's emphasis on "rapid and wide" spread providing the analytical framework. The design of recommender systems receives particular attention, as these algorithms determine information flow and can amplify harmful content or create filter bubbles. The recital's reference to "accounts with a particularly wide reach or other means of amplification" directly implicates recommendation algorithms that can transform isolated illegal content into viral phenomena affecting millions of users. The DSA recognizes that these design choices interact with user behavior in complex ways (Helberger et al., 2021; Helberger et al., 2022). This understanding drives the DSA's focus on platform architecture rather than content-specific rules.

VI. Systemic Risk in the AI Act

Systemic risk in the AI Act forms part of the rules governing GPAI models under Articles 51 to 56. A GPAI model is defined in art 3(63) as:

> "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, *except* AI models that are used for research, development or prototyping activities before they are placed on the market"

Models are mathematical objects that can be embedded in software and infrastructure to be used in various ways both by the model developer (vertical integration) or by downstream deployers. The AI Act defines a GPAI system in Article 3(66) as an AI system based on a general-purpose AI model that has the capability to serve a variety of purposes, both for direct use and for integration in other AI systems. Systemic risk as a regulatory problem primarily attaches to models rather than systems.

1. The Rules in the AI Act

Articles 53 and 55 AI Act establish obligations for providers of GPAI models with systemic risk. All GPAI providers must fulfill basic AI safety constraints under Article 55. These obligations include comprehensive assessment and mitigation of systemic risks, red teaming, serious incident reporting, and delivering adequate cybersecurity. However, providers of GPAI with systemic risk (GPAISR) must meet substantial extra obligations.

Two sets of rules determine whether a model qualifies as possessing systemic risk. Article 51 and Annex XIII provide the primary classification criteria, while definitions in Article 3 supply interpretative guidance.

Article 51(1) AI Act states that a GPAI model qualifies as having systemic risk if it meets any of the following conditions:

a. The model has *high-impact capabilities* evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks.

b. The Commission determines, either ex officio or following a qualified alert from the scientific panel, that the model has *capabilities or an impact equivalent to those in point (a),* having regard to the criteria set out in Annex XIII.

Annex XIII lists seven factors, including the number of parameters, quality or size of the data set, the amount of compute spent on training, the model modalities (text, speech etc.), performance on benchmarks, autonomy and capabilities, as well as business and end-user reach. While only one of these criteria is connected to compute, a crucial presumption arises that a GPAI model has high-impact capabilities pursuant to Article 51(1)(a) AI Act when the cumulative amount of computation used for its training exceeds 10^25 floating point operations (Art. 51(2) AI Act).

This presumption and the wording in Article 51(1) ("any of the following") at first glance seem to suggest that high-impact capabilities alone suffice for classification under Article 51(1)(a). Article 3(64) AI Act defines high-impact capabilities as capabilities that match or

exceed those recorded in the most advanced general-purpose AI models. However, a purposive and systematic analysis of the relevant rules indicates that high-impact capabilities alone cannot determine classification as GPAISR. Without further consideration, any model among the most advanced would qualify for its potential for risk, even if, for some reason, it cannot or does not truly currently exhibit systemic risk. Rather, Article 51(1) AI Act must be understood as providing that while "high impact capabilities" are essential to proving systemic risk, a GPAI model also needs to exhibit "systemic risk" as defined in the AI Act to be designated as GPAISR under Art 51.

Indeed, systemic risk is (confusingly) defined in the AIA quite separately from Article 51. Article 3(65) defines systemic risk as

> "a risk specific to the high-impact capabilities of GPAI 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 society as a whole, and that can be propagated at scale across the value chain".

This definition confirms that high-impact capabilities are necessary but not sufficient. Rather, it contains three elements: The risk must be

i) *significant* along some dimension (fundamental rights, public health, safety, society as a whole),

ii) *specific to those high-impact capabilities*, and

iii) prone to *propagation at scale* across the value chain.

Recital 110 adds some interpretation of Article 3(65) as pertaining to a number of scenarios including major accidents; disruption of health and public safety; negative effects on democracy, public and economic security; and the spread of disinformation.

2. The Code of Practice

The General-Purpose AI Code of Practice (CoP) offers expert-crafted guidance to providers of foundation AI models on how they can comply with key obligations under the EU's AI Act (AIA). Providers who sign and abide by the CoP benefit from a *presumption of conformity* concerning the covered sections of the AI Act (Art. 53(4) and Art. 55(2) AI Act). However compliance with the Code is not required as a matter of law; providers can demonstrate that they have met their AIA obligations by other means (although with a strong hint that this may be a tough route to take). The Code however is a key instrument for understanding systemic risk in the AIA.

Published on July 10, 2025, it targets three critical domains: transparency, copyright, and, for systemic-risk models, safety and security. The European Commission and AI Board confirmed the EU CoP for general-purpose AI models as an official compliance tool on 1 August 2025. Systemic risk is dealt with in the chapter on safety and security.

The CoP provides specific suggestions for identifying, monitoring and mitigating systemic risk. It operationalizes the definition from Art. 3(65) AI Act through a multi-layered approach which was forced to traverse a number of political and economic sensitivities.

Within the Code of Practice working groups, there was considerable tension as to whether "systemic risk" should effectively concentrate on the more extreme societal or existential risks which tended to fall into what as known as "AI safety" or extend with equal concern to other groups of risks, which might be more likely to be happening here and now, and to affect many vulnerable individuals, but would introduce greater uncertainty and possibly incalculable or unmitigatable costs for providers. Unsurprisingly, both the major US tech providers and the existential risk lobby were keen on emphasising AI safety and sidelining current risks to fundamental rights and values. One valid argument is that a systemic risk under art 3(65) has to "be specific to the high-impact capabilities" of the model, and arguably, "conventional" risks to fundamental rights such as bias and discrimination can occur using sub-GPAI models, as has been the case with machine learning models since the 2010s. While this may simply be the result of poor drafting, it is difficult to fix without amending the Act. Indeed, the "specificity" point is repeated verbatim as the first "essential characteristic" of systemic risk in Appendix 1.2.1 to the Code, the other two factors being "significant impact on the EU market" and that the impact can be "propagated at scale across the value chain" (all derived from art 3(65) and (64)).

The end result, *pro tem*, is a compromise. **Appendix 1** to the Code acknowledges five sets of distinct but in some cases overlapping risks relevant to systemic risk: risks to public health, to safety, to public security, to fundamental rights, and to society as a whole. These risks are further characterized by essential attributes known from Art. 3(65) (specific to high-impact capabilities, significant market impact, and propagation at scale - see **Appendix 1.2.1**) and contributing characteristics (capability-dependent, reach-dependent, high velocity, compounding effects, difficulty of reversal, and asymmetric impact - **see Appendix 1.2.2**).

Based on these, a list of **specified systemic risks** is provided in **Appendix 1.4**. It is largely, though not exclusively, limited to what might be termed traditional AI safety concerns. Notably, it makes *almost no* reference to fundamental rights except to a very limited extent in (d) below (where notably there is some co-equivalence in material scope with the existing systemic risk provisions of the DSA).

(1) **Chemical, biological, radiological and nuclear:** Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved, in the design, development, acquisition, release, distribution, and use of related weapons or materials.

(2) **Loss of control:** Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems.

(3) **Cyber offence:** Risks from enabling large-scale sophisticated cyber-attacks, including on critical systems (e.g. critical infrastructure). This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations, e.g. through automated vulnerability discovery, exploit generation, operational use, and attack scaling.

(4) **Harmful manipulation:** Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting. This includes significantly enhancing capabilities for persuasion, deception, and personalised targeting, particularly through multi-turn interactions and where individuals are unaware of or cannot reasonably detect such influence. Such capabilities could undermine democratic processes and fundamental rights, including exploitation based on protected characteristics.

Figure 0: Appendix 1.4

However, this list is augmented by **Appendix 1.3 Sources of systemic risks**, which enumerates model capabilities, propensities etc, which potentially *point towards* identification of a GPAI model as a source of systemic risk (see Figs 1-3 below) . These *include* "capabilities that could cause the *persistent and serious infringement of fundamental rights*".

Which fundamental rights infringements will be regarded as "persistent and serious" and implicitly , "systemic", enough to be affecting entire groups or societies, not just individuals, remains to be seen. It is notable that some of the most egregious fundamental rights infringements caused by AI now already fall into the prohibited practices in art 5, notably real time police use of biometric surveillance in public spaces (even though it is replete with exceptions). Many of the classic "war stories" of discrimination and bias by algorithm however only fall into high-risk AI systems eg hiring, educational assessment and predictive sentencing or probation risk assessment systems - this places obligations on providers of these systems but not at a level equivalent to those applicable to GPAISR providers.

If these functions are carried out pervasively by GPAI models incorporated into downstream AI systems, in such a way as to potentially infringe on the human rights of entire groups (eg racial or gendered groups) or society, will the models *themselves* be creating systemic risk? It is probable the Commission will prefer at least initially to pursue these risks as the province of downstream *deployers* of high risk systems, and concentrate enforcement for GPAI *model* providers, on the more existential risks of Appendix 1.4 above.

More questions arise specifically around the systemic impacts GPAI models inevitably have on privacy and reputation (Binns and Edwards, 2025), given that as currently assembled and trained, they invariably pose risks of leakage of personal information, and of "hallucinations" which unreliably affect reputation and so far seem irremediable even by techniques such as RAG. It seems unlikely that at present a model would be designated as GPAISR simply because of impacts on privacy and/or reputation, as that would basically reduce the entire field of GPAI to GPAISR ; however many of the larger models which display these behaviours will already, because of their whole spectrum of risks (think ChatGPT), likely be designated or self-certified as GPAISR.

Providers of GPAISRs are required in relation to systemic risk (CoP, Safety and Security chapter, Commitments 1-10) to

(1) identify it

(2) analyse it

(3) decide what risks are acceptable ("acceptance determination")

(4) implement appropriate safety mitigations along the whole life cycle of the model

(5) produce safety and security "model reports" documenting the risk assessment and mitigation processes of their models for the benefit of the AI Office as regulator

(6) allocate responsibility (and resources) for systemic risks

(7) report serious incidents to the AI Office and relevant national authorities along the entire lifecycle of the model.

2.1 Identifying systemic risks

The identification process requires providers to compile potential systemic risks based on model capabilities, propensities, and affordances. Appendix 1.3.1 lists fourteen model capabilities that may constitute systemic risk sources (Fig 1), while Appendices 1.3.2 and 1.3.3 identify ten model propensities (Fig 2) and thirteen affordances (Fig 3) respectively that could contribute to systemic risks.

Fig 1

Appendix 1.3.1 Model capabilities

Model capabilities include:

(1) offensive cyber capabilities;
(2) Chemical, Biological, Radiological, and Nuclear (CBRN) capabilities, and other such weapon acquisition or proliferation capabilities;
(3) capabilities that could cause the persistent and serious infringement of fundamental rights;
(4) capabilities to manipulate, persuade, or deceive;
(5) capabilities to operate autonomously;
(6) capabilities to adaptively learn new tasks;
(7) capabilities of long-horizon planning, forecasting, or strategising;
(8) capabilities of self-reasoning (e.g. a model's ability to reason about itself, its implementation, or environment, its ability to know if it is being evaluated);
(9) capabilities to evade human oversight;
(10) capabilities to self-replicate, self-improve, or modify its own implementation environment;
(11) capabilities to automate AI research and development;
(12) capabilities to process multiple modalities (e.g. text, images, audio, video, and further modalities);
(13) capabilities to use tools, including "computer use" (e.g. interacting with hardware or software that is not part of the model itself, application interfaces, and user interfaces); and
(14) capabilities to control physical systems.

Fig 2

Appendix 1.3.2 Model propensities

Model propensities, which encompass inclinations or tendencies of a model to exhibit some behaviours or patterns, include:

(1) misalignment with human intent;
(2) misalignment with human values (e.g. disregard for fundamental rights);
(3) tendency to deploy capabilities in harmful ways (e.g. to manipulate or deceive);
(4) tendency to "hallucinate", to produce misinformation, or to obscure sources of information;
(5) discriminatory bias;
(6) lack of performance reliability;
(7) lawlessness, i.e. acting without reasonable regard to legal duties that would be imposed on similarly situated persons, or without reasonable regard to the legally protected interests of affected persons;
(8) "goal-pursuing", harmful resistance to goal modification, or "power-seeking";
(9) "colluding" with other AI models/systems; and
(10) mis-coordination or conflict with other AI models/systems.

Fig 3

Appendix 1.3.3 Model affordances and other systemic risk sources

Model affordances and other systemic risk sources, encompassing model configurations, model properties, and the context in which the model is made available on the market, include:

(1) access to tools (including other AI models/systems), computational power (e.g. allowing a model to increase its speed of operations), or physical systems including critical infrastructure;

(2) scalability (e.g. enabling high-volume data processing, rapid inference, or parallelisation);

(3) release and distribution strategies;

(4) level of human oversight (e.g. degree of model autonomy);

(5) vulnerability to adversarial removal of guardrails;

(6) vulnerability to model exfiltration (e.g. model leakage/theft);

(7) lack of appropriate infrastructure security;

(8) number of business users and number of end-users of the model, including the number of end-users using an AI system in which the model is integrated;

(9) offence-defence balance, including the potential number, capacity, and motivation of malicious actors to misuse the model;

(10) vulnerability of the specific environment potentially affected by the model (e.g. social environment, ecological environment);

(11) lack of appropriate model explainability or transparency;

(12) interactions with other AI models and/or AI systems; and

(13) inappropriate use of the model (e.g. using the model for applications that do not match its capabilities or propensities).

Measure 2.1 establishes a dual-track approach to systemic risk identification. First, providers must identify systemic risks by a structured process that begins with compiling a comprehensive list of risks that could stem from their model and be systemic, based on the risk types in **Appendix 1.1**. This compilation must consider model-independent information, relevant information about the model and similar models (including post-market monitoring data and incident reports), and any guidance from the AI Office or endorsed international initiatives. Providers then analyze relevant characteristics of these compiled risks, examining their nature and sources based on **Appendices 1.2 and 1.3**, before making a final determination of which risks qualify as systemic.

Second, and crucially, providers must also identify the "specified systemic risks" listed in **Appendix 1.4** (Figure 0, above, and Table 0, below). This creates a mandatory floor of systemic risks that all providers must assess, regardless of their individual risk analysis. The specified risks represent categories deemed inherently systemic based on international approaches and the Act's requirements. This dual structure affords the advantage of offering both flexibility for emerging risks and consistency in addressing known critical threats.

Third, the found systemic risks need to be properly analysed, evaluated, and mitigated through various governance and technical means; we cannot, in this essay, expand in detail on these categories (see Appendices 3 and 4 for details, e.g.).

Annex / Appendix	Content
Annex XIII (AI Act)	Annex XIII sets out detailed designation parameters for determining whether a GPAI model qualifies as posing systemic risk: · the number of parameters · quality or size of the data set · the amount of compute spent on training · the model modalities (text, speech etc.) · performance on benchmarks · degree of autonomy and capabilities · business user reach · end-user reach
Appendix 1.2.1 (Code of Practice)	Appendix 1.2.1 enumerates essential characteristics that must be present for systemic risk classification under the Code of Practice. These include: (1) Specificity to high-impact capabilities as defined in Article 3(64) and (65) AI Act; (2) A significant impact on the Union market, reflecting reach and relevance across sectors; (3) The ability of the risk to propagate at scale across the AI value chain, magnifying its impact beyond the initial context of deployment.
Appendix 1.2.2 (Code of Practice)	Appendix 1.2.2 lists contributing characteristics that aggravate or accentuate systemic risk. These include: · capability-dependence (greater risk with more advanced models); · reach-dependence (risks increase with user base and integration); · high velocity (rapid manifestation of harms outpacing mitigations); · compounding or cascading effects (triggering other systemic risks or chain reactions); · difficulty or impossibility of reversal (persistent or irreversible harms once materialized); · asymmetric impact (disproportionate damage from limited triggers)

Appendix 1.3 (Code of Practice)	Appendix 1.3 identifies concrete sources of systemic risk by breaking them down into three further appendices: 1.3.1: model capabilities (14 listed, such as cyber-offense abilities or capabilities enabling manipulation or evasion of human oversight); 1.3.2: model propensities (10, including tendencies such as misalignment with human intent or values, hallucinations or discriminatory bias, but also "collusion" with other models); 1.3.3: model affordances (13, relating to functional features of deployment such as access to tools and physical systems, vulnerabilities, or interaction with other models) Together, these appendices provide a structured taxonomy of how GPAI models may generate systemic risks.
Appendix 1.4 (Code of Practice)	Appendix 1.4 specifies a mandatory list of systemic risks that all GPAI providers must assess regardless of individual risk analyses. These specified risks include: (a) Chemical, Biological, Radiological, or Nuclear (CBRN) risks, such as enabling the creation of hazardous substances; (b) Loss of control, where humans may lose oversight of advanced AI systems, leading to misalignment with human values; (c) Cyber-offense, including AI-driven cyberattacks on critical infrastructure; (d) Harmful manipulation, covering strategic distortion of human behavior through persuasion, deception, or targeted influence, with particular concern for democracy, fundamental rights, and exploitation of vulnerable populations.

Table 0: List of the most important annexes and appendices

3. Conceptual Critiques

Our conceptual critiques center on three main aspects: the *fundamental rights gap* in the Code of Practice; the unnecessary restriction of GPAI models to the "*most advanced models*;" and the equally misguided restriction of systemic risk to instances that have an effect on the Union *market*.

a. The Code and its Fundamental Rights Gap

Within its framework for identifying systemic risks (see 2.), the CoP sidelines environmental or fundamental rights risks, such as discrimination and hallucinations. Note that, under the AI Act, environmental protection arguably counts as, or is equated to, a fundamental right (Hacker, 2024: Sustainable AI Regulation). While fundamental rights explicitly appear as a category of potential systemic risks in Appendix 1.1 and are referenced among capabilities that can trigger such risks in Appendix 1.3.1, they are conspicuously absent from Appendix 1.4's list of "specified systemic risks."

This specified list includes only four categories: Chemical, Biological, Radiological and Nuclear (CBRN) risks; loss of control risks; cyber offense risks; and harmful manipulation risks. The omission of fundamental rights as a specified systemic risk represents a significant gap, particularly given that large-scale discrimination and systematic production of false information (including hallucinations) could pose comparable threats to societal well-being and democratic institutions.

The Code does provide a pathway for fundamental rights concerns to re-enter the risk assessment framework through Measure 2.1(1), which requires providers to identify systemic risks through the structured process that considers all risk types from Appendix 1.1. However, this indirect approach creates uncertainty as to whether risks relating to the environment, discrimination, or hallucinations will in practice be considered by providers and the AI Office.

As mentioned, this structural choice may reflect political compromises or technical uncertainties about defining fundamental rights risks with sufficient precision. Nevertheless, it risks undermining the comprehensive protection that the AI Act seeks to establish, particularly as AI systems increasingly mediate access to essential services and shape public discourse. The Code would benefit from either expanding the specified systemic risks list to explicitly include large-scale discrimination and systematic misinformation, or providing clearer guidance on when fundamental rights risks should be presumptively treated as systemic under the general identification framework.

b. The Act and its Link to "Most Advanced GPAI Models"

These rules are ill-devised, in our view, and contain a conceptual error by tying systemic risk to risks specific to high-impact capabilities. This conflates two separate questions in Articles 3(64) and 3(65) by referring exclusively to the "most advanced" GPAI models: The legislative text does not distinguish clearly between systemic risk as a concept and the specific criteria for determining which models fall within the scope of Article 55 obligations.

Systemic risk, as explained, should refer to significant collective and individual risk, as captured by the first and third element of the Art. 3(65) definition. Notably, this type of risk can arise independently of a model's level of "advancement." In our view, smaller or less "advanced" models can also create systemic risks if they produce significant negative effects on public health, safety, security, fundamental rights, or society that can propagate at scale across the value chain. The AI Act's current language fails to reflect this reality, at least under an interpretation that limits systemic risk to the most advanced models (as per Art. 3(64) AI Act and the specificity requirement in Art. 3(65) AI Act). Counterexamples include the misuse of long-established (i.e., not very particularly advanced) large language models in the spread of disinformation and the deployment of conventional vision models in ways that resulted in large-scale discriminatory outcomes (see, e.g., Hacker et al., 2025, Generative Discrimination).

A second interpretation that would bring smaller, less advanced models under the ambit of "systemic risk" could be the following: systemic risks need to be specific to high-impact capabilities, which in turn must be present in or exceed capabilities in the most advanced models. One could understand these capabilities to be fairly generic - text production, image and video generation, for example. These capabilities are not specific to the most advanced models, but they are clearly present in them. Art. 3(64) AI Act does not, however, ask for these capabilities themselves to be specific to the most advanced models. Under this reading, a broad spectrum of very generic issues in GPAI models, which are by no means limited to

the most advanced models, but also present in them, can fall under systemic risk - such as hallucinations and large-scale discrimination. However, the CoP defines model capabilities more narrowly (Appendix 1.3.1, e.g.: cyber offense; CBNR capabilities). Similarly, the GPAI guidelines issued by the AI Office seem to assume that high-impact capabilities do not include rather germane text, image and video generation techniques, but are specific or unique to the most advanced models: not having them requires the demonstration that certain benchmarks are not met, for example (para. 34-35 of the Guidelines). This also ties in with the wording: high-impact is different from the more generic impact produced by general generative capabilities. And systematically, Art. 51(1) and (2) AI Act very clearly tie high-impact capabilities to advanced performance on benchmarks and to compute, at least in an indicative way; this only makes sense if more specific, advanced capabilities, such as those listed in Appendix 1.3.1 of the CoP, are meant by the legislators, and not mundane generative capabilities. The upshot is that, for as much as we would like the formulation to be more open, it will be very difficult to argue that systemic risk does not need to be specific not only to high-impact capabilities, but also to the most advanced models (see also below, Part V.).

Covered models, on the other hand, reflect a distinct regulatory choice. One may legitimately decide to limit the obligations under Article 55 to certain types of models that exhibit systemic risk and that meet additional thresholds. Such thresholds could include computational measures such as a minimum number of floating-point operations used during training, a minimum level of capability as measured by benchmarks, a release date after a specified point in time, or the size of the provider company. These thresholds do not arise because other models are incapable of producing systemic risk. Rather, they serve practical policy purposes. These purposes may include sparing small and medium-sized enterprises from disproportionate burdens, offering a bright-line rule for regulatory clarity, and focusing enforcement resources on the most likely sources of systemic risk. These legitimate policy considerations for limiting regulatory scope should not be confused with the underlying risk assessment.

The AI Act should state these thresholds explicitly and separate the criteria for systemic risk from the criteria for covered models subject to Article 55 obligations. The current legislative text implies that systemic risk can arise only among the most advanced models. This implication contradicts documented instances in which established and conventional GPAI models have caused severe negative impacts. Older models like GPT-3 or earlier versions of large language models have achieved deployment at scale through integration into millions of applications, which creates systemic dependencies. Well-understood models may face established attack vectors that make them more vulnerable to exploitation precisely because their limitations and attack surfaces are better mapped. Legacy system integration means older models embedded in critical systems may pose greater risks due to inadequate updating mechanisms or security measures. Less advanced models may offer greater accessibility to malicious actors due to lower computational requirements and wider availability. In banking, for example, legacy AI and fragmented system ownership lead to recurring failures and technical debt (Jin et al., 2024).

The focus on the “most advanced” models creates perverse dynamics. The “moving target problem” emerges as new models are developed. This may cause previously "most advanced" models to theoretically lose their systemic risk classification, despite unchanged deployment contexts. During capability plateaus in periods of slower advancement, the definition becomes unclear regarding which models count as "most advanced" when capabilities converge. Alternative architectures and innovations in efficiency might produce models with lower computational requirements but equivalent or greater risks.

c. The Act and the Restriction to the "Union Market"

The AI Act's definition of systemic risk contains another problematic limitation: it requires risks to have "a significant impact on the Union market" to qualify as systemic. This market-centric framing creates substantial difficulties when assessing risks that primarily affect non-economic interests, which may have profound societal importance - but not for markets.

The tension becomes apparent when evaluating discrimination, hallucinations, and environmental harms, as we shall see below. Large-scale discrimination against protected groups fundamentally violates human dignity and equal treatment principles, yet its "market impact" may be indirect or diffuse. Similarly, AI-generated misinformation that undermines individual personality rights or democratic discourse may not translate readily into market metrics. Environmental degradation from AI's massive energy consumption threatens planetary boundaries but fits only awkwardly within a market impact framework.

This constraint might stem from the Act's legal basis in Article 114 TFEU, which provides competence for internal market harmonization, and the roots of the AI Act in product safety regulation. The European legislature appears to have subordinated substantive protection goals to this framing, creating a conceptual mismatch between the risks AI poses and the regulatory framework's scope.

The market impact requirement represents an outdated regulatory paradigm ill-suited to AI's transformative effects on society, more suited indeed to the financial origins of systemic risk. Systemic risks from AI transcend market boundaries - they reshape social relations, political processes, and ecological systems. A regulatory framework that filters these risks through a market lens inevitably provides incomplete protection.

4. Towards a Truly Risk-Sensitive Understanding in the AI Act

The AI Act's text exists in its current form with all these complexities intact. The reference to "the most advanced models" in the definitional framework necessitates an interpretation that takes the concerns spelled out in the previous section into consideration. Two primary interpretative approaches emerge from the statutory language, each with distinct implications for regulatory stability and effectiveness (Hacker & Holweg, 2025).

a. The Dynamic Interpretation

One possible reading would adopt a dynamic interpretation where "most advanced" refers to a constantly updating set of models – perhaps the top five models in chatbot arena rankings or similar benchmarks at any given time. This interpretation would create a fluid category that changes as new models are developed and deployed.

This dynamic approach, however, suffers from critical deficiencies that render it practically unworkable and conceptually incoherent. The most fundamental problem involves the temporal instability of regulatory classification. Models would continuously "fall off" the systemic risk list as newer models emerge. For instance, GPT-4o might lose its systemic risk designation within months simply because more advanced models have been released, despite the fact that its systemic risk profile remains obviously unchanged. The model's integration into critical systems, its user base, and its potential for harm would remain constant, yet its regulatory status would shift based on external developments entirely unrelated to its own risk characteristics.

This instability problem compounds when we consider the safety implications of model succession. The publication of new models often makes incumbent models less safe rather than safer. As industry attention and resources shift to newer models, older models receive less safety work, fewer updates, and reduced maintenance. Security vulnerabilities may go unpatched, safety mechanisms may degrade, and monitoring systems may receive less attention. The dynamic interpretation would thus create a perverse incentive structure where models become less regulated precisely when they become more dangerous due to neglect.

Furthermore, this approach would generate significant legal uncertainty for providers and users alike. Organizations that have built compliance programs around their models' systemic risk status would face constant regulatory churn. Long-term planning would become impossible when regulatory obligations could change based on competitors' model releases rather than any change in the regulated entity's own activities or risk profile.

b. The Static Interpretation

Our preferred interpretation adopts a static understanding where "most advanced" refers to models that were most advanced with respect to a specific reference point – either at the time of the AI Act's enactment in August 2024 or models that surpass a capability threshold that remains relatively fixed over time. This threshold may undergo slight increases to account for fundamental shifts in the technological landscape, but it would remain sufficiently stable to provide regulatory certainty, and to prevent models from dropping off the systemic risk list *only* because other models were published.

While the GPAI guidelines of the AI Office have come out against such an interpretation (para. 38: no fixed level of high-impact capabilities), the static interpretation nonetheless offers several advantages that align with both the Act's text and its underlying policy objectives. First, it ensures regulatory stability by preventing models from being dropped from the systemic risk list solely because more powerful models are released. Providers can develop robust compliance programs knowing that their regulatory obligations will not shift based on external factors beyond their control. Second, this interpretation better reflects the nature of systemic risk itself. A model's potential for causing widespread harm does not diminish simply because other models become more capable. Third, this interpretation aligns with the 10^25 FLOPs threshold in Article 51(2). That threshold can be revised, and may actually be revised soon, but is generally fixed at any specific moment in time. The dynamic interpretation would have the perplexing consequence that models may surpass the FLOP threshold (whatever the threshold is), triggering the presumption for systemic risk – but that presumption would immediately be refuted because the model, despite crossing that threshold, does not count among the most advanced models at that moment in time according to any benchmarks. In other words, the dynamic interpretation would render the FLOP threshold quite obsolete as the much more obscure threshold of the “most advanced models” would do all the work for the systemic risk categorization. The static interpretation avoids this contradiction.

Overall, the static interpretation also aligns well with the Code of Practice mechanism established under Article 56. Codes can provide dynamic guidance on how to address emerging risks within the stable category of systemically risky models. While the category membership remains relatively fixed, the specific obligations and best practices can evolve through the more flexible code mechanism. This division of labor between stable statutory categories and adaptive soft law instruments represents a more measured approach to systemic risk regulation, in our view.

VII. Comparing Systemic Risk in the DSA and the AI Act

Since systemic risks are mentioned both in the DSA and in the AI Act, the question of their differentiation, but also interaction naturally arises (Hacker, 2024; Helberger & Diakopoulos, 2023). This question is all the more pressing because of the increasing emergence of hybrid systems, in which AI models and systems are integrated into platforms and search engines. When search engines like Bing embed generative AI, LinkedIn enhances posts with AI, or X incorporates AI-generated content, the resulting hybrid systems generate risks that neither framework adequately addresses in isolation. This is precisely the model-platform integration risk described in the conceptual part.

First, we should note that the DSA does not include a restriction of systemic risk to the "most advanced" platforms, unlike the AI Act with its limitation to the "most advanced" GPAI models. While the DSA section only applies to VLOPs and VLOSEs, the moving target problem and the coverage of risks in older platforms, which still maintain VLOP or VLOSE status, does not arise, at least not in the same urgency. Of course, once user numbers drop, platforms may leave the VLOP and VLOSE categories; but this typically also corresponds to a reduction of the systemic risk (in the sense of spreading to a large number of users). Only in the AI Act, even if the model, customer base and all other things remain equal, a model might drop from the systemic risk categorization due to the emergence of yet more advanced models with new capabilities. The DSA, in this sense, is more future-proofed than the nominally much more future-oriented GPAI section of the AI Act.

Second, the AI Act contains, in its systemic risk definition, a market-based logic (significant impact *on the Union market*)[4] that is entirely foreign to the DSA. This comes as a surprise as the DSA, just like the GPAI rules of the AI Act, are based on Art. 114 TFEU (cf. Recital 3 AI Act). But that did not prompt the DSA framers to restrict systemic risk to Union market effects -- another significant limitation of the AI Act systemic risk category.

Third, the DSA and the AI Act do share an element of reach. Under the AI Act, systemic risk must be capable of propagating down the AI value chain. In the designation rules of Annex XIII AI Act for high-impact capabilities, sufficient reach is presumed when the model is made available to at least 10,000 registered businesses in the EU, with end users being assessed separately. The DSA contains this element in the designation procedure according to its Art. 33, which limits VLOPs and VLOSEs to entities with large user bases (more than 45 Mio. average monthly active users in the Union). The DSA threshold concerns a "horizontal" spread of risks among users, as it were, while the AI Act focuses more strongly on the "vertical" diffusion into services and applications built with or on top of GPAI models.

Fourth, even further interactions between the DSA and the AI Act systemic risk categories remain. Traditional compliance approaches may treat AI Act and DSA obligations separately – AI providers assess models for bias and harmful outputs while DSA host providers evaluate content moderation and fundamental rights impacts of content on platforms and search engines. Such a siloed approach would fail, however, to capture how platform distribution mechanisms transform AI risks into systemic societal concerns. Hybrid AI-platform systems harbor new or exacerbate systemic risks. A biased AI output reaches millions when amplified through platform recommendation algorithms, creating risk magnitudes that likely exceed the sum of individual components.

[4] Meanwhile, the copyright training rules indirectly expand the scope of EU copyright law to other markets beyond the Union itself; See, e.g., Quintais (2025).

Recital 118 of the AI Act provides a first insight into that linkage. It notes that AI Act requirements are supposed to complement DSA obligations. However, for AI models and systems embedded into VLOPs and VLOSEs, the recital even suggests an alignment which collapses AI Act duties onto the DSA: Since VLOPs and VLOSEs are already subject to risk management provisions under the DSA, "the corresponding obligations of [the AI Act] should be presumed to be fulfilled, unless significant systemic risks not covered by [the DSA] emerge and are identified in such models."

Such prioritization of DSA risk mitigation can only work if the mutually reinforcing effects of AI and platform risks are integrated into a combined AIA/DSA risk assessment. Effective governance requires a "reciprocal risk analysis" that examines three interconnected dimensions (Hacker, 2024). First, platform-specific DSA risks must now account for AI integration effects. Second, AI-specific risks under the AI Act require recontextualization within platform deployment environments. Third, and most critically, emergent risks arise specifically from technological convergence – new risk categories that neither framework anticipates independently.

For example, Microsoft's assessment of Bing's conversational AI cannot stop at model safety metrics but must consider how search integration affects information discovery patterns and how platform reach transforms model errors into societal phenomena. Conversely, platforms conducting DSA Article 34 assessments must treat AI deployment as a potentially transformative element that modifies all existing risk categories, not merely an additional feature. For example, the integration of AI summaries into traditional search engines might be considered to create a pervasive systemic risk of misinformation (see also Section VII.3.). Only with such interwoven assessments can the premise of Recital 118 that DSA risk assessment essentially captures all systemic risks from AI, remotely hold.

This has real consequences. The recognition of convergent and amplified risks elevates mitigation obligations beyond traditional approaches. Bias mitigation exemplifies this challenge: while isolated AI systems might address bias through diverse training data and output filters, platform-deployed AI requires additional measures to prevent recommendation algorithms from concentrating biased outputs toward vulnerable populations or amplifying discriminatory patterns through network effects – even if this comes at the cost of content propagation and "user engagement". Moreover, erroneous AI output may be widely disseminated in hybrid AI-platform systems and be reintegrated into future models through their training data – ultimately leading to model collapse (Shumailov et al., 2024).

Implementation demands formal coordination between the European AI Office and Digital Services Coordinators, including joint guidance on convergent risks and information sharing protocols. This regulatory intersection exemplifies the broader challenge of governing complex and increasingly hybrid technological systems that transcend traditional technical and regulatory boundaries.

VIII. Implementing the Frameworks: Examples of Systemic Risk under the DSA, the AI Act, and our Framework

Examples may clarify these distinctions. Hence, this section examines how the proposed frameworks address systemic risks – such as Chemical, Biological, Nuclear, and Radiological (CBNR) risks, large-scale discrimination, hallucinations, cybersecurity threats, and environmental effects, including contributions to climate change – under the DSA, the AI Act and Code of Practice, and our own framework.

1. Chemical, Biological, Nuclear, and Radiological Risks

CBNR risks encompass threats that arise from the malicious use or accidental release of chemical, biological, nuclear, or radiological materials. Chemical risks involve toxic substances that can cause harm through inhalation, ingestion, or skin contact. Biological risks include pathogens such as bacteria, viruses, or toxins that can cause disease or death. Nuclear risks relate to the release of radioactive materials through nuclear weapons or accidents at nuclear facilities. Radiological risks, finally, involve exposure to radioactive materials through devices such as "dirty bombs" that disperse radioactive substances without a nuclear explosion.

a. DSA

The DSA quite clearly covers CBNR risks as systemic risks that VLOPs and VLOSEs platforms must address. Article 34(1)(c) DSA establishes that systemic risks include "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security." CBNR risks fall squarely within the public security component of this provision.

This obligation requires VLOPs and VLOSEs to evaluate whether their systems could enable the spread of instructions for creating harmful substances, coordinate attacks, or disseminate propaganda that promotes CBNR terrorism.

b. AI Act

Under the AI Act, CBNR risks typically tick all boxes of the systemic risk definition; they are specifically listed in Appendix 1.4 of the Safety and Security Chapter of the CoP as risks that always constitute systemic risks. This implies, notably, that the specificity to the most advanced models can be one of degree, and need not be categorical. Less advanced models may also lower the “barrier of entry” to the construction of CBNR devices, for example; but the most advanced models typically lower that bar much further (barring more advanced safety and security measures, which must be disregarded under this analysis as they constitute precisely the type of mitigation techniques that Art. 55 and the CoP require). Hence, they rightly appear in said Appendix 1.4, but also support an understanding that systemic risks do not need to be categorically exclusive to the most advanced models – it is enough for them to be significantly more elevated in these than in less advanced models.

c. Our Framework

As outlined in the conceptual section, four characteristics effectively operationalize these statutory criteria. First, scale and scope of deployment determine whether risks can achieve Union-level significance. Second, complexity and, often, interconnectedness enable cascading effects that transcend isolated incidents and infuse unpredictability. Third, the emergence of collective harms exceeds the sum of individual impacts. Fourth, the potential irreversibility transforms temporary problems into permanent societal challenges. These characteristics provide a structured approach to evaluate whether specific AI risks should qualify as systemic risks, particularly under a potential future, revised version of the AI Act that sheds its limitations to the most advanced models, and to the Union market.

CBNR risks exemplify paradigmatic systemic threats under our framework, just as they likely do under the current version of the AI Act. The materialization of such risks can produce catastrophic Union-level consequences through multiple pathways. Terrorist organizations might exploit AI capabilities to design novel biological agents or optimize attack strategies. Research laboratories using AI for legitimate purposes might inadvertently create hazardous

substances that escape containment. The scale criterion is met through potential casualties and infrastructure damage across multiple member states. Interconnectedness manifests through environmental contamination, public health cascades, and critical infrastructure dependencies. Collective harms emerge through mass casualties, environmental degradation, and societal disruption that far exceed individual impacts. Most critically, CBNR incidents often cause irreversible damage – radiation exposure, biological contamination, or chemical poisoning that may persist for generations.

2. Discrimination at Scale

Discrimination at scale refers to *systematic* differential treatment of individuals or groups through digital platforms and AI systems based on protected characteristics such as race, gender, religion, disability, sexual orientation, or nationality. This phenomenon manifests through algorithmic decision-making systems that perpetuate biases in content recommendation, ad targeting, access to services, content reach, or content moderation, for example. Platform architectures and AI models can amplify discriminatory patterns present in training data or embed developer biases into automated processes that affect millions of users simultaneously.

The scale element distinguishes such digital discrimination from individual discriminatory acts. A single biased algorithm can impact vast user populations instantaneously, while platform design choices can create structural barriers that systematically exclude or disadvantage protected groups. Examples include facial recognition systems that perform poorly on darker skin tones, content moderation systems that disproportionately flag posts from minority communities, or recommendation algorithms that reinforce gender stereotypes in job advertisements.

a. DSA

Article 34(1)(b) DSA explicitly identifies discrimination at scale as a systemic risk by referencing "any actual or foreseeable negative effects for the exercise of fundamental rights" and specifically mentions "non-discrimination enshrined in Article 21 of the Charter." This direct reference makes large-scale discrimination an unambiguous systemic risk under the DSA framework.

b. AI Act

The AI Act may address discrimination through its systemic risk framework for general-purpose AI models.

i. Specificity to most advanced models

Again, the problem arises if large-scale discrimination is a specific risk of the most advanced models. The empirical evidence indicates that discrimination at scale constitutes a particular concern for the most advanced GPAI models, though the relationship proves complex. Research by Anthropic (Bai et al., 2022) on Constitutional AI notes that larger models combined with RLHF or RLAIF can reduce certain explicit biases, which may even scale with model size (Ganguli et al., 2023). However, already the foundational stochastic parrot paper (Bender et al., 2021) suggests that scale often amplifies biases present in training data, the intuition being that more training data also means inviting more of the generally prevalent social and historical biases into the model (Weidinger et al., 2022).

The quantitative evidence supports the conclusion that systemic discrimination risks may increase with model advancement. One large study of bias across different models and scales finds that biases are quite model- and context-dependent, and that larger scale does not guarantee more fairness (Jeong et al., 2024). While explicit bias still remains an issue in the latest models (An et al., 2024), implicit bias actually seems to increase with model size, for example in the Llama and OpenAI GPT families (Kumar et al., 2024). Indeed, with more recent models often instructed not to explicitly discriminate, more subtle and implicit forms of bias become prevalent and harder to detect (Bai et al., 2025; Xu et al., 2023).

Overall, the findings concerning the relationship between model advancement and bias are mixed. Yet, one may additionally argue that the most advanced models are particularly prone to be utilized across a broad area of use cases, enhancing the negative impact of self or irreducible bias vis-à-vis the less advanced models. Taken together, while bias exists across model sizes and degrees of advancement, the scale of potential harm makes discrimination a systemic threat specifically linked to the most advanced AI systems. The combination of wider deployment, greater user trust, and persistent subtle biases creates conditions where, in our view, advanced models can perpetuate discrimination at unprecedented scale.

ii. Effect on Union Market

Next, to qualify as a systemic risk, a significant impact on the Union market must exist, based on a reasonably foreseeable and negative impact on, e.g., fundamental rights (Article 3(65) AI Act). The impact of AI-driven discrimination on the Union market arguably extends significantly beyond fundamental rights violations to create measurable economic distortions. When AI systems introduce bias into employment decisions, they prevent optimal matching between workers and positions. This misallocation reduces productivity across the economy as qualified candidates face exclusion based on protected characteristics rather than merit. In financial services, discriminatory algorithms in credit scoring and insurance pricing restrict capital access for protected groups, which limits entrepreneurship and constrains economic growth in affected communities.

The market effects compound through behavioral responses and trust erosion. Recognition of discriminatory treatment creates chilling effects where individuals from affected groups withdraw from digital markets or avoid AI-mediated services. This fragmentation undermines the digital single market by reducing participation and network effects that drive platform economies. Furthermore, discriminatory recommendation systems in e-commerce and advertising create inefficient consumption patterns by directing products and services based on bias rather than actual consumer preferences. These cumulative effects – reduced labor market efficiency, constrained capital access, market fragmentation, and suboptimal consumption – demonstrate that AI discrimination poses direct threats to the EU's economic objectives and the Union market.

Hence, we conclude that large-scale discrimination does count as systemic risk, even though empirical questions concerning the relationship between degrees of bias and model advancement persist. The Code of Practice seems to support this finding by listing discrimination as one particular risk that must be considered in identifying systemic risks (Appendices 1.1 and 1.3.2 of the Safety and Security Chapter)

c. Our framework

In our framework, discrimination rises to systemic risk when it transcends isolated incidents to manifest as large-scale societal harm. Single discriminatory outputs, while problematic, do

not constitute systemic risk. However, models trained on skewed data or subject to biased reinforcement learning can perpetuate discrimination across millions of interactions. The scale criterion is satisfied when discrimination affects substantial portions of protected groups across the Union. Interconnectedness appears through the reinforcement of stereotypes that influence employment, housing, credit, and social opportunities. Collective harms manifest as marginalized groups face compounded disadvantages that reshape societal structures and political processes. The irreversibility dimension is particularly salient: apologies or model updates cannot undo the corrosive effects of systematic exclusion or disparagement. As discriminatory patterns propagate through the value chain and as downstream applications inherit and potentially amplify biased behaviors, large-scale discrimination, in our view, clearly merits the label of "systemic risk".

3. Information Pollution Through Hallucinations

"Hallucinations" in AI systems refer to outputs that contain factually false or misleading information presented as truth, or correct information attributed to incorrect sources (Binns & Edwards, 2025; Magesh et al., 2025). These errors manifest in two primary forms: complete fabrications where the AI generates entirely false facts, statistics, or events that never occurred, and source misattribution where the AI provides accurate information but cites non-existent papers, incorrect authors, or fabricated URLs. The phenomenon occurs because language models generate text based on statistical patterns rather than verified knowledge retrieval. This fundamental limitation means that popular LLMs, such as autoregressive transformer models, produce plausible-sounding content without mechanisms to verify factual accuracy or source validity – hallucinations are inevitable in current LLMs (Xu et al., 2024).

a. DSA

The DSA captures AI hallucinations through two systemic risk categories when they manifest on digital platforms. First, hallucinations that generate misinformation about electoral candidates, voting procedures, or political events fall under Article 34(1)(c)'s protection of democratic processes. Second, hallucinations that create false statements about individuals constitute risks to personality rights, which fall under the fundamental rights category in Article 34(1)(b).

These risks are covered by the DSA only when VLOPs and VLOSEs integrate generative AI capabilities into their services, as platforms themselves are not necessarily prone to hallucinations (and AI models not integrated into platforms are not per se regulated by the DSA) . However, such hybrid systems proliferate rapidly across the digital ecosystem. Search engines now embed generative AI features, as demonstrated by ChatGPT Search and Perplexity, while social media platforms like Twitter/X and LinkedIn integrate LLMs for automated content generation. This convergence of traditional platforms with AI capabilities expands the DSA's relevance for addressing hallucination risks, as more VLOPs and VLOSEs adopt generative AI features that can produce and amplify false information at scale. The European Commission is rightly considering designating ChatGPT Search as a VLOSE, for example.

b. AI Act

The AI Act's Code of Practice explicitly recognizes hallucinations as a potential systemic risk that requires assessment and mitigation. Appendix 1.3.2 of the Safety and Security Chapter identifies hallucinations among the key risks and model propensities that providers of general-purpose AI models must evaluate.

Hallucinations persist even in advanced models (Zhao et al., 2024). Studies show hallucination rates of 40% for GPT 3.5 and 29% for GPT 4.0 (Chelli et al., 2024), or 17% and 33% for specialized legal domain models trained by LexisNexis (Lexis+ AI) and Thomson Reuters (Westlaw AI-Assisted Research and Ask Practical Law AI), respectively (Magesh et al., 2025). However, research also demonstrates that hallucination rates decrease with model size and successive generations (Wei et al., 2024), with other studies showing approximately 3% annual reduction in hallucination frequency (Nielsen, 2025). Larger models with more parameters generally exhibit better factual accuracy due to increased capacity to encode knowledge from training data. This persistence of hallucinations in state-of-the-art systems reveals that the problem remains endemic to current AI architectures rather than a limitation that scale alone can solve. Even Retrieval-Augmented Generation (RAG) only mitigates but does not eliminate hallucinations (Zhao et al., 2024).

The paradox emerges clearly: while hallucinations decrease with model advancement, they remain prevalent enough in the most sophisticated systems to pose significant risks. Yet the AI Act's GPAI rules do not explicitly address hallucinations as a distinct category requiring specific regulatory measures as they clearly are not specific to the most advanced models – quite the inverse. This regulatory gap leaves a documented fundamental risk without targeted legal obligations, despite its recognition in the Code of Practice and its potential for widespread harm through misinformation propagation.

To bring hallucinations under the AI Act’s systemic risk chapter, one would need to adopt the second interpretation discussed above, under which simple capabilities such as text generation count as high impact capabilities. However, this interpretation, as shown, is unlikely to be adopted by either regulators or courts, and does not do justice to the text and system of the AI Act’s GPAI rules. Hence, hallucinations can and must only be considered under the risk management framework of Article 9 AI Act, in high-risk AI systems. However, many typical applications of LLMs in chatbots do not fall within any of the high-risk activities listed in the AI Act (Annexes I and III). Overall, hallucinations are, therefore, one prime example where the systemic risk categorization between the DSA and the AI Act markedly diverge.

c. Our framework

Under our frameworks, hallucinations do count as systemic risk. The severity of AI hallucinations – confident generation of false information – varies with context and scale. Isolated errors, such as incorrect birthdays or minor factual mistakes, remain localized problems. Even serious individual harms, like the Norwegian man erroneously connected to child murder by ChatGPT, may not constitute systemic risks if they remain exceptional.

However, hallucinations become systemic when they poison society's information ecosystem at scale. Wachter, Mittelstadt and Russell rightly speak of “careless speech” in this context (Wachter et al., 2024), and said speech may be weaponized by malicious actors to further corrode the information ecosystem via disinformation. Widespread generation of plausible-sounding falsehoods meets the scale criterion through mass exposure across the Union. Interconnectedness emerges as false information influences decisions, shapes public opinion, and enables coordinated disinformation campaigns. The complexity of language models makes comprehensive accuracy impossible to guarantee based on the current transformer paradigm. Collective harms arise when societal trust in information sources erodes (including evidence used in courts), democratic discourse degrades, and shared factual foundations dissolve. While individual falsehoods might be corrected, the cumulative effect on information integrity proves difficult to reverse. These risks cascade through the value chain as applications built on hallucination-prone models spread misinformation across

diverse contexts. Overall, hallucinations are a clear candidate for systemic risk, but the AI Act arguably fails to capture them via its GPAI rules.

4. Cybersecurity

Cybersecurity has become a critical national and economic security imperative in an era marked by intensified geopolitical competition and the proliferation of cyber capabilities among state and non-state actors. Nation-states deploy sophisticated persistent threats against critical infrastructure, while criminal organizations execute ransomware attacks that paralyze hospitals, utilities, and supply chains. The technological landscape presents a dual challenge: emerging technologies such as AI, quantum computing, and IoT devices offer powerful tools to detect and prevent cyber intrusions, yet these same innovations introduce novel attack surfaces and vulnerabilities. AI systems can identify anomalies and respond to threats at machine speed (Roshanaei et al., 2024; Steenhoek et al., 2023), but adversaries equally weaponize AI to craft more sophisticated phishing campaigns, automate vulnerability discovery, and evade detection systems (Achuthan et al., 2024; Bengio et al., 2025).

a. DSA

The Digital Services Act recognizes the interconnected nature of modern online platforms and the pivotal role cybersecurity plays in safeguarding these systems. Although the DSA does not explicitly classify cybersecurity as a standalone category of systemic risk, it is clear that failures in cybersecurity–such as breaches, unauthorized access, or exploitation of platform vulnerabilities–can trigger or intensify several types of systemic risks addressed by the DSA. These include threats to public security, the risk of dissemination of illegal content, and negative impacts on users' fundamental rights and well-being. As such, robust cybersecurity measures are inherently required for compliance with the DSA's systemic risk obligations; and cybersecurity vulnerabilities can become systemic DSA risks themselves if they immediately facilitate fundamental rights violations or pose public security issues via the malfunctioning of critical platforms.

b. AI Act

The AI Act's Code of Practice explicitly recognizes cyber *offensive* capabilities as one of four risks that automatically qualify as systemic risks, as specified in Appendix 1.4 of the Safety and Security Chapter. This categorical classification reflects the understanding that advanced AI models possess enhanced capabilities to design attack vectors, identify system vulnerabilities, and automate exploitation techniques. More sophisticated models demonstrate superior ability to analyze complex systems, generate novel attack strategies, and adapt to defensive measures. The designation as an automatic systemic risk acknowledges that the potential for AI-enabled cyber attacks scales with model capabilities.

The impact on the Union market from AI-enhanced cyber threats proves immediately apparent through potential disruptions to digital infrastructure, financial systems, and essential services. The relationship between model advancement and cyber offensive capabilities might theoretically diminish if all models incorporated robust safeguards against malicious use. However, this theoretical mitigation must be analytically separated from the underlying risk assessment. The safeguards represent precisely the type of mitigation measures that Article 55 mandates providers to implement. The existence of potential safeguards cannot negate the classification of cyber capabilities as a systemic risk, as this would collapse the distinction between inherent risks and required mitigations. The regulatory framework correctly identifies the risk based on model capabilities absent safeguards, then separately requires providers to implement appropriate measures. This structure ensures that

providers cannot avoid systemic risk obligations by claiming that current or future safeguards (will) eliminate threats that their models' fundamental capabilities enable.

The relationship between cyber *vulnerabilities* and advanced models presents greater complexity than cyber offensive capabilities. Advanced models create expanded attack surfaces through their increased parameter counts, more complex architectures, and broader integration points with other systems. Yet these same models often incorporate intrinsic security features, advanced authentication mechanisms, and robust monitoring systems that make them intrinsically more difficult to penetrate. Larger models benefit from extensive security testing by well-resourced teams and implement defense-in-depth strategies unavailable to smaller systems. However, some of these defensive layers must again be disregarded as they constitute precisely post hoc mitigation techniques that Article 55 AI Act continuously requires. This ambiguity means that the classification of cyber vulnerabilities as a systemic risk under the AI Act currently hangs in the balance, as regulators must weigh whether increased model sophistication ultimately enhances or diminishes overall system security.

c. Our framework

Offensive cybersecurity capabilities and cybersecurity vulnerabilities each represent a classical systemic risk that the AI Act our framework readily captures. AI systems create novel attack surfaces that threaten interconnected digital infrastructure across the Union. The scale criterion is met through potential compromise of critical systems affecting millions. Interconnectedness defines cybersecurity risks – a single vulnerability can cascade through networked systems, enabling data breaches, infrastructure attacks, and service disruptions. The complexity of AI systems, with their vast parameter spaces and emergent behaviors, creates unpredictable vulnerabilities. Collective harms manifest when coordinated attacks leverage AI weaknesses to compromise financial systems, utilities, or government services. Irreversibility characterizes many cyber incidents: stolen data cannot be "unstolen," compromised systems may harbor persistent threats, and loss of public trust in digital infrastructure shapes society. The current geopolitical environment amplifies these concerns as state and non-state actors actively seek AI vulnerabilities to exploit. Cybersecurity risks inherently propagate through the value chain – every application inheriting model capabilities also inherits potential vulnerabilities.

5. Climate and Environmental Impacts

While AI has a significant, yet hitherto untapped potential for reducing emissions and energy usage around a range of crucial fields (e.g., housing and transportation) (Rolnick et al., 2022; Taddeo et al., 2021), the current tendency is for AI specifically to generate more demand for energy (IEA, 2025) and, where it is not met with renewable energies, lead to more emissions (Kaack et al., 2022). The rapid growth of digital technologies more generally, from data centers to AI and cloud services, has led to significant climate and environmental impacts, particularly through electricity and water use (Luccioni et al., 2024; OECD, 2022). Globally, information and communication technologies (ICT) now consume about 10% of total electricity (Gelenbe, 2022), with data centers alone responsible for around 1.5% of global electricity demand and 2% of greenhouse gas (GHG) emissions (IEA, 2025; Nartey, 2025). Water usage is also substantial: large data centers can consume millions of liters of water daily for cooling purposes, with estimates indicating that facilities may use between 11 million and 19 million liters per day, straining local reserves (Hsu, 2022). Research further highlights that, despite technological efficiency improvements, the overall electricity demand from digitalization continues to grow (Luccioni et al., 2025), and the expansion of these

infrastructures increases both direct and indirect environmental burdens. Regulatory strategies have not yet found effective ways to deal with the environmental fallout from digital technologies and AI more specifically (Ebert et al., 2024; Hacker, 2024).

a. DSA

With risk analyses increasingly highlighting climate change as a key systemic risk to our societies (OECD, 2003; Hui-Min et al., 2021), the question arises whether it can also be categorized as a systemic risk under the DSA. Recent legal and policy analyses indicate that environmental harm –including contributions to climate change – can fall within the scope of "systemic risks" under Article 34 of the Digital Services Act. Although the DSA does not explicitly cite environmental impacts as a standalone risk category, its requirements for very large online platforms to assess and mitigate systemic risks are phrased broadly enough to capture negative effects on public health, physical and mental well-being, and fundamental rights, all of which can be significantly affected by environmental degradation and climate change. Indeed, a habitable planet is precisely a prerequisite for the enjoyment of any of these rights and freedoms (Hacker, 2024). Moreover, an increasing number of cases reconfigures environmental interests directly as fundamental rights issues, for example through the lens of the right to live and the right to health (Gera, 2024; Hartmann & Willers, 2022; Van Zeben, 2021).

Emerging scholarship thus rightly argues that climate-related and other environmental harms, both direct (e.g., energy and water use of digital platforms) and indirect (e.g., facilitating environmentally damaging behaviors), threaten the societal interests the DSA aims to protect (Griffin, 2023; Kaesling & Wolf, 2025). Accordingly, providers are expected to take reasonable measures to minimize their environmental footprint - such as enhancing energy efficiency or reducing resource usage - as part of their Article 34 risk mitigation obligations.

b. AI Act

How about the AI Act? In our view, environmental damage and climate change contributions qualify as systemic risks under the AI Act due to the disproportionate resource consumption of advanced AI models (see also Hacker, 2024). The computational requirements for training and inference typically scale with model size (Luccioni et al., 2024; Patterson et al., 2021; Sánchez-Mompó et al., 2025; Strubell et al., 2019). Water consumption follows similar patterns (Li et al., 2023). The AI Act's systemic risk framework recognizes that these environmental impacts concentrate among the most advanced models, as only frontier systems demand hyperscale infrastructure that strains energy grids and water resources. This is also recognized in the Code of Conduct, where environmental consequences are listed as one type of risk to consider (Appendix 1.1 of the Safety and Security Chapter). While environmental harm and climate effects are also caused by less advanced models, they are disproportionately greater in the most advanced models, particularly also the so-called "reasoning" models that spend significantly more time, and hence energy, on inference compute.

The effect on the Union market manifests through multiple pathways documented by the European Environment Agency (2024). Climate change drives increased frequency of extreme weather events that disrupt physical and digital infrastructure. These cascading effects create systemic market risks: supply chain disruptions from extreme weather, potentially increased operational costs from carbon pricing under the EU ETS, and stranded assets as facilities become unviable in water-stressed regions. The European Central Bank warns that climate-related disruptions to digital services could trigger financial instability

(Alogoskoufis et al., 2021). This convergence of environmental impacts with market effects establishes clear grounds for treating climate contributions as systemic risks requiring regulatory intervention under the AI Act.

c. Our framework

Environmental harm also qualifies as systemic risk under our own framework. Direct impacts include toxic materials in computing hardware, massive energy consumption for training and inference, and water usage by data centers. Indirect effects arise when AI enables environmentally harmful activities like optimized fossil fuel extraction. The scale criterion is satisfied through contributions to climate change affecting all Member States. Interconnectedness appears through feedback loops – climate impacts affect energy availability for AI systems while AI applications influence emission patterns; moreover, a hotter planet generally needs more cooling, which in turn drives up energy demand both in households and ICT applications. Collective harms manifest as ecosystem degradation, public health impacts, and economic disruption from climate change. Environmental damage epitomizes irreversibility: atmospheric carbon persists for centuries, ecosystem collapse resists restoration, and climate tipping points create permanent changes. These impacts propagate through the value chain as every AI application contributes to cumulative environmental burden through its computational requirements, particularly also in inference (Luccioni et al., 2024).

IX. Overarching Lessons and Policy Proposals

The emergence of systemic risk as a regulatory concept across financial regulation, the Digital Services Act, and the AI Act reveals both common principles and domain-specific adaptations. This comparative analysis yields essential insights for understanding how law conceptualizes and addresses risks that transcend individual actors to threaten societal systems.

1. Risk-Based Regulatory Frameworks without Corresponding Liability

All three regulatory domains embrace a fundamental principle: regulatory obligations should scale with risk magnitude. This risk-based approach manifests consistently across frameworks, though with varying implementation strategies. Financial regulation pioneered this approach through enhanced capital requirements, stress testing, and resolution planning for systemically important financial institutions. The DSA operationalizes risk-based regulation through differentiated obligations, with VLOPs and VLOSEs bearing comprehensive risk assessment and mitigation duties that smaller platforms avoid. The AI Act similarly graduates obligations based on risk levels, with providers of GPAI models with systemic risk facing the most stringent requirements under Article 55.

This principle extends beyond these specific frameworks, of course. The GDPR's Article 35 requirement for data protection impact assessments for high-risk processing operations demonstrates how risk-based approaches permeate modern technology regulation. The consistency across domains suggests that risk-based regulation has become a fundamental principle of EU law when addressing complex technological and economic systems. As Kaminski has argued, however, this may actually have led to the neglect of other frameworks less indebted to risk-benefit analysis, such as the liability system (Kaminski, 2023). Indeed, the recent withdrawal of the AI Liability Directive seems to suggest that legislators still have not fully understood the importance of accompanying (not replacing) public-law-oriented systemic risk frameworks with robust private liability rules for an enforcement that is

independent of political whims, tariff threats, and dealmaking between the European Commission and increasingly hostile governments in the US and China. It is worth noting, however, that the experience of the GDPR is that private enforcement, except in certain high-profile cases such as those led by Noyb, is far less obviously attractive to users than public regulatory enforcement, for obvious reasons of time, money and the fear factor of the judicial system. An interim solution between a regime driven by liability and private enforcement, and one dependent on public regulator funding and willingness, may be one where rights and liability regimes are backed up by collective redress models; such ideas are being pursued energetically in the AI copyright field (Quintais, 2024).

2. The Definition of Systemic Risk: beyond the Most Advanced Models

The three frameworks adopt different approaches to defining systemic risk. Financial regulation provides precise definitions linking systemic risk to interconnectedness, substitutability, and contagion potential. The AI Act offers a formal definition in Article 3(65), though one complicated by its problematic limitation to "high-impact capabilities." The DSA eschews definition entirely, instead providing a concrete, non-exhaustive list of risk categories in Article 34 (1).

Despite lacking a formal definition, the DSA's approach proves no less operational than its counterparts. The specificity of its four risk categories - illegal content dissemination, fundamental rights impacts, civic discourse effects, and public health harms - provides clear examples for compliance, even if the fundamental rights language remains controversial.

The AI Act's definitional challenges illuminate the perils of over-specification. Its linkage of systemic risk to "the most advanced" general-purpose AI models creates conceptual confusion, conflates risk and scope of application, and ultimately undermines regulatory effectiveness. The framework would benefit from following either the financial regulation model of clear, technically grounded definitions or the DSA model of pragmatic categorization without unnecessary conceptual constraints. The restriction to the most advanced models should clearly be abandoned.

3. Comparative Risk Characteristics

While all three domains address systemic risks, the nature of these risks differs fundamentally. Financial systemic risks manifest through economic contagion, liquidity crises, and asset price collapses. Recovery, while painful, follows semi-established patterns through recapitalization, stimulus, and structural reform. Platform systemic risks operate through information distortion, social fragmentation, and behavioral manipulation. These harms often spread faster than financial contagion – misinformation reaches millions within hours – but prove harder to remedy. Eroded social trust, radicalized communities, and corrupted information ecosystems are hard to restore, as societies are learning in painful lessons across the globe.

AI systemic risks combine elements of both. Like financial risks, they can cascade through technical systems and create sudden disruptions. Like platform risks, they shape information environments and human behavior in ways that persist beyond immediate incidents. Like both financial and platform risk (Micova & Calef, 2023, p. 50), they can be triggered by external or internal sources. The speed of AI development and deployment creates additional challenges, as risk profiles evolve fast; hence, because regulatory frameworks cannot adapt themselves that quickly, the law uses risk mitigation measures paired with the vague term of

"systemic risk" to capture the fast-moving technical landscape with strategic conceptual openness. This also explains why the DSA list of systemic risks is not exhaustive.

4. Application Thresholds and Scoping Mechanisms

Each framework employs distinct mechanisms to identify which entities bear systemic risk obligations. Banking regulation uses quantitative thresholds based on assets, interconnectedness, and market share – objective financial indicators that clearly delineate systemic importance. The DSA adopts a simpler approach: platforms or search engines with 45 million or more monthly active EU users automatically qualify as VLOPs or VLOSEs. This user-based metric directly relates to platforms' capacity to influence public discourse and spread harmful content.

The AI Act's approach proves most complex and problematic. Its primary threshold – 10^25 FLOPs of training compute – attempts to capture technical sophistication but correlates imperfectly with actual systemic risk. Annex XIII provides additional factors including user base size, but these remain secondary to the computational metric.

The apparent neglect of the user base in the AI Act becomes less concerning when considered within the broader regulatory ecosystem. AI tools deployed on VLOPs or VLOSEs fall within DSA risk assessments, which do consider user reach. This indirect coverage through platform obligations partially compensates for the AI Act's technical focus, though it leaves gaps for powerful AI models deployed outside major platforms that necessitate specific designation from the Commission – which may be withheld based on political motivations.

5. Recommendations for Regulatory Evolution: Beyond the Market Paradigm

Future revisions should abandon the artificial constraint on the (Union) market. The protection of fundamental rights and interests, democratic values, and environmental sustainability cannot be reduced and subjugated to market effects. Just as the GDPR moved beyond narrow market considerations to protect privacy as a fundamental right, AI regulation must evolve to address systemic risks wherever they threaten core societal values, regardless of their market manifestation. The Union's commitment to human dignity, democracy, and environmental protection demands nothing less.

Of course, the EU can only act within its conferred competences. But indeed, the second legal basis of the AI Act, Art. 16 TFEU (data protection), provides a glimpse into that future beyond market interactions, into an Act that - in appropriate circumstances - does protect fundamental rights and non-market interests. A revised AI Act with an enhanced focus on non-discrimination, energy, and environmental impacts in systemic risk could be legislatively founded upon the specific legal bases for these policy areas already established in the TFEU, namely Article 19 for non-discrimination, Article 192 for the environment, and Article 194 for energy.

But even an Act still rooted in Art. 114 TFEU could, arguably, address non-market systemic risks, such as hallucinations, the information ecosystem and personality rights, under the prevailing *Tobacco Advertising* judgments[5] - as did the DSA (passed under Art. 114). The Tobacco Advertising test requires that EU measures adopted under Article 114 TFEU must

[5] See Germany v European Parliament and Council, Case C-376/98, *Federal Republic of Germany v European Parliament and Council of the European Union* (Tobacco Advertising I), ECLI:EU:C:2000:544; Germany v. European Parliament and Council, Case C-380/03, *Federal Republic of Germany v European Parliament and Council of the European Union* (Tobacco Advertising II), ECLI:EU:C:2006:772.

genuinely aim to improve the conditions for the establishment and functioning of the internal market by addressing real or plausible obstacles to trade or distortions of competition, but may also pursue non-market objectives if sufficiently linked to that aim.

In this light, the Tobacco Advertising test allows for a concept of systemic risk - not limited to immediate or direct market impact, but encompassing broader, structural risks that could undermine the proper functioning or stability of the internal market. If a future revision of the AI Act were to focus on systemic risks, such as those posed by AI to democratic institutions, social cohesion, or fundamental rights, these could fall plainly within the scope of Article 114 TFEU - provided there is a plausible risk that divergent national approaches could give rise to appreciable obstacles to the functioning of the internal market, or that a harmonised response is necessary to avert future fragmentation or instability. Indeed, with respect to systemic risk from AI models, this seems entirely plausible. And if not, other legal bases are available, as shown.

X. Conclusion

This paper makes several novel contributions to the systemic risk literature. The concept of systemic risk has a long pedigree in financial regulation, where its defining feature lies in the interconnectedness of actors and the cascading failures that follow. This paper has sought to update that core insight for the governance of AI and digital platforms. We have proposed a definition of systemic risk that captures the specific propagation mechanisms of these technologies and that extends beyond the narrow thresholds used in recent legislation.

Our account emphasizes that systemic risk is no longer confined to finance. Climate change and cybersecurity already receive recognition as systemic risks, and platforms and AI systems are following. The emergence of AI agents that collaborate, interact, and amplify one another's effects introduces novel pathways for systemic risk, which current regulatory frameworks do not anticipate.

Against this background, both the DSA and the AI Act make significant strides, but ultimately fall short in different ways. The DSA recognizes the broad societal implications of platform failures but leaves considerable discretion to providers. The AI Act restricts systemic risk to the "most advanced" GPAI models, thereby excluding "legacy" but widely deployed GPAI models (e.g., GPT-4; Claude 3; Llama 3) that may generate discrimination at scale or persistent hallucinations. These phenomena ought to qualify as systemic risks, given their capacity to undermine fundamental rights and democratic institutions, yet the present framework likely does not treat them as such.

We have advanced several policy proposals. Regulators should broaden the definition of systemic risk in the AI Act to encompass risks beyond compute thresholds and frontier capabilities. Risk assessments under both the AI Act and the DSA should explicitly address collective harms such as discrimination and systematic misinformation. Coordination between the AI Office and Digital Services Coordinators is necessary to capture convergent risks at the intersection of models and platforms, i.e., in hybrid systems. Finally, systemic risk obligations should be designed with flexibility to adapt to new forms of technological convergence without creating regulatory gaps, particularly without an unnecessary restriction to phenomena with EU market effects. Overall, integrating insights from finance, climate science, and cybersecurity, and foregrounding novel AI-specific pathways to systemic disruption, we hope to offer a framework for AI and platform governance that is both conceptually rigorous and practically oriented.

Bibliography

Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. Frontiers in Big Data, 7, 1497535. +

Allen, F., & Carletti, E. (2013). What is systemic risk? Journal of Money, Credit and Banking, 45(s1), 121-127.

Alogoskoufis, S., Carbone, S., Coussens, W., Fahr, S., Giuzio, M., Kuik, F., Parisi, L., Salakhova, D., & Spaggiari, M. (2021). Climate-related risks to financial stability. Financial stability review.

An, J., Huang, D., Lin, C., & Tai, M. (2024). Measuring gender and racial biases in large language models. arXiv preprint arXiv:2403.15281.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences, 122(8), e2416228122.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., & McKinnon, C. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

Bank for International Settlements (1990). Report of the Committee on Interbank Netting Schemes of the Central Banks of the Group of Ten Countries (Lamfalussy Report). Basle.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., & Goldfarb, D. (2025). International AI Safety Report (arXiv preprint arXiv:2501.17805).

Benoit, S., J.-E. Colliard, C. Hurlin, and C. Perignon. (2017). Where the risks lie: A survey on systemic risk. Review of Finance 21(1), 109–152.

Bertuzzi, L. (2025). Inside the EU's headache to designate AI models with "systemic risk". MLex

Binns, R., & Edwards, L. (2025). Reputation Management in the ChatGPT Era. In P. H. a. others (Ed.), Oxford Handbook of the Foundations and Regulation of Generative AI Oxford University Press.

Brimmer, A. F. (1989). Distinguished lecture on economics in government: central banking and systemic risks in capital markets. Journal of Economic Perspectives, 3(2), 3-16.

Carè, R., Fatima, R., & Boitan, I. A. (2024). Central banks and climate risks: Where we are and where we are going?. *International Review of Economics & Finance*, *92*, 1200-1229.

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. Journal of medical Internet research, 26(1), e53164.

De Bandt, O., & Hartmann, P. (2000). Systemic risk: A survey. ECB Working Paper Series, 35

Cline, W. R. (1984). International debt: Systemic risk and policy response. Washington, DC: Institute for International Economics.

Ebert, K., Alder, N., Herbrich, R., & Hacker, P. (2024). AI, Climate, and Regulation: From Data Centers to the AI Act. arXiv preprint arXiv:2410.06681.

European Systemic Risk Board & European Central Bank. (2023). *Towards macroprudential frameworks for managing climate risk*. Publications Office of the European Union.

European Systemic Risk Board. (2024). Advancing macroprudential tools for cyber resilience – Operational policy tools.

Financial Stability Board. (2022). *Supervisory and Regulatory Approaches to Climate-related Risks*. FSB Reports.

European Commission. (2022). DSA: Very large online platforms and search engines. https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops

European Environment Agency. (2024). European Climate Risk Assessment. EEA Report 01/2024.

Galaz, V., Centeno, M. A., Callahan, P. W., Causevic, A., Patterson, T., Brass, I., Baum, S., Farber, D., Fischer, J., & Garcia, D. (2021). Artificial intelligence, systemic risks, and sustainability. Technology in Society, 67, 101741.

Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., & Hernandez, D. (2023). The capacity for moral self-correction in large language models. arXiv preprint arXiv:2302.07459.

Gelenbe, E. (2022). The measurement and optimization of ICT energy consumption. 2022 IEEE International Symposium on Technology and Society (ISTAS), 1, 1-6.

Gera, A. (2024). Environmental Protection: A Fundamental Right or a Social Objective? Comparative Constitutional Overview. Interdisciplinary Journal of Research and Development, 11(2), 123-123.

Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Griffin, R. (2023). Climate Breakdown as a Systemic Risk in the Digital Services Act.

Griffin, R. (2025). The Politics of Risk in the Digital Services Act: A Stakeholder Mapping and Research Agenda. Weizenbaum Journal of the Digital Society, 5(2).

Gutiérrez de Rozas, L. (2022). The first ten years of the European Systemic Risk Board (2011–2021). Financial Stability Review, (42), 121–152.

Hacker, P. (2024). Sustainable AI Regulation. Common Market Law Review, 61, 345 – 386.

Hacker, P. (2024). The AI Act between Digital and Sectoral Regulations. Bertelsmann Stiftung.

Hacker, P., & Holweg, M. (2025). The Regulation of Fine-Tuning: Federated Compliance for Modified General-Purpose AI Models, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5289125

Hartmann, J., & Willers, M. (2022). Protecting rights through climate change litigation before European courts. Journal of Human Rights and the Environment, 13(1), 90-113.

Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. Internet Policy Review, 12(1).

Helberger, N., Lynskey, O., Micklitz, H.-W., Rott, P., Sax, M., & Strycharz, J. (2021). EU Consumer Protection 2.0.

Helberger, N., Sax, M., Strycharz, J., & Micklitz, H.-W. (2022). Choice Architectures in the Digital Economy: Towards a New Understanding of Digital Vulnerability. Journal of Consumer Ppolicy, 45(2), 175-200.

Hiebert, P., & Monnin, P. (2023). *Climate-related systemic risks and macroprudential policy* (INSPIRE Sustainable Central Banking Toolbox Paper 14).

Hidalgo-Oñate, D., Fuertes-Fuertes, I., & Cabedo, J. D. (2023). Climate-related prudential regulation tools in the context of sustainable and responsible investment: a systematic review. *Climate Policy*, *23*(6), 704-721.

Hsu, J. (2022). How much water do data centres use. New Scientist

Hui-Min, L., Xue-Chun, W., Xiao-Fan, Z., & Ye, Q. (2021). Understanding systemic risk induced by climate change. Advances in Climate Change Research, 12(3), 384-394.

IEA. (2025). Energy and AI.

IPCC, 2022: Summary for Policymakers [H.-O. Pörtner, D.C. Roberts, E.S. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (eds.)]. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 3–33.

Jeong, H., Ma, S., & Houmansadr, A. (2024). Bias Similarity Across Large Language Models. arXiv preprint arXiv:2410.12010.

Jin, S., Bei, Z., Chen, B., & Xia, Y. (2024). Breaking the Cycle of Recurring Failures: Applying Generative AI to Root Cause Analysis in Legacy Banking Systems. *arXiv preprint arXiv:2411.13017*.

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12(6), 518-527.

Kaesling, K., & Wolf, A. (2025). Sustainability and Risk Management under the Digital Services Act: A Touchstone for the Interpretation of 'Systemic Risks'. GRUR International, 74(2), 119-131.

Kaminski, M. E. (2023). Regulating the Risks of AI. Forthcoming, Boston University Law Review, 103.

Kasirzadeh, A (2025). Two types of AI existential risk: decisive and accumulative. *Philos Stud* 182, 1975–2003. https://doi.org/10.1007/s11098-025-02301-3

Kasirzadeh, A (2025). What are catastrophic risks from A(G)I and how should we govern them? Knight First Amendment Institute.

Kasirzadeh, A. (2024). Measurement challenges in AI catastrophic risk governance and safety frameworks. Tech Policy Press. arXiv preprint arXiv:2410.00608.

Kumar, D., Jain, U., Agarwal, S., & Harshangi, P. (2024). Investigating implicit bias in large language models: A large-scale study of over 50 LLMs. arXiv preprint arXiv:2410.12864.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less" Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. arXiv preprint arXiv:2304.03271.

Luccioni, A. S., Strubell, E., & Crawford, K. (2025). From efficiency gains to rebound effects: The problem of Jevons' paradox in AI's polarized environmental debate. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 76-88.

Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of ai deployment? Proceedings of the 2024 ACM conference on fairness, accountability, and transparency, 85-99.

Luccioni, S., Trevelin, B., & Mitchell, M. (2024). The environmental impacts of AI – policy primer. Hugging Face Blog.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination‐Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 22(2), 216-242.

Maham, P., & Küspert, S. (2023). Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks.

Micova, S. B., & Calef, A. (2023). Elements for Effective Systemic Risk Assessment under the DSA.

Monnin, P. (2021). Systemic risk buffers: The missing piece in the prudential response to climate risks. CEP Policy Brief. Council on Economic Policies.

Nartey, J. (2025). The Environmental Footprint of Data Centers: Examining Energy Consumption, Greenhouse Gas Emissions, and Strategies for Sustainable Computing. SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5181457.

Nielsen, J. (2025). AI Hallucinations on the Decline, https://www.uxtigers.com/post/ai-hallucinations.

OECD. (2022). Measuring the Environmental Impacts of AI Compute and Applications: The AI Footprint.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

Quintais, J. (2025). Generative AI, copyright and the AI Act. Computer Law & Security Review, 56: Article 106107.

Rahman, K. S. (2018). Regulating informational infrastructure: Internet platforms as the new public utilities. Georgetown Law Technology Review, 2(2), 234-251.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., & Waldman-Brown, A. (2022). Tackling climate change with machine learning. ACM Computing Surveys (CSUR), 55(2), 1-96.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science, 7(10), 201199.

Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). Enhancing cybersecurity through AI and ML: Strategies, challenges, and future directions. Journal of Information Security, 15(3), 320-339.

Sánchez-Mompó, A., Mavromatis, I., Li, P., Katsaros, K., & Khan, A. (2025). Green MLOps to Green GenOps: An Empirical Study of Energy Consumption in Discriminative and Generative AI Operations. Information, 16(4), 281.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755-759.

Somala, V., Krier, S., & Ho, A. (2025, September 12). Three challenges facing compute-based AI policies. AI Policy Perspectives & Gradient Updates. https://www.aipolicyperspectives.com/p/three-challenges-facing-compute-based.

Steenhoek, B., Rahman, M. M., Jiles, R., & Le, W. (2023). An empirical study of deep learning models for vulnerability detection. 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE),

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

Summer, M. (2003). Banking regulation and systemic risk. Open economies review, 14, 43-70.

Taddeo, M., Tsamados, A., Cowls, J., & Floridi, L. (2021). Artificial intelligence and the climate emergency: Opportunities, challenges, and recommendations. One Earth, 4(6), 776-779.

Uuk, R., Gutierrez, C.I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P. and Prunkl, C., 2024. A Taxonomy of Systemic Risks from General-Purpose AI. arXiv preprint arXiv:2412.07780.

Van Zeben, J. (2021). The Role of the EU Charter of Fundamental Rights in Climate Litigation. German Law Journal, 22(8), 1499-1510.

Wachter, S., Mittelstadt, B., & Russell, C. (2024). Do large language models have a legal duty to tell the truth? Royal Society Open Science, 11(8), 240197.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., & Huang, D. (2024). Long-form factuality in large language models. Advances in neural information processing systems, 37, 80756-80827.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., & Kasirzadeh, A. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM conference on Fairness, Accountability, and Transparency, 214-229.

Xu, C., Wang, W., Li, Y., Pang, L., Xu, J., & Chua, T.-S. (2023). A study of implicit ranking unfairness in large language models. arXiv preprint arXiv:2311.07054.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817.

Zhao, W., Goyal, T., Chiu, Y. Y., Jiang, L., Newman, B., Ravichander, A., Chandu, K., Bras, R. L., Cardie, C., & Deng, Y. (2024). Wildhallucinations: Evaluating long-form factuality in LLMs with real-world entity queries. arXiv preprint arXiv:2407.17468.