Regulating Artificial Intelligence in a World of Uncertainty

Bronwyn Howell

OCTOBER 2024

AMERICAN ENTERPRISE INSTITUTE

Executive Summary

New and increasingly capable artificial intelligence applications are a fact of life. They offer great promise of advances in human welfare but also have engendered fears of misalignment with human values and objectives, leading at best to harm to individuals and at worst to catastrophic societal outcomes and even threats to human survival. Consequently, considerable attention has been given to whether AI applications should be subject to regulation and, if so, what form that regulation should take. In the EU and the US, the focus has been on using risk management processes to ensure safe development and deployment and establishing confidence in AI use.

Risk management processes and safety regimes draw on a long history of developing computer applications based on models of mathematical, scientific, and engineering precision—and this is likely satisfactory for managing risks associated with "good, old-fashioned" symbolic AI. Nevertheless, a new generation of generative AIs (GAIs) that have been pretrained are not well suited to governance and management using risk management processes because their very basis is toward continuous adaptation and infinite variety rather than constraint and increased precision. They will also likely intersect with complex dynamic human systems, leading to great uncertainty. Managing uncertainty is different from managing risk, so a different sort of regulatory framework is needed for GAIs.

This report explores the distinctions between risk and uncertainty in AI. It illustrates why existing risk management arrangements are insufficient to prevent truly unexpected harms from GAIs. It argues that what is required is a set of arrangements for managing the consequences of harm arising, without chilling the incentives for innovative development and competitive deployment of GAIs. Arguably, insurance arrangements for managing outcome uncertainties provide a more constructive way forward than do risk management regimes, which presume knowledge of outcomes that is just not available.

Regulating Artificial Intelligence in a World of Uncertainty

Bronwyn Howell

The release of OpenAI's ChatGPT large language model (LLM) on November 30, 2022, became a turning point in public awareness of artificial intelligence applications and their potential and actual impact.[1] Previously, AI had operated largely beneath the radar of average human consciousness. However, the rapid expansion in uptake and use of ChatGPT—arguably the fastest diffusion of any new information and communications technology in human experience[2]—has brought both anticipation of gains and concerns about the potential harms of widespread AI deployment.

Some see AI as a means to achieving efficiency and productivity gains by automating mundane and repetitive tasks, freeing up humans for higher-level work.[3] They anticipate the technology will drive innovation and find solutions to long-standing problems in many industries, such as health care, transportation, and education.[4] They also expect it to lead to an explosion of creative written, audio, and video output while enabling highly personalized experiences and content tailored to end users' needs and preferences.[5]

However, some fear job losses as AI applications displace human workers.[6] They are also concerned about the extent to which AI outputs may sometimes be factually erroneous, exhibit biases, and lack coherence. And while AI systems are good at recombining existing data, they may lack the ability to generate completely new ideas and concepts.[7] Furthermore, there are fears that some may use AI systems unethically or maliciously[8] and that the companies developing very large models may act strategically to deepen organizations' reliance on their tools, compromising end user autonomy.[9] At the extreme, some fear that advanced AI systems could become misaligned with human values and objectives, potentially leading to catastrophic outcomes and even threats to human survival.[10]

Consequently, considerable attention has been given to whether AI applications should be regulated and, if so, what form that regulation should take. The European Union was an early leader on this, beginning with a draft act proposed in April 2021,[11] drawing extensively on the EU's long experience with regulations based on the precautionary principle (PP), risk management, and product standardization. Indeed, some hoped that early adoption would lead to the EU laws quickly becoming a global standard.[12]

After modifications, the act was approved by the European Parliament in March 2024 and came into force on August 1, 2024. The act will be implemented gradually, with full enforcement expected by August 2026.[13] Similar risk-based regulation was proposed for Canada in November 2022,[14] but it has not yet been passed into law.

In the United States, at the federal level, explicit legislation has been eschewed in favor of voluntary compliance with industry-led risk management standards and guidelines, notably those developed by the National Institute of Standards and Technology (NIST).[15] Notwithstanding, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, issued October 30, 2023,[16] requires government department use of AI to comply with guidance issued by the Office of Management and Budget (OMB),[17] which is strongly influenced by the NIST framework.

Individual states, however, have enacted or proposed specific constraints. For example, Colorado's SB24-205,[18] signed into law on May 17, 2024, and effective starting February 1, 2026, requires specific disclosures and NIST-style risk management practices by operators of applications deemed high-risk. California's AI legislation, passed by the State Legislature but subsequently vetoed by the governor, consisted of several bills regulating various aspects of AI. SB-896 required state agencies using generative AI (GAI) to disclose AI interactions and conduct NIST-style risk evaluations for GAI systems.[19] Other bills dealt with training data disclosure (AB-2013),[20] privacy and data protection (AB-3048 and SB-1076),[21] algorithmic discrimination (AB-2930),[22] and watermarking and provenance (AB-3211).[23] One bill also contained a so-called kill switch provision (SB-1047), requiring a mechanism enabling the full shutdown of an AI system in case it acts unexpectedly.[24]

A common theme linking all these legislative efforts is reliance on elements of risk management by AI application developers and operators to protect consumers and, ultimately, society from potential harmful effects. The presumption, rooted in the principles of negligence and product liability, is that developers owe a duty of care to ensure the safety of users of their products—particularly, to protect them from harm due to product defects.[25]

Developers are responsible for harms caused from any reasonably foreseeable conduct on their part and for foreseeable misuse of their products by others. They have a duty to warn consumers about any dangers inherent in using the product, including when it is misused, and to design the product in such a way as to prevent misuse and minimize harms from misuse.[26] This necessitates a risk management approach, in which developers must identify and analyze all reasonably foreseeable potential harms (i.e., risks), prioritize and treat them (by either avoidance, transference, reduction, or acceptance), communicate the risks to those likely to be affected, and continuously monitor and review the risk management results.[27]

However, this presumes that the likely harms are, in fact, foreseeable and amenable to management by developers alone, using traditional risk management frameworks. One of the characteristics of AI applications, and in particular generative pretrained transformers (GPTs—also called GAIs) such as LLMs, is that unpredictable outcomes are an inherent design feature. Whereas classical logic-based AI algorithms such as those used in processing big data were predicated on creating even more precise or better results (e.g., more accurate forecasts and more correct object classifications), GAIs are instead measured by their ability to create outputs with near-infinite variety.[28] An object classifier should produce the same results every time it is given the same inputs, but an LLM will have "failed" if it produces identical outputs when presented the same prompt twice.

A specific issue for risk-based regulation of GAIs is that their propensity for infinite variety and novelty rather than precise explainability means that not even the developers or the AI systems themselves know or can articulate how the applications arrive at specific outputs. This suggests that a risk-based product-safety regulation framework may not satisfactorily protect consumers (i.e., end users and society) from harm. If developers cannot anticipate ex ante a truly unexpected harmful output that a GAI subsequently creates, then a risk management strategy to address that outcome cannot be prepared. No amount of risk management activity can protect end consumers from that specific, but nonetheless real, harm.

This draws to attention the distinction between risk and uncertainty. As Frank Knight articulated in 1921,

> Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which it has never been properly separated. . . . It will appear that a measurable uncertainty or "risk" proper . . . is so far different from an unmeasurable one that it is not in effect an uncertainty at all.[29]

If the concerns engendered by AI systems, and GAIs in particular, pertain to uncertainties rather than risks, meaning classical product-safety-based risk management regulation is not going to provide the sorts of assurances sought by consumers and policymakers, then what sorts of regulation should be pursued?

This report examines this question. The first section begins with a discussion of risk and uncertainty. It examines the challenges of managing risky and uncertain environments and human responses and biases when faced with making decisions in the face of uncertainty. The second section discusses regulatory and managerial approaches to risk management and decision-making under uncertainty, notably the PP and enterprise risk management as codified under the international standard ISO 31000.

The third section evaluates the content of the EU regulations and the US NIST-based approach against classical risk management practices and the challenges posed by GAIs. This section finds that despite its espoused risk management focus, the EU act does not follow classical risk management principles, and the US approach is more consistent with ISO 31000 practices. However, neither is well suited to address unexpected outcomes from GAIs. Both incur substantial transaction costs but will be unable to address the consequences of truly unanticipated outcomes. Indeed, the presumption of strict liability on AI developers for harms caused when they have no realistic means of anticipating or managing them ex ante is neither reasonable nor just, and it will inevitably have a chilling effect on AI innovation.

The fourth section proposes a way forward using insurance-based means of sharing the costs of harm from truly unexpected AI outcomes between the developers and society. Such an approach is likely more supportive of innovation than the current arrangements and recognizes the shared responsibilities and benefits arising from AI systems as "general purpose technologies." The fifth section concludes.

Risk and Uncertainty

Knight's distinction between risk and uncertainty is foundational. Knightian uncertainty pertains to "unknown unknowns"—the things that humans both do not know and even can't possibly know. Knightian uncertainty is real and necessarily poses challenges for decision theory and regulatory practice.[30] It arises because of limits to human cognition about the current and future states of the world, especially those arising from the complex, dynamic interactions of multiple systems.

By contrast, Knightian risk reflects the situation in which the probability and magnitude of an outcome are known or can be estimated. This enables logical (i.e., rational) management of the relevant situation. Even if exact quantification is not possible, logical decision-making is feasible so long as it is possible to rank the magnitudes of and preferences for outcomes. This conceptualization of risk has enabled the development of classical risk management practices, as it facilitates the systematic specification of the boundaries within which an outcome will lie, the identities of those who will benefit or be harmed (as a class, even if not individually), and the (comparative) trade-offs arising from different possible outcomes (including the costs and benefits of adopting different risk management strategies).

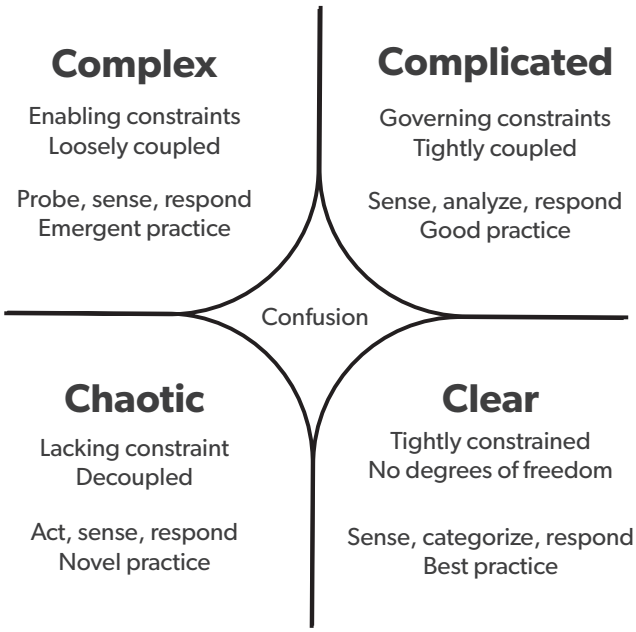
Knightian risk thus pertains to situations in which there is a significant degree of certainty—if not of the exact outcomes, at least of the statistical parameters within which they will occur. That is, it pertains to bounded systems. Classical product-safety laws presume such a bounded system. There is a single, well-defined product with identifiable consumers; the processes under which it is designed and made (notably, the bounds for safe production) are known and within the control of the designer and manufacturer (including communication with consumers regarding safe use). Under strict liability, it can be presumed that

harm caused by product defects is the responsibility of the designer and manufacturer, because of failure to design and produce the good within known safe boundaries.

These concepts derive from the scientific application of engineering principles to design and manufacture. The physical conditions are well-known, definable, and understood, with clear cause-and-effect relationships and, when ranges and tolerances can be confidently prescribed, statistical accuracy.

Such concepts also attend the application of the PP to safety regulation. While the principle supports a "better safe than sorry" approach, lack of full scientific certainty of outcomes or understanding of all cause-and-effect relationships should not be used as a reason to postpone deployment of an application with an associated threat of serious or irreversible damage (so long as the proponent rather than the public bears the burden of proof of safe boundaries, the full range of alternatives has been considered, and the process under which the application is made available is developed under an open, informed, and democratic process including all potentially affected parties).[31] However, this presumes that the state of the world is simple and straightforward and that uncertainties can be resolved or refined by further scientific inquiry.

Figure 1. The Cynefin Framework



Source: Figure adapted from Cynefin Company, "The Cynefin Framework," https://thecynefin.co/about-us/about-cynefin-framework.

What Can Be Known (and What Cannot). In reality, the world into which AI applications are being deployed and the applications themselves are far from simple. Application contexts are complex adaptive systems, constantly changing and exhibiting emergent, unpredictable behaviors. The applications, too, given their ability to update themselves and learn from use, are complex adaptive systems. Combining them increases, rather than decreases, the number of dimensions of uncertainty. Managing such environments is extremely problematic.

The Cynefin Framework (Figure 1) maps the states of the world by what is known, unknown, and unknowable. The "clear" quadrant depicts a simple world of simple scientific or engineering precision. Everything relevant is known clearly by all concerned. The systems are tightly constrained, with no degrees of freedom. The outcomes are few, and they are easily predictable and quantifiable, based on known and static cause-and-effect relationships. Decision-making takes place in the realm of reason. Classic risk management sits easily within this quadrant.

In the "complicated" quadrant, the governing constraints are tightly coupled but still knowable and predictable in advance—at least by someone. The cause-and-effect relationships are still static and may require analysis and expertise to tease out, but the problem remains bounded. An example is using the Deep Blue computer to play a chess game. The programming required is logical and explainable, though complicated. Classic risk management can be applied to this quadrant, but it is more difficult because the number of outcomes to consider is necessarily much larger.

In the "complex" quadrant, the cause-and-effect relationships are no longer static; the choice sets are

constantly changing as the systemic elements interact with each other, although the contexts may be loosely coupled. It may be possible to identify ex post how the cause-and-effect relationships of a single scenario played out, but it is impossible to predict ex ante what may occur. There are no "right answers" amenable to scientific inquiry because it is not known ex ante which of a near-infinite range of contextual characteristics will apply at each stage of interaction. Risk management is impotent to address circumstances in this quadrant; the number of possible scenarios is too vast to contemplate, and the probability is low that any given scenario selected and analyzed is the one that actually emerges. At best, GAIs are operating in this quadrant, alongside and with complex, adaptive human societal and environmental systems.

In the "chaotic" quadrant, there are no discernible constraints and no apparent identifiably reliable coupling of the relevant contexts. Cause and effect are unclear, even in an ex post analysis. It is too confusing, and waiting to learn more about the situation will not be helpful, as the underlying relationships are inherently unknowable.

Clearly, different management strategies are required for each of these states of uncertainty and complexity. When it is feasible that the cause-and-effect relationships can be known, even if they are complicated, then ex ante precautions are feasible. However, if the cause-and-effect relationships are unknown and unknowable ex ante (complex or chaotic), then any conditions imposed can address only factors already known or knowable; true uncertainty pertains to factors that are either unknowable or knowable only after the fact.

A risk-averse approach of not releasing an application with unknown outcomes into a complex environment certainly prevents unexpected harm arising from that application. But it does not prevent harms that would have arisen anyway, including those that it was known could have been averted by the application. While the application developer cannot be held liable for harms arising anyway, there is no corresponding liability for regulators who knowingly prioritized avoiding unknown harms over ameliorating known harms.

Human Responses to Uncertainty. Sometimes, decisions must be made despite inherent limitations in the context—for example, when only some relevant information is available, a short time frame precludes gathering more, or limits in human cognition constrain accuracy and decision quality. The mere presence of uncertainty cannot be a justification for not acting. Humans have been making decisions and taking actions in the face of uncertainty since the beginning of the species. However, these actions have been limited by physical elements of human information processing and decision-making.

Over time, humans have developed a range of cognitive decision-making heuristics to address these problems. However, these heuristics have led to systemic biases, which can lead to suboptimal decisions. Psychologist Daniel Kahneman and economist Amos Tversky pioneered work showing that humans value expected losses more highly than expected gains of the same magnitude (i.e., prospect theory), even though mathematically, they should be indifferent.[32]

Kahneman subsequently theorized that humans have two separate approaches to decision-making: System 1, in which decisions are made quickly, automatically, and intuitively using patterns and past experiences, and System 2, requiring slow, deliberate, and conscious (i.e., intentional) effort. The vast number of decisions humans must make and the comparatively high costs of exerting conscious effort mean that System 1 governs most daily behavior.[33] However, relying on System 1 for high-stakes decisions can lead to suboptimal outcomes. Kahneman suggests System 2 should operate as a constraint on System 1 in these instances, but the cognitive effort required means System 2 can sometimes be "lazy."

Building on this work, economists and psychologists have now cataloged a wide range of decision-making biases. In particular, humans faced with uncertainty are neurologically programmed to

- Overestimate the probabilities and costs of harm from low-probability, high-cost events;
- Overemphasize recent experiences and information;

- Fear losses more than they value gains;
- Wish to avoid irreversible events (e.g., death);
- Substitute a problem for which they do know the answer for the one for which they don't; and
- Act precipitously when waiting may have been the better thing to do.[34]

These effects can be seen in historical regulatory decision-making, even when well-reasoned System 2–type thinking should have been used for such decision-making. This can be exacerbated by the political imperative to appear to be doing something in the face of perceived or real threats.

Take, for example, the regulatory requirements in the United Kingdom and the United States following the introduction of the motor vehicle (known at the time as the "horseless carriage"). Motor vehicles were prevented from using public roads unless accompanied by a man holding a red flag walking ahead of them as they traveled. Knowledge of the known harms arising from the legacy technology (horse-drawn carriages) was used to regulate the frontier technology (motor vehicles).

The major public-safety concern with horse-drawn vehicles was that when traveling fast, the driver could lose control of the propulsion power, causing the carriage to run amok among pedestrians and other road users. The man with the flag prevented the new technology from traveling fast, thereby averting the danger posed by the unregulated legacy technology. However, this also prevented the gains from faster travel. The regulation also failed to recognize that the motor vehicle driver had far more control over propulsion than did the horse-drawn carriage driver. An internal combustion engine did not possess a mind of its own, so it was in fact less likely to escape the driver's control than was a horse.

This was eventually realized, but not before the behavior of other road users had been conditioned by the presence of the man with the flag. Lulled into a false sense of security by his presence (i.e., recent experience), other road users were unaware of the actual speeds motor vehicles could achieve, so many injuries were caused by people not getting out of motor vehicles' way fast enough when the regulation requiring the man with the flag was retired.

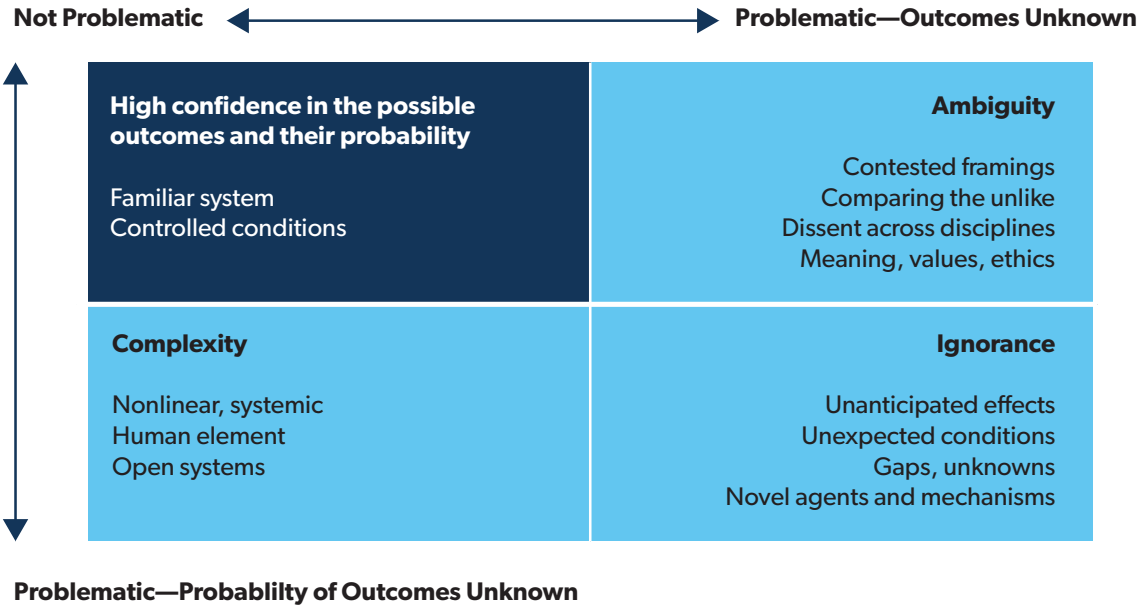
Note also that decision-making under uncertainty encompasses both short-term and long-term concerns and uncertainties. The introduction of the motor vehicle led in the short term to the rapid replacement of horse-drawn carriages, largely because motor vehicles addressed a significant (and growing) environmental health and safety problem—streets filling with horse manure as populations increased and cities grew. However, the long-term environmental effects of using petroleum products to fuel motor vehicles were not known at the time. With the benefit of time and the growth of scientific knowledge, more information has become available that casts doubt on the long-term wisdom of encouraging motor vehicle use.

That said, a truly precautionary approach of not allowing a technology to be deployed because of potential unknown future effects would ensure that no advances are ever taken, ossifying the state of technological endowment at the status quo. Two feasible strategies have achieved widespread acceptance for enabling gains from technological progress: allowing implementation so long as the best available knowledge at the time provides assurances that harms are manageable and protecting those taking the decisions at the time from the dangers of unwarranted ex post allegations of negligence. Using a regulatory process to hold someone accountable for outcomes ex post that could not reasonably have been foreseen ex ante simply hurts the person held to account and does not address the harms that have arisen.[35]

Risk Management and Safety Regulation

Knightian risk addresses the state in which both the probability and magnitude of an outcome can be known or estimated. The EU Artificial Intelligence Act defines risk as "the combination of the probability of an occurrence of harm and the severity of that harm."[36] The ISO 31000 risk management standard defines risk as "the effect of uncertainty on

Figure 2. Stirling's Sources of Uncertainty



Source: Andrew Stirling, "Risk, Precaution and Science: Towards a More Constructive Policy Debate," *EMBO Reports* 8, no. 4 (2007): 309–15, https://www.embopress.org/doi/full/10.1038/sj.embor.7400953.

objectives,"[37] whether positive or negative. While the latter acknowledges some things are not or cannot be known (i.e., uncertainty), to manage risks efficiently, it must be possible to at least identify scenarios in which the effects arising from not knowing can be quantified. This allows the manager to rank risks and have some sense of where the costs of managing risks exceeds the level of harm avoided.

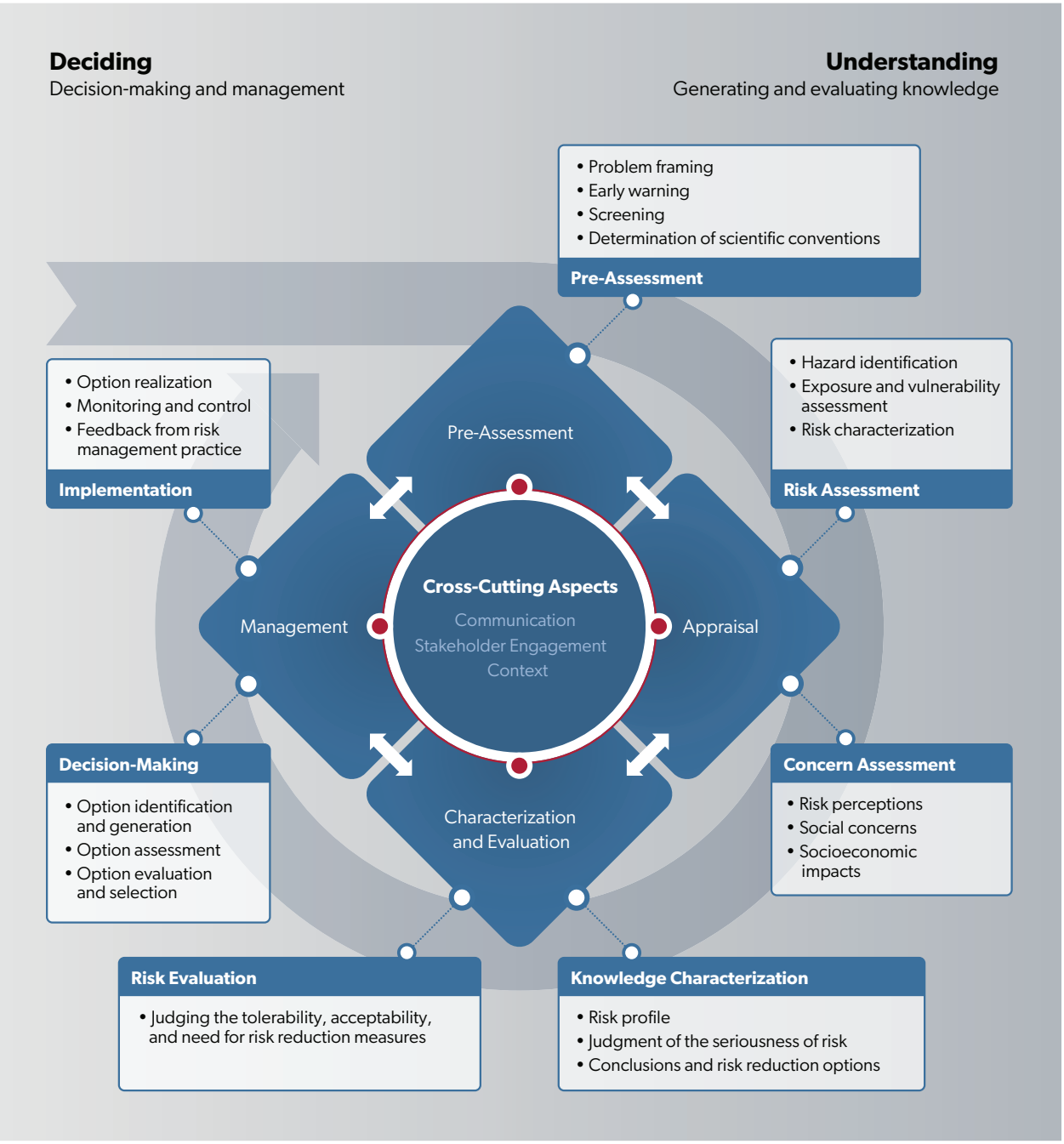
Andrew Stirling's framework (Figure 2) has been used to incorporate the PP into risk management for, especially, environmental interventions in the EU. This was done by linking the sources of uncertainty with their magnitude, especially identifying their probabilities, to determine whether the extent of uncertainty justifies legislative intervention. Didier Bourguignon identifies three levels of interpretation:

- **First/Minimal Interpretation.** Uncertainty does not justify inaction and warrants legislation despite the absence of complete scientific evidence concerning a particular hazard.
- **Second/Median Interpretation.** Uncertainty justifies action and warrants legislation even if the link between cause and effect has not been fully established.
- **Third/Maximal Interpretation.** Uncertainty necessitates legislation until the absence of hazard has been proved.[38]

Bourguignon then illustrates how these can be linked to assessed Bayesian probabilities of outcomes in environmental examples.[39]

From a Precautionary Approach to Risk Governance. Translating a precautionary approach into the practice of risk management and the application of legislated support invokes a process of risk governance, whereby society takes and implements collective decisions on activities with uncertain consequences. The International Risk Governance Council's risk governance framework (Figure 3) is one example, emphasizing that

Figure 3. The International Risk Governance Council's Framework



Source: International Risk Governance Council, *Introduction to the IRGC Risk Governance Framework, Revised Version*, 2017, 10, https://infoscience.epfl.ch/entities/publication/ff0a38ac-9e4c-4b70-a854-b9176bcf230a.

- The identification, assessment, management, evaluation, and communication of risks occurs in the context of plural values and distributed authority;
- All important actors need to be involved; and
- Good governance principles include transparency, effectiveness, efficiency, accountability, strategic focus, sustainability, equity and fairness, and respect for the rule of law.

On the one hand, such frameworks suggest a need for wide stakeholder engagement in understanding the relevant context, making decisions about deployment of a new technology, and managing the associated risks. But in practice, various product-safety laws based on the PP that have been enacted place most responsibility on the firm proposing to implement the product or technology. Legislative frameworks may set boundaries within which application may occur (using the best available scientific advice at the time), and the developer bears responsibility for demonstrating that the product fits within these boundaries and, for subsequent refinement of legislative boundaries, that the body of scientific knowledge has advanced to the point that these can be redefined.

Such processes are feasible because, for the most part, the uncertainties concerned follow the long-observed scientific and engineering pattern in which new knowledge builds directly on what has already been discovered, serving to better define the bounds for legislative protection. For example, more scientific investigation can enable the safe level of human exposure to a chemical to be more precisely defined. For a new drug, the range of conditions for which it may be prescribed can increase, but in each case, the target population can be tightly defined by having been diagnosed first with the relevant condition.

However, the defining feature, consistent with the PP, is that the primary responsibility lies with the developer for proving product safety within the limits and refining the limits. This also tends to include the specification ex ante of the processes that will be followed to ensure compliance and risk management plans to specify what will occur should the limits be approached or breached.

From Risk Governance to Enterprise Risk Management. The apparently standardized process of refining scientific and engineering knowledge to reduce uncertainty, combined with the responsibility placed on developers to demonstrate safety and reliable risk management practices, has led to the development of enterprise-level risk management standards. The most widely promulgated are those of the International Standards Organization, ISO 31000.

The core principles of ISO 31000 include the following:

- Integrate risk management into all organizational processes and activities;
- Use a structured approach to risk management for consistent and comprehensive results;
- Customize risk management to an organization's specific needs and context;
- Involve all stakeholders' knowledge, views, and perceptions in a timely and inclusive manner;
- Anticipate, detect, acknowledge, and respond to changes in the organization's risk context;
- Use the best available information;
- Acknowledge that human factors significantly influence all aspects of risk management; and
- Continually improve risk management through learning and experience.

The ISO 31000 risk management process consists of the following six steps:

1. **Communication and Consultation.** Engage with stakeholders continuously.

2. **Scope, Context, and Criteria.** Define the scope of risk management activities and understand the internal and external context.

3. **Risk Assessment.** Risk assessment consists of identification, analysis, and evaluation.
   - *Risk Identification*. Find, recognize, and describe risks.
   - *Risk Analysis.* Comprehend the nature and level of risk.
   - *Risk Evaluation*. Compare analysis results with risk criteria to determine treatment.

4. **Risk Treatment.** Select and implement options for addressing risks.

5. **Monitoring and Review.** Continually check, supervise, and critically observe to identify changes or improvements needed.

6. **Recording and Reporting.** Document and communicate risk management activities and outcomes.[40]

ISO 31000 addresses all aspects of risk governance; it is far broader than the narrow risk management activities shown in Figure 3. In this regard, an entity complying with ISO 31000 has assumed significant responsibility for societal risk governance. This would also be consistent with the assignment of responsibilities to developer firms under the PP and safety regulation.

ISO 31000 and Uncertainty. The obligation to continually monitor, review, and when necessary update the scope, context, and criteria of risk management processes allows the framework to accommodate new learning as it emerges. Nevertheless, the framework does not specifically acknowledge the implications of what is not or cannot be known about outcomes—for example, the environmental effects of burning fossil fuels in internal combustion engines, in the regulatory situation mentioned above. The framework is constrained by the limits of human knowledge at any given point in time. It assumes that in the risk assessment process, all relevant risks have been articulated. It also assumes the ability to quantify the outcomes and their probabilities and prioritize and design effective treatments.

ISO 31000's inherent limitation is that it is capable of addressing only risks and relationships that are already known. It functions in the top left quadrant of Stirling's Sources of Uncertainty (Figure 2) or the clear and complicated quadrants of the Cynefin Framework (Figure 1). If applied outside these contexts (ambiguity, complexity, and ignorance in Stirling's framework and complexity and chaos in Cynefin), then it will not provide the necessary assurances of safety and effective risk management practice for wider societal governance purposes. Indeed, to apply it in these circumstances, if it is known that these are the realms of operation, is irresponsible. The fundamentals of these contexts are such that no assurances of safety based on risk management practices can be offered.

Arguably, the application of risk management practices like ISO 31000 in these situations of genuine uncertainty could be a case of human decision-making bias: substituting a problem to which the answer is known for the complex, uncertain one that is really the subject to be addressed. However, if the extent of uncertainty is actually known (i.e., the decision maker knows they do not or cannot know relevant information), then this should be disclosed. Former Federal Trade Commissioner Maureen Ohlhausen has called for such exercise of "regulatory humility"[41] in regulatory decision-making contexts—that is, acknowledging what isn't known and how that bears on the decisions taken. This can and should be extended to risk management contexts.

For example, when deep uncertainty exists, a typical risk management strategy is to identify a series of plausible scenarios and substitute these for identified risks.[42] If probabilities are not known, a typical strategy is to assume they are equally probable. As they are based on limited human experience, they are a less-than-complete sample of the true range of outcomes, so probabilities assigned to them will be overstated. Furthermore, the selection of scenarios and probabilities assigned is almost certain to

be influenced by human biases in the face of uncertainty. High-cost, low-probability scenarios are likely to be overemphasized in the selection. Expert assessments can be relied on too much, as these experts are sometimes no better positioned to anticipate truly unknown outcomes than anyone else.[43]

A Brief History of Computer Engineering and Risk Management Evolution. The evolution of risk management into the ISO 31000 framework owes much to engineering and its role in linking scientific knowledge of the physical world with financial management practices underpinning the insurance sector.

The earliest applications of the PP and the interaction of safety factors and reliability can be traced to civil engineering in the 18th century.[44] The growing body of scientific knowledge associated with the physical sciences enabled increasing precision in engineering specifications and for manufactured goods. This supported a regulatory approach assuming an increase in scientific knowledge that served to reduce engineering uncertainty (i.e., enable more precise specifications within known boundaries). This was consistent with risk management regimes, in which new knowledge served predominantly to reduce risks or enable more precise specifications of the systems within which they operated. The even greater precision that could be applied to physical production processes enabled the expansion of risk management principles to the wider environment in which a firm operated. This underpins the wide, holistic scope of the ISO 31000 risk management principles.[45]

A narrower subset of risk management evolution can be seen in computer technology engineering. Developing and deploying computer applications that were integrated into business and societal contexts led to an awareness of the emergence of new risks, notably those associated with system security and data processing. In the US, the National Bureau of Standards (NBS, which is now the NIST) began work in the 1970s on measuring and analyzing risks to computer systems.

In response to a request from the OMB in 1978, the NBS developed a standardized risk management framework based on the work of Robert J. Courtney Jr. at IBM, which became the basis for federal projects and systems. Automated risk management tools emerged in the late 1980s, enabled by the increased amount of data available and new algorithms for identifying and mitigating risks in complicated technological environments. This allowed risk management to expand into the more integrated and comprehensive enterprise-wide and project-life-cycle scope indicated in ISO 31000.[46]

For the most part, advances in computer engineering have followed a similar path to those in civil engineering. As computers have become more capable and reliable, operational risks have reduced in importance and been superseded by a focus on wider system risks at the interface of the systems with wider society—such as data provenance (e.g., privacy concerns), quality (e.g., bias), and security (e.g., protections from leaks)—and external threats from malfeasant actors (i.e., cybersecurity).

Automating the detection and mitigation of risks using masses of stored data has been possible insofar as the algorithms can be relied on to use information about past events to predict future outcomes and systematically apply known mitigations that will respond predictably when enacted. However, this presumes a stable environment with known (or at least knowable) and predictable interactions based on known cause-and-effect relationships. Again, it is not obvious that these systems are suitable for navigating uncertainties of the sort canvased in the complex and chaotic quadrants of Figure 1.

This Time, Computing Is Different. For the majority of humans' engagement with digital computers, the focus has been on developing hardware and software following logic-based rules amenable to mathematical tractability. A key feature of these rules is that they reliably provide the same results when given the same inputs. Indeed, the scientific process itself requires recording experiments and new learning in such a manner as to allow others to reliably reproduce the findings. A computer program's accuracy and reliability is measured by whether it will provide the same outputs every time it is given the same inputs.

The early era of AI application development was underpinned by algorithms grounded in rule-based decision-making amenable to mathematical (i.e., symbolic) representation—"good, old-fashioned AI" (GOFAI).[47] The majority of big-data applications creating forecasts or allocating inputs to output classes came under this category.[48] These systems outperformed human decision makers because their outputs were reliably reproducible, while the outputs of human decision makers are subject to variations, biases, and other frailties.[49]

GOFAI systems assumed "logic" was synonymous with "intelligence." Often, these systems relied on the brute force of ever-increasing computing power to trial all possible combinations to find the optimal one. This would be equivalent to evaluating all possible combinations of risk scenarios in an ISO 31000 situation rather than having to rely on a tractable sample of four to five scenarios for human analysis. Importantly, GOFAI systems' logic could be documented and explained (even if it was complicated).

However, the new generation of AI applications—GPTs or GAIs—are not logic based. They are designed to resemble the "intuitive" decision-making of human beings (and, so far, beyond logical explanation, even by the best neuroscientists and psychologists). LLMs are but one example, based on probabilistic recombination along a vast number of dimensions of human language. Their responses may appear human and be attuned to addressing prompts and questions (within their inherent data limitations; they can only reflect the qualities of the data on which they were trained), but the paths taken to the outputs are intractable to their developers and even the AI systems themselves. As mentioned above, the hallmark of LLMs (unlike reproducible GOFAIs) is their near-infinite creativity, and an LLM producing the same output twice to the same input has failed to perform to expectations.[50]

Is Classical Risk Management Feasible in a GAI Context? The discussion so far has established that "risk management" is not suitable for governing decision-making in a state of uncertainty. For risk management to be effective, it assumes that all outcomes are known or at least definable ex ante, have reasonably assessable probabilities, and can be observed when they occur. This is necessary for the liable parties to control (or at least reliably observe and understand) the circumstances in which their technology or product is being deployed. The natures of the technology and outcomes need to be known, understood, and reliably predictable with scientific, engineering, and technological precision. The deployer must be able to define, observe, and influence (i.e., control) all use contexts and have the ability to anticipate and intervene when one of the identified risk scenarios emerges.

These conditions apply to the majority of "traditional" computer applications, operating within controlled and limited circumstances. This would apply to most enterprise operations (e.g., a bank computer system and a government department information system), including those deploying big-data AI models. However, they do not pertain to GAIs, whose infinite variety and novelty render the selection and articulation of meaningful risk management scenarios intractable.

All GAI systems are vulnerable to an outcome that was not and could not have been foreseen. As stated previously, even when an outcome has occurred, neither the developers nor the AI itself can explain how it happened. As there is no logic or reason to apply, ex post analysis of outcomes is not useful for refining and improving the risk management system. No amount of effort put into documenting and reporting use will prevent a truly unexpected outcome from occurring.

A further consideration arises in that the LLMs already in use are also general-purpose technologies. They are created with a view that other users will use and adapt them for a wide variety of purposes. Many rely on open-source software elements as inputs, and the code is placed in the open-source market for others to use. This is different from the proprietary environment in which most classical computer systems have been developed and operate, and it creates a special challenge for LLM (and other GAI) developers. Unless they control the entire value chain, both upstream and downstream, it will be impossible for them to either exert control over how the applications

are used or obtain the necessary information to conduct risk management activities.

Indeed, the open-source business model introduces significant uncertainty into the GAI environment. On the one hand, the potential for innovative new welfare-enhancing applications is increased, as the number of other developers able to make applications and adaptations is greater. However, requiring the original developers to be responsible for application risk management favors a "walled garden" business model in which the original developer maintains close and continual contractual control over application uses. The number of beneficial innovative uses will be substantially less, but the developers are shielded from the risk that unexpected outcomes will arise and that they will be held legally liable for harms occurring. Competition in enhancing and developing new GAIs will also be reduced.

Effectively, such contractual restrictions would result in a similar chilling of competition and innovation as observed historically in the telecommunications sector. Before the 1968 Carterfone case,[51] service quality obligations induced network operators to contractually prevent the attachment of devices not approved and supplied by them to the network, for fear of regulatory liability if the devices unexpectedly reduced network service quality. What AT&T deemed rational risk management practice was determined by the Federal Communications Commission to be inhibiting competition. In this case, the ability to use predictable scientific and engineering inquiry enabled demonstration of the fact that the actual risk to AT&T was small and manageable via other means.

However, the nature of GAIs means that it is not possible for scientific inquiry to provide such assurances for original GAI developers. Arguably, removing uncertainty about liability in these circumstances is necessary to foster innovation. A parallel is the effect on internet innovation of Section 230 of the Communications Decency Act of 1996, protecting "providers and users of an interactive computer service" from legal liability as publishers of content provided by another party.[52] Explicit removal of the risk of liability for actions by those outside the control of application providers facilitated the development of an open, vibrant, and innovative internet ecosystem when the only alternative was a walled garden with limited, and only proprietary, innovation.

EU and US Approaches to AI Governance

Given the discussion of the preceding two sections, I will now evaluate the content of the EU and US regulatory governance regimes concerning AI.

All countries have laws governing competition, fair trading, consumer protection, content censorship, copyright protection, and similar subjects to protect consumers, even in the event of there being no specific AI regulations or policies. These apply equally to firms providing AI applications as to any other commercial or other offerings. And the firms developing AI applications have also engaged in extensive voluntary industry self-governance (i.e., self-regulation).

At an international level, large developer firms and other stakeholders have actively engaged in civil society groups to develop and share standards, best practices, and other learnings. Examples include the AI Alliance (comprising Dell, IBM, Meta, Oracle, and many universities) and the AI Governance Alliance (including Amazon, ByteDance, Cisco, Google, Meta, Microsoft, and OpenAI). Given the high degree of technical knowledge required to develop AI systems, these entities are arguably better placed than government regulatory bodies to develop effective codes and standards and monitor and enforce compliance with them. These industry bodies have proved highly influential over history in developing and testing governance rules in emerging industries. Competition among such collectives surfaces information about which rules work best, thereby facilitating the subsequent incorporation of those proven successful into legislation (e.g., financial market regulation).[53]

Although people often express concerns about the use of industry self-regulation to protect incumbent members from competition provided by new entrants (membership of the "club" being a prerequisite for

industry participation), these risks are mitigated in new industries by the presence of multiple collectives among which new entrants can choose. Arguably, the risks to competition and innovation are greater when a single regulatory agency imposes a single, high-cost set of rules that incumbents have already met but that become a barrier to new entrants. And while some express concerns that industry self-governance may be used to exploit consumers for commercial success, the interests of end users must be mostly aligned with those of the developers; what is good for consumers is profitable for deployers. Again, allowing consumer choice among developers and governance regimes mitigates this risk.

The EU AI Act and EU AI Liability Directive. The EU AI Act was first proposed in 2022 and signed into law in February 2024. It draws heavily on a long EU history of product-safety regulation and the objective of creating a single, harmonized market across the European Union, in which products and services can be freely traded. It builds on the foundation created by the General Data Protection Regulation (GDPR)[54] and the EU Digital Markets Act (DMA),[55] which focus strongly on the control and use of data. A key presumption of the DMA is that data are a key source of market power for large technology firms; the GDPR aims to empower consumers' control over how and when their data can be used.[56]

Content. The EU AI Act defines and binds developers, deployers, importers, distributors, and operators of AI applications available for use in the European Union, regardless of where the application is developed or hosted. This is intended to ensure protection from harm for all application users within the EU. The starting assumption is that while "most AI systems pose limited to no risk and can contribute to solving many societal challenges, certain AI systems create risks that we must address to avoid undesirable outcomes."[57]

Risk is defined as "the combination of the probability of an occurrence of harm and the severity of that harm."[58] "AI system" means

> a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[59]

The EU claims its regulatory framework is risk based. It identifies four levels of risk (Figure 4). Systems deemed to pose unacceptable risk are banned. Banned systems include subliminal techniques to impair decision-making, exploiting vulnerabilities to cause users to harm others, social scoring, predicting criminal behavior, facial recognition from internet scraping, inferring emotions in the workplace or an educational setting, and real-time biometric identification in public places.[60] Systems posing high risk are subject to strict obligations and must be approved by the EU Commission's AI Office (for general-purpose AI applications—namely, GAIs) or national regulatory bodies (for other applications) before being made available in the EU.

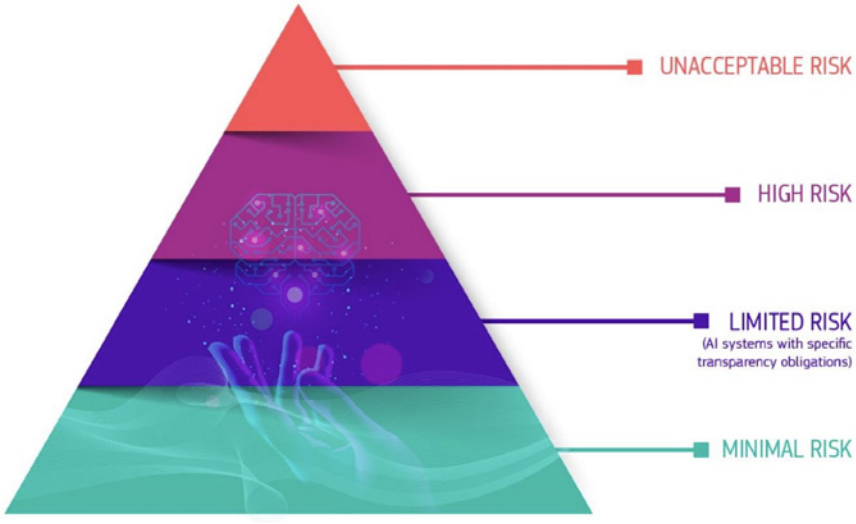
Limited risk refers to risks associated with a lack of transparency in AI usage. Providers of all AI applications are required to ensure that AI-generated content is identifiable. The transparency requirement is intended to allow users to make an informed decision about using the application or content.[61] Unrestricted use of minimal-risk applications is permitted so long as they meet the transparency obligation. It is likely the majority of AI applications will be in this category.

The definition of high-risk AI systems is complex and determined predominantly by the use case. These include the following:

- Critical infrastructures (e.g., transport), which could put citizens' lives and health at risk;
- Educational or vocational training, which may determine access to education and the professional course of someone's life (e.g., scoring of exams);

- Safety components of products (e.g., AI applications in robot-assisted surgery);
- Employment, management of workers, and access to self-employment (e.g., resume-sorting software for recruitment procedures);
- Essential private and public services (e.g., credit scoring that could deny citizens the opportunity to obtain a loan);
- Law enforcement, which may interfere with people's fundamental rights (e.g., the evaluation of the reliability of evidence);
- Migration, asylum, and border control management (e.g., the automated examination of visa applications); and
- Administration of justice and democratic processes (e.g., AI solutions to search for court rulings).[62]

Figure 4. The EU Risk-Based Approach



Source: European Commission, "AI Act," August 8, 2024, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Providers considering offering an AI system referred to in the act's Annex III that does not pose high risk (e.g., it performs only a narrow procedural task or does not replace human decision-making) must document the assessment of the system and register themselves and the system in the EU database before making it available. All other high-risk systems are subject to strict obligations before being put on the market. The requirements for high-risk systems include the following:

- The provider must have a comprehensive, continuous, and iterative risk management system that
  - ▲ Identifies and analyzes known and reasonably foreseeable risks to health, safety, and fundamental rights;
  - ▲ Estimates and evaluates the risks under the intended purpose and any reasonably foreseeable misuse; and
  - ▲ Includes measures for post-market evaluation and management.
- The system should eliminate or reduce relevant risks as much as technically feasible, and when appropriate, the provider should implement adequate mitigation and control measures when they cannot be eliminated.
- The provider must document

  - ▲ Technical knowledge, experience, education, and training and the presumable context in which the system is intended to be used; and
  - ▲ Data governance, including quality criteria (e.g., design choice, collection, preparation, formulation of assumptions, availability, biases, and shortcomings) and representativeness.
- The provider must detail technical specifications, including
  - ▲ Hardware and software specifications;
  - ▲ Documentation of the design process, development, oversight and testing, and monitoring and performance-metric records; and
  - ▲ Recordkeeping (e.g., event logs, data, and natural persons involved).
- The provider must be transparent about contact persons, characteristics (including purpose, metrics, data specifications, and instructions on interpretation of outputs), and oversight measures.
- The system must document human-oversight specifications.
- The system must include accuracy, robustness, and cybersecurity provisions.

Providers must also document and operate a quality management system, keep documentation for at least 10 years, immediately investigate problems and disable or withdraw the system as appropriate, and cooperate with EU authorities. All systems must have named representatives or authorities appointed for ensuring EU conformity; these individuals will be held liable in the event of breaches. The EU Liability Directive specifies strict liability for failure to comply with regulations in the event of harm.

Ex ante certification of compliance from one of the member state authorities must be obtained before the AI system can be made available for use. All member states are required to have processes for overseeing and issuing this certification. They must also operate regulatory sandboxes for testing; priority access to sandboxes will be given to new entrants to facilitate competitive entry.

Separate provisions apply for AI systems deemed to be GAI models with systemic risk.[63] These models are defined by compute power used during training (greater than 10^{25} floating-point operations) or an EU Commission assessment. These provisions include additional model evaluation and testing obligations, documentation and reporting requirements, and requirements to abide by yet-to-be-determined codes of practice.[64]

The act excludes open-source AI models from some obligations, provided that the applications are not monetized. Open-source general-purpose models are not required to provide technical documentation to the AI Office unless they pose a systemic risk. This exemption recognizes the collaborative nature of open-source development and is aimed at avoiding stifling innovation in this sector.[65] Systems developed for scientific research and development are also exempt from many requirements.

Analysis. While the EU claims the AI Act is risk based, the risk classification is not based on ISO 31000 principles. Under the EU Act, most AI applications won't be subject to any regulatory obligations other than disclosure. Not even basic risk management practices are required of those who develop and deploy them.

Identifying unacceptable and high-risk applications is somewhat arbitrary, based on specific use cases, sectors, and application types. There is no systematic means of identifying risk based on either the magnitude of harm or probability of occurrence, as is expected in traditional risk management. Using compute power as a proxy for defining general-purpose and high-systemic-risk applications is also grossly imperfect. There is no reason to believe that systems trained with 10^{24} flops are any safer than those trained with 10^{25}. Some large applications may

pose no real prospect of harm but will be caught up in costly compliance obligations. There is a real cost to end users, in that these applications will take longer to get to market (i.e., the opportunity cost of lost benefits), and the higher costs will translate into higher prices for use.

The requirements for high-risk system documentation resemble those for ISO 31000, but it is not clear that these will help in either identifying or preventing harm arising from truly unexpected events. The adequacy of both the documentation and assessment of systems as high-risk by the certifying authorities relies on their (human) assessment and selection of the scenarios for all of risk identification, assessment, treatment, and monitoring.

The selection of scenarios for risk management will be subject to the usual human biases, with overemphasis on managing the risks of high-cost, low-probability outcomes at the expense of lower-cost, higher-probability outcomes. This applies to both application developers and regulators. If a harmful outcome has not been evidenced in the past or has not been contemplated by these human actors in their anticipation of the future, then it will not be included. Thus, outcomes with aggregate material harm may not even form part of the active risk management processes.

These scenarios can be included in the risk management processes only after the harm has been incurred. This highlights that, no matter how much it might be promulgated that the provisions will "keep end users safe," real harms will occur because these scenarios have not been part of the initial and certified processes. The use of regulations to engender trust will also be compromised: Once information about these harms is communicated widely, confidence in both AI and the regulations will fall.

To some extent, automated risk management systems canvas a wide range of possible outcomes, including the ones that occur. But if the future occurrence is in no way dependent on any past observations, then even these systems will not be able to assign an appropriate probability to the event. The regulatory processes cannot be relied on to protect end users from these harms eventuating.

Red teaming (i.e., hiring individuals unconnected with a system's development to test it in situations inside and outside its design parameters) and testing systems in sandboxes have been proposed as means of identifying harmful outcomes that developers may have overlooked. However, these too are constrained by limits to human experience. Red teams tend to be biased in their efforts toward testing for more recently experienced harms (e.g., a current focus on testing cybersecurity stresses appears to dominate red teaming). The requirement for national regulatory bodies to prioritize limited sandbox spaces to new entrants, rather than focusing on principled risk assessment, means resources are likely to be spent on applications with eventually small or insignificant reach. Meanwhile, high-impact applications with real benefits from existing providers could be delayed because regulators operating the sandboxes are not able to sufficiently demonstrate that certification is warranted.

The exemption for open-source models is of particular interest. There is a long history of new, innovative, and highly successful applications coming from this community. In part, this may be because individuals in this community are not necessarily constrained by past uses in considering new applications. The unique qualities of GAIs (which count as EU general-purpose systems) mean they can be used in a wide range of ways that have not yet been contemplated.

Complex AI systems, the outputs of which are unpredictable, are incorporated with complex human systems that are themselves not well understood. Given these dynamics, the exemption for open-source systems means that the probability of an unexpected event occurring is much higher in this exempted space than in the regulated space. On the one hand, exemption from regulation mitigates the risk of these applications' developers being strictly liable for the harms caused, as is the case for their regulated counterparts (which will encourage more innovation, as the exemption intends). On the other hand, without a requirement to undertake at least some rudimentary risk management activities, it may prove difficult to determine whether the harms

arose from developer negligence or a truly random event. Moreover, assigning responsibility for compensation may be problematic.

By way of illustration, consider Facebook. It was developed using open-source tools in a college dorm, and its initial use was not commercial. Its algorithms, prioritizing the distribution of content, were developed and tested in this context. Real harms associated with the use of the tool for bullying and sharing fake news undoubtedly took place during this stage. Yet the developers did not focus on them because they were not yet seen as meaningful risks to be managed in the context of a college dorm. (They might have been if the developer had been the university itself, but only if these were university governance priorities at the time.)

If Facebook were commercialized today, then these risks would be addressed because of subsequent learning. Some new applications developed in similar circumstances today will almost certainly exacerbate negative effects of human behavior that will not necessarily be identified at the present because of lack of awareness or understanding among those responsible.

Sometimes these effects become evident only as emergent behavior and new and different user communities engage with the application. It is difficult in these circumstances to determine exactly who is responsible for the harms that arise. Arguably, the fact that the EU AI Act explicitly exempts open-source developers could be seen as an implicit acceptance of some degree of regulatory responsibility for these truly unexpected outcomes.

US Federal-Level AI Governance Initiatives. To date, no explicit legislation governing AI has been implemented at the federal level in the US. The White House issued its Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence in October 2023.[66] This led to the OMB issuing instructions for government departments and agencies on March 28, 2024.[67] The NIST, as part of the Department of Commerce, prepared the initial AI risk management standards in its January 2023 Risk Management Framework.[68]

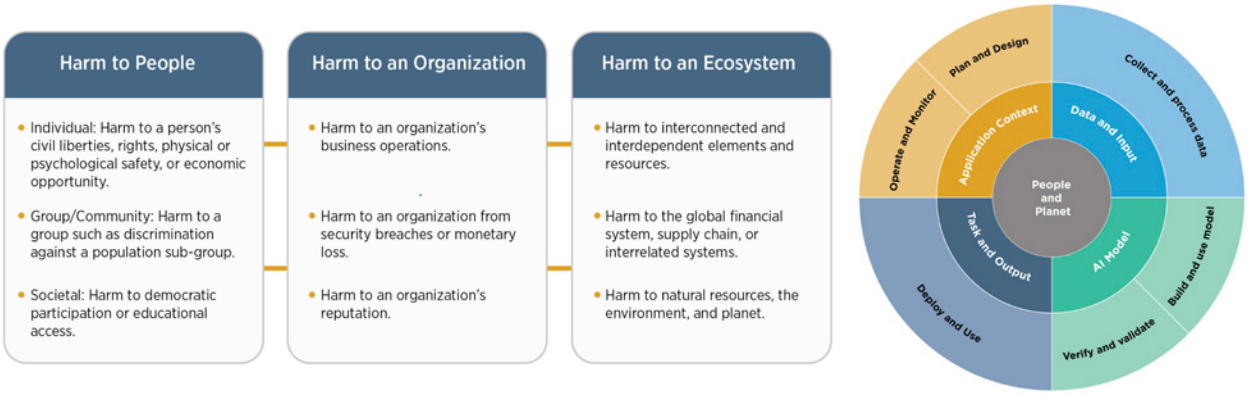
This was followed on July 26, 2024, with a set of risk management standards for GAIs,[69] partly in fulfillment of an obligation under the executive order. The National Telecommunications and Information Administration (NTIA), also part of the Department of Commerce, produced a set of accountability policy guidelines in March 2024, which address disclosure and auditing of AI applications to assure that extensive risk management processes are implemented and maintained to acceptable standards.[70]

The NIST and NTIA frameworks, though not mandatory, have become a de facto standard for US AI development and deployment. They have been formulated following a detailed and open consultation process among industry stakeholders.[71] Extensive stakeholder participation in these processes has helped ensure the frameworks developed have wide support across both developer and end user communities. They largely reflect the risk management processes adopted by large developers as part of their own internal governance responsibilities (e.g., Microsoft).[72] Voluntary compliance is widespread. The codes of practice developed by industry self-governance entities such as the AI Alliance and the AI Governance Alliance build on this base. Undoubtedly, the European standardization organizations will be informed by NIST standards when formulating and harmonizing their standards.[73]

Content. The NIST Risk Management Framework closely follows the ISO 31000 risk management standards, and indeed it derives its definition of risk management from that source. The framework is "*voluntary*, rights-preserving, non-sector-specific, and use-case-agnostic" to enable flexibility. (Emphasis in original.) It is intended to be practical and "adapt to the AI landscape as AI technologies continue to develop, and to be operationalized by organizations in varying degrees and capacities so society can benefit from AI while being protected from its potential harms."[74]

The framework recognizes that the risks posed by AI systems may differ from those encountered in traditional software and information-based systems. It

Figure 5. The NIST Risk Management Framework: A Broad Scope



Source: US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, January 2023, Figures 1 and 2, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

acknowledges that they are "trained on data that can change over time, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to explain." The AI systems' complexity and the contexts in which they are deployed are known to make it "difficult to detect and respond to failures when they occur."[75]

The framework acknowledges that the inherent socio-technical nature of AI systems is influenced by social dynamics and human behavior. It clearly articulates that

> AI risks—and benefits—can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed.[76]

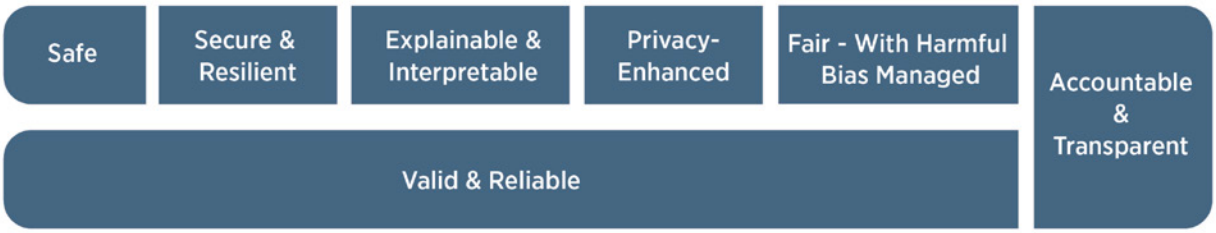
In other words, it expects unexpected emergent behavior.

Yet despite these risks making AI technology challenging to deploy and use, what is proposed is still a classical risk management system, albeit one that endeavors to emphasize human centricity, social responsibility, sustainability, and an awareness of the need to "think more critically about context and potential or unexpected negative and positive impacts."[77]

The framework recognizes the necessity of a wide view for assessing the potential harms arising from AIs. It recognizes the difficulty of measuring risks, the availability of reliable metrics, the need to track emergent risks at different stages of the AI life cycle, and the challenges posed by third-party software, hardware, and data (Figure 5). However, it does not distinguish between the management of risk and decision-making in the face of uncertainty. Rather, it focuses on the characteristics of AI systems that make them trustworthy (Figure 6).

It recognizes that trustworthiness is a social concept and will necessitate trade-offs across a wide spectrum of factors. A system will be only as trustworthy as its weakest characteristics. The definitions of the characteristics "safe" and "secure and resilient" come from ISO standards; "explainability and interpretability" recognize the distinction between describing the AI system's mechanisms and understanding the meaning of the system's outputs in its intended purpose; and "fair—with harmful bias managed" is broader than simple demographic balance, with a view to systemic, computational, statistical, and human-cognitive biases.

Figure 6. NIST AI Risks and Trustworthiness



Source: US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, January 2023, Figure 4, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

The NIST framework is operationalized around four core components, mapping directly onto ISO 31000 activities: govern, map, measure, and manage (Figure 7). These are cataloged into 19 activity categories and 72 subcategories.

A fundamental component of the NIST framework lies in accountability and transparency. This includes matters such as the provenance of training data and the easy availability of all relevant documentation for audit and assurance purposes. The March 2024 NTIA *Artificial Intelligence Accountability Policy Report* provides a comprehensive set of processes for auditing AI systems and evaluating the activities of the developer and operator firms (Figure 8). These, too, rely on the development of standards and benchmarks for both internal and external evaluation, including AI risk hierarchies; acceptable risks and trade-offs; performance of AI models, including for fairness, accuracy, robustness, reproducibility, and explainability; data quality, provenance, and governance; internal governance controls; stakeholder participation; security; internal documentation and external transparency; and testing, monitoring, and risk management.

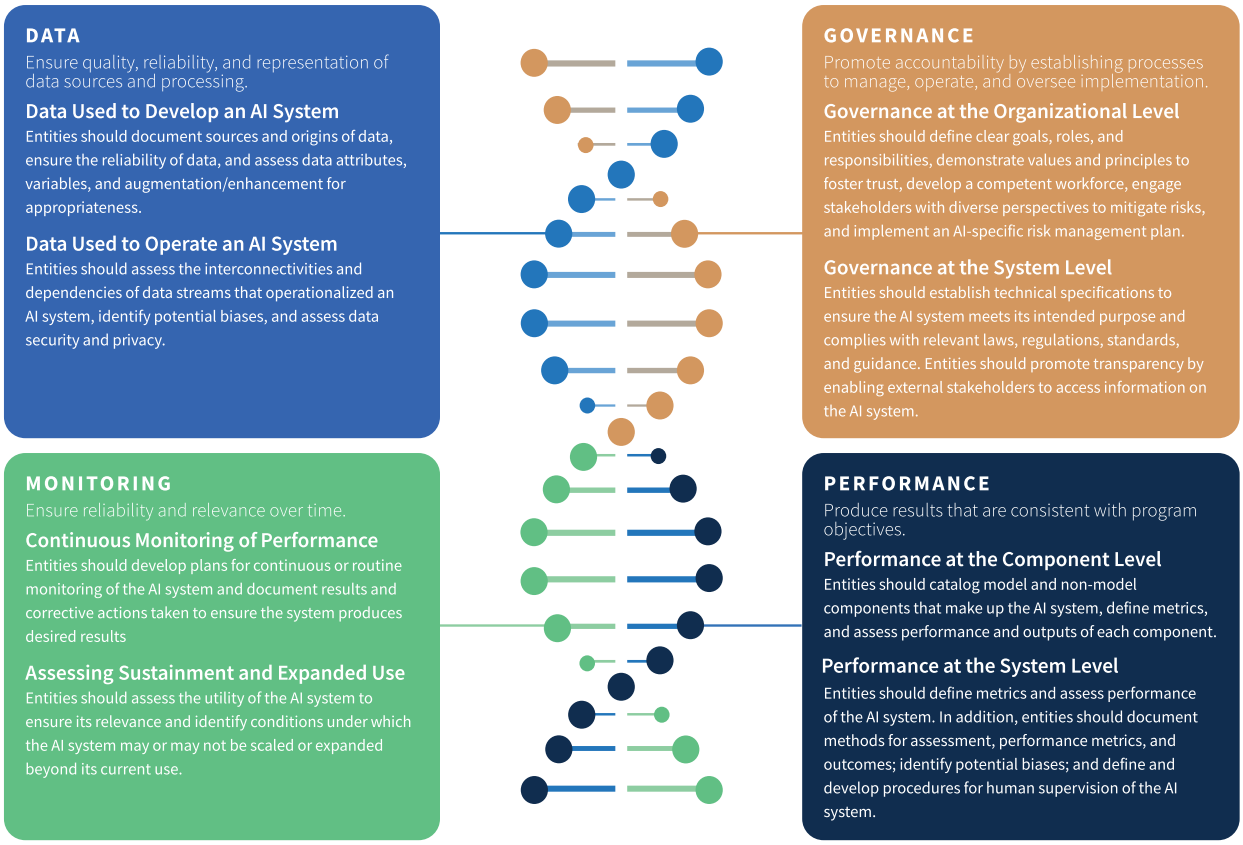
Figure 7. The NIST AI Risk Management Framework Core Activities

Source: US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0),* January 2023, Figure 5, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

The OMB's specific requirements for government departments and agencies[78] specify transparency and accountability obligations in alignment with the NIST framework. Similarly to the EU AI Act, these require each entity or department to have nominated individuals to be held accountable for AI system development and operation and ensure appropriate accountability processes are followed.

NIST, NTIA, and OMB acknowledge that standards for model performance may not yet be available and

Figure 8. NTIA Accountability Framework



Source: US Department of Commerce, National Telecommunications and Information Administration, *Artificial Intelligence Accountability Policy Report*, March 2024, 38, https://www.ntia.doc.gov/sites/default/files/publications/ntia-ai-report-final.pdf.

will continue to be developed as new systems emerge. Academic research into these is critical, and the OMB recognizes the key role that government funding can have in facilitating such research. The Stanford Institute for Human-Centered Artificial Intelligence has been foundational in developing an array of tests and benchmarks, reported annually since 2017, to assist in this measurement process.[79] These tests have become de facto reporting standards for many existing AI systems. The ability for such development to continue relies on AI firms allowing academic researchers access to their systems to develop and conduct tests; AI firms' willingness to do so stands as an added signal of their transparency and trustworthiness, compared with those taking a more proprietary approach.

NIST specifically addressed the particular challenges of GAIs in the US context, as part of the requirements of the executive order, in the July 2024 Generative Artificial Intelligence Profile. The document defines risks that are novel or exacerbated by the use of GAIs. The Executive Order 14110 definition of generative AI is used: "The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content." While not all GAI is derived from foundation models, GAI generally refers to generative foundation models. The foundation model subcategory of "dual-use foundation models" is defined by Executive Order 14110 as "an AI model that is trained

on broad data; generally uses self-supervision; contains at least tens of billions of parameters; [and] is applicable across a wide range of contexts."[80]

The NIST GAI framework recognizes that

> some GAI risks are unknown, and are therefore difficult to properly scope or evaluate given the uncertainty about potential GAI scale, complexity, and capabilities. Other risks may be known but difficult to estimate given the wide range of GAI stakeholders, uses, inputs, and outputs. Challenges with risk estimation are aggravated by a lack of visibility into GAI training data, and the generally immature state of the science of AI measurement and safety today.[81]

This document focuses on

> risks for which there is an existing empirical evidence base at the time this profile was written; for example, speculative risks that may potentially arise in more advanced, future GAI systems are not considered. Future updates may incorporate additional risks or provide further details on the risks identified below.[82]

It therefore knowingly limits the framework to the bounds of current knowledge, abstracting away from the consideration of true Knightian uncertainty.

Nonetheless, the GAI framework identifies 214 specific action areas for managing GAI risks. (An earlier draft, in April 2024, identified 467.)[83] This compares to 74 for standard AIs. Specific characteristics include confabulation; dangerous, violent, or hateful content; harmful bias and homogenization; ease of intellectual property infringement; obscene, degrading, or abusive content; and nontransparent or untraceable upstream third-party components in the value chain.

Analysis. The US approach, unlike that in the EU, is truly risk management based. While compliance is voluntary, all AI developers and users are encouraged to develop awareness of the risks and their management processes, regardless of the size or characteristics of the application concerned. Providers that voluntarily comply and are open and transparent in their development, operation, and accountability processes will offer greater assurances of their systems' trustworthiness.

This approach encourages users to familiarize themselves with the assurance tools, so that they can make knowledgeable choices regarding the AI systems they use. However, this requires significant consumer education to be effective. The NTIA has also identified the pressing shortage of people with the requisite skills even for application development, risk management, and regulation and called for government scholarships, subsidies, and extensive in-house education and training to plug the gap.[84] Just where consumer education sits within this call on resources is unclear.

While there are no explicit regulations, the leadership exhibited by NIST, NTIA, and OMB should be commended. The wide stakeholder engagement follows best practices and has ensured that the resulting frameworks build on existing industry and civil society initiatives and have widespread support. However, the lack of a single set of standards has proved challenging for some developers, who report being required to meet multiple sets of subtly different obligations with attendant additional costs. The competition among different sets of standards, though, is arguably better given the technologies' early stage of development and deployment, during which there is still much to learn about the applications and their possible uses.[85]

Clearly, US politicians have been under considerable pressure to implement legislative provisions. So far, this pressure has been averted, albeit with some intervention from the White House. However, this has favored a less intrusive and more collaborative industry-led approach. This has the advantage of being more flexible and capable of amendment and, if necessary, pivoting, should new information come to hand. The permissive approach is more likely to support an innovative environment in which new uses can be put through trials and deployed rapidly.

While some have expressed concerns that this gives too much power to developers—and to Big Tech in particular—it must be recognized that the

resources required for GAI development are significant. These firms have invested billions of dollars in their applications and have much at risk in future earnings and reputation if they are found to be operating unethically. Creating a culture of transparency by design, as has occurred with the NIST and NTIA processes, has allowed specific companies to use their willingness to be open to scrutiny from independent evaluators as an additional signal of trustworthiness and governance quality.

On the other hand, there are concerns about the apparent assurances these measures provide about the safety and trustworthiness of the applications, rather than the companies that develop and deploy them. The executive order claims in its title to assure "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The risk management framework alone cannot deliver on this objective, especially regarding GAIs. While the governance of these applications seems to take a wide view of risk types, the NIST document expressly states that the risks addressed are only those of which there is currently knowledge and awareness. These systems cannot manage the risks of unknown, unexpected, and unanticipated outcomes that will arise from their use, especially because they are going to be used in the complex and uncertain areas of human interaction and are inherently unknowable and unexplainable in their operations.

The extensive risk management activities undertaken will relate to and manage known risks only. They may succeed in averting some harms, but at high cost, as all activities must be observed and managed to avoid harms arising from a small subset of them. Additionally, all these monitored activities must be reviewed and audited to ensure the firms are compliant with the standards. And this will have no effect on reducing the probability of the harm arising from a truly unexpected event.

The EU system is less costly, as only a handful of pre-certified systems will be required to undertake the detailed risk management processes incurred by all firms abiding by the US guidelines. More harms can be expected from the insufficient management of known risks by the majority of firms in the EU not subject to the risk management provisions; these harms will be less likely to arise in the US, but overall total administration costs of the processes will be higher. There will be a lot more information available from the US operations, allowing better refinement of what is known about the risks (of which humans are already aware and about which automated systems are already keeping data) and their subsequent management. The EU firms and regulators may gain some spillover benefits from this new knowledge. But neither system can claim that its governance arrangements will keep end users safe.

The one certainty that exists is that we can expect unanticipated harms.

This suggests that, in line with Ohlhausen's prescription, a degree of humility is warranted for regulators and AI developers. It is dishonest of legislators and regulators to hold that AI regulations or even voluntary codes of practice will keep citizens safe, any more than road rules keep road users and pedestrians safe. Unexpected outcomes will occur in both contexts. The social biases monitored for in AI applications are those currently of interest. In the future, new biases will become important, but we do not know in advance what these will be, so AI systems programmed not to bias against people on demographic bases won't have learned to avoid the newly preferred biases. Humans are unpredictable, and we don't know all possible uses of a motor vehicle (i.e., new ones continue to evolve as human ingenuity expands). The same can be expected with AI—and even more so with GAI, for which we can expect complex interactions of complex systems with emergent behaviors.

At best, all that can be done is to manage the likelihood of events that we can anticipate. Developer firms have an obligation to be transparent about what they can reasonably foresee and, when possible, manage those risks responsibly. But there are limits to what they can do—both in terms of the costs of managing risks and the ability to actually prevent harms. Developers cannot prevent all harms; they should be held liable only for those that they could reasonably foresee and failed to manage adequately. There is little point in holding them accountable for outcomes they

could not reasonably foresee—as discussed above, doing so will have a chilling effect on innovation.

While AI developers stand to benefit financially from their innovations, over the fullness of time, the majority of benefits arising from new technologies accrue to end users and society—in the form of higher standards of living from advances nearly impossible to imagine at the outset. Historically, society has shared the risks with developers as new technologies are deployed, learned about, and improved. Some of those learnings lead to better regulations, but for the most part, the gains have outweighed the costs and end users have managed their risks of harm sufficiently. Thus, end users must actively be part of the risk management activity, not just passive recipients (or victims) of its consequences.

Other US State Legislation. For completeness, this section considers the state laws passed in Colorado and those passed in California except one—SB-1047—which was vetoed by the governor.[86] In principle, both of these are constrained by the same limitations discussed for the EU AI Act and US federal arrangements. Both derive considerably from the EU AI Act.

The Colorado act makes the same distinction between high-risk and other applications as the EU. Risk management provisions apply for high-risk applications, and disclosure applies for all applications. The obligations for deployers are more extensive, requiring them to notify end users when actual decisions are made using AI.

The California acts and bills are also more extensive than the EU act. As for Colorado, more extensive disclosure is required when AI is used for automated decision-making, and individuals will be allowed to opt out of solely AI-based decisions when feasible. Companies would be required to watermark all AI-generated content and make decoders available to verify AI content. Algorithmic discrimination is specifically targeted: Those who develop and deploy algorithms would be prohibited from using or making available automated decision tools that result in algorithmic discrimination. Advanced AI applications would be tightly controlled, with a special regulatory division overseeing their development and training.

The Colorado act has been passed, albeit with the governor expressing reservations upon signing, urging the legislature to "fine tune the provisions and ensure that the final product does not hamper development and expansion of new technologies in Colorado that can improve the lives of individuals."[87] Controversially, the California governor vetoed SB-1047, sending it back to the senate, because it failed to take into account "whether an Al system is deployed in high-risk environments, involves critical decision-making or the use of sensitive data." He asserted that while a California-specific set of regulations might be warranted, "it must be based on empirical evidence and science."[88] Consistent with the PP and safety regulation principles, uncertainty associated with a new technology is not sufficient to justify regulation without an assessment of both the risks and the benefits.

In both cases, some have expressed concern that the provisions make the states comparatively less attractive for AI developers and deployers compared with states without explicit regulations—and that firms might migrate accordingly. This is of especial concern for some California legislators and firms, given the state economy's reliance on innovation from Silicon Valley.[89] Some advocates saw passing a state act as a means of putting pressure on the federal government to take explicit action. Neither appeared to be cognizant of the limitations of risk management in a context of uncertainty.

A Way Forward?

The preceding three sections have demonstrated that while risk management processes and regulations may be suitable for narrowly defined applications, constrained populations, and cases in which increased precision rather than greater variety is the objective, these conditions do not apply to the development and deployment of GAIs. Rather, the degree of uncertainty associated with these applications' potential uses and the emergence of truly unexpected outcomes when complex GAIs interact with complex, poorly understood human systems suggests that, as

identified above, the one certainty that exists is that we can expect unanticipated harms.

At best, one would hope that legislators and regulators will explicitly recognize their limitations when it comes to regulating these applications to "keep people safe" and "engender trust in AI." This is not realistic for GAIs. Instead, legislators and regulators should avoid taking actions that engender a false sense of assurance or safety in end users. This would be an irresponsible use of regulation.

Rather, some regulatory humility is required. At the least, legislators and regulators should ensure that the end user population is as educated as possible about the technologies and how to responsibly interact with them. Also, they should limit regulatory intervention to specific use cases in which there is some knowledge of the interaction of complex human and AI systems (e.g., in medical research for the generation of new protein types), which may be more helpful in limiting harm than overarching, generic AI rules.

A Societal Problem. When such great uncertainty exists, there is a real risk that regulatory arrangements holding developers to account for outcomes that neither they nor anyone else could have foreseen will have a chilling effect on innovation. Society is the largest loser because when no one can predict whether an application will be benign or harmful, the risk is too great for the developer to take it to market. Extreme precaution is extremely harmful in terms of lost benefits; extreme caution by monitoring each and every step is also extremely costly, as costs are incurred in closely monitoring applications that did not need to be so closely observed, meaning the successful applications cost substantially more to take to market than was necessary.

If society is a major beneficiary of the successful applications, then a proportionate or just system would efficiently share the risks arising from uncertainty among the parties concerned. As it stands in the EU, under the strict liability imposed in the EU AI Liability Directive, the presumption is that harms are the developer's responsibility, unless it can be demonstrated that they have not been negligent. This applies to all high-, low-, and minimal-risk applications, even though the latter two face less rigorous regulatory obligations.

In the US, tort law developed from product-safety cases would likely similarly hold the developer accountable in the first instance, though the process of determining whether negligence occurred would be highly contentious. In effect, the status quo ante is that in both jurisdictions, in the event of harm, developers will be held guilty until proven innocent. Given the levels of uncertainty involved, the likelihood of a developer being found liable will be something of a lottery.

If it is known in advance that this will be the likely outcome of harm inevitably occurring, then the relevant question to address is not about apportioning liability or blame in the first instance but determining compensation for those harmed. I presume here that restitution will not likely be possible, because the harms caused will likely be irreversible. But that does not preclude the development of institutional arrangements to enable some form of redress.

Historically, in such circumstances, society has looked to insurance arrangements to manage the cost incurred by harm from truly unpredictable outcomes. Rather than holding motor vehicle manufacturers accountable for the harms to others from motorists' accidental or deliberate misuse of the vehicle (which we cannot predict precisely in advance), we require motorists to purchase public liability insurance to pay compensation for harm when it occurs. The premiums are paid by those in control of the vehicle. Motor vehicle manufacturers also purchase insurance (or are sufficiently large to self-insure) against the risks of unanticipated harm arising due to their design or manufacturing processes.

In this way, third parties are assured that if they are unlucky enough to be harmed, there will be compensation available. In some arrangements, this compensation is made available to those harmed on a no-fault basis, ensuring harm can be compensated even though it may take time for legal proceedings to clarify which fund or individual should ultimately be liable.

An Institutional Solution? Why, then, could an insurance arrangement not be contemplated for compensating truly unexpected harms arising from the deployment of AI systems, particularly GAIs?

The immediate problem is that, in the short to medium term, the information necessary to price such arrangements is not available. It is constrained by the same uncertainty as the AI systems themselves. Private markets will not develop to manage these risks (as they have for motor vehicles) until a sufficiently large number of incidents have occurred.

We see that at play even now in embryonic cybersecurity insurance markets. Despite some 30 years of experience in the use of open internet systems, the requisite information for pricing such contracts is still not satisfactory. In part, this is due to the high costs and low probabilities associated. It is also complicated by the environment constantly changing as new software platforms bring new vulnerabilities into play and human actors are constantly responding to the changes with increasingly innovative responses.

However, such uncertainties have not prevented the development of insurance funds to compensate for losses due to other high-cost and low-probability events, such as earthquakes. New Zealand, for example, has a state-owned and -managed insurance fund to address the risks of earthquake damage to land and other associated losses (over and above the damage to built infrastructure, which is expected to be covered by private insurance). Premiums paid into this fund are shared between the government and property owners. Property owners pay via a levy collected through insurance premiums; government pays via an annually budgeted sum. A process is determined ex ante as to how the fund will be disbursed in the event of an earthquake.

While administrative processes to make the payments must be established after the event, as much certainty as possible is provided ex ante about access to the funds. This has the effect of reducing uncertainty for individuals when making decisions about investing in businesses and homes in earthquake-prone locations. Without access to such insurance, it may prove too risky for investments to take place, and the New Zealand economy is the loser. The fund goes some way to manage the uncertainties about harm from low-probability, high-cost earthquakes.

How might such a fund work in an AI environment? First, the fund's costs should be shared between society and application developers, not borne by application developers alone. As the majority of benefits from successful AI systems will be nonmonetary benefits or benefits to society that are difficult to quantify, some of the risks of developing and deploying them must be borne by society. It is not just to impose the full cost onto developers.

Second, such a fund should not reduce developers' responsibilities for taking due care given the current state of knowledge when bringing a new application to market. The fund should pay out only when it is demonstrated that all reasonable steps were taken to manage known risks.

Third, given the multinational application of AI systems, the fund's administration must cross national boundaries. This suggests administration at an international rather than national level for the process of collecting premiums and making disbursements.

Importantly, such a fund would free GAI developers from the need to constrain downstream uses of their applications to manage liability for unanticipated uses, in the same manner as insurance frees motor vehicle manufacturers from responsibility for decisions taken by vehicle owners who cause harm to third parties.

Under current arrangements, GAI developers must manage for uses not anticipated when allowing third parties to use and build on their applications. As long as they are likely to be held liable in the event of harm from these downstream uses, they are best protected by either maintaining development in-house or maintaining strict contractual control over future uses. This will have distinctly chilling effects on competition and innovation. But with an insurance provision in place, so long as all due care is taken, the risks of liability are reduced. New uses will then occur (albeit with their developers also being liable to contribute to the fund), but

with the costs of downstream monitoring and competition constraints reduced.[90]

I strongly recommend that such an insurance fund be investigated, if not in respect of all AIs collectively, then at least in respect of specific AIs in specific contexts.

Transparency for Model Assessment. Given that it is impossible ex ante to know the outcomes of the intersection of GAIs and human systems, but also that expert knowledge is required to undertake assessment, then independent assessment of GAIs—for example, by expert university laboratories—must be considered. While regulatory sandboxes provide some information, as their outcomes form part of a regulatory process, they cannot be truly independent. However, allowing universities and similar third parties access to the models and the data on which they were trained will make it possible to test and assess the models and develop new tests and benchmarks for performance against safety and other objectives.

On the one hand, in a voluntary compliance context, developers with little to fear will have no problem making such access available. Indeed, their doing so is an additional signal of the quality of their intentions and their confidence in the quality of their own processes. On the other hand, if such voluntary transparency is not forthcoming, then there may be some merit in considering whether regulatory provisions to make models and data available should be implemented.

Conclusion

This report has investigated how uncertainty affects the ability to effectively regulate AI systems and reduce potential harms.

The first section explored the distinction between risk and uncertainty. Whereas classical risk management requires the ability to define and quantify both the probability and occurrence of harm, in situations of uncertainty, neither of these can be adequately defined or quantified, particularly in the case of GAIs. Management and decision-making under uncertainty differ from management and decision-making under risk. When faced with uncertainty, humans tend to make simplifying biases that can lead to unsatisfactory outcomes. This includes substituting a problem for which answers or solution methods are known for the problem to which the answer is unknown and substituting regulations used for situations for which outcomes are predictable for those for which they are not.

The first section also identified complex and chaotic contexts in which we cannot assume that cause-and-effect relationships are known or understandable. In these contexts, regulation and management strategies must consider that it is impossible to rely on thinking based on cause-and-effect relationships. It is impossible to predict in advance what outcomes will occur; at best, it may be possible to explain ex post what has occurred.

The second section explored the classic risk management systems historically used to govern product safety, assuming known and predictable scientific principles and a narrow range of identifiable subjects. These situations have been largely applicable in the development of computer systems based on predictable scientific and engineering principles. This includes mathematically tractable systems known as "good, old-fashioned AI," including most big-data models.

However, GAIs do not conform to the assumptions of classical risk management. Rather, they are characterized by the intersection of complex AI systems, which have unknown and unpredictable outcomes, with complex human systems, which have unknowable and unpredictable outcomes. Historic risk management systems are unlikely to safeguard end users and society from unexpected harms.

The third section discussed the regulatory and governance arrangements for AI development and deployment in the EU and US and evaluated their ability to prevent the outcomes identified in the second section. The EU arrangements do not conform to classical risk management principles, in that only a handful of applications in high-risk use cases are required to undertake complete risk management activities. We should expect unexpected harms, especially in the

application of open-source models, which are exempt from most risk management obligations.

The US arrangements, although not mandatory, do follow standard risk management processes. Firms following the US guidelines will provide greater assurances and harm reduction than those following the EU regulations. However, the costs of compliance will be higher. Neither set of arrangements is well suited to managing the unexpected outcomes arising from GAI deployment and use. Consequently, we should expect unexpected outcomes—and harms.

The fourth section outlines some measures that could enable the development and deployment of GAI models in as competitive and beneficial an environment as possible. First, regulators need to be honest about their limitations in regulating to prevent harm and engender confidence in AI systems. They should focus on educating end users and society about the AI environment and their role in managing personal exposure. However, there may also be some benefit in considering the extent to which GAI developers make their models and training data available to independent third parties for evaluation.

If voluntary cooperation is not sufficient, mandatory rules may be needed. However, given that we can expect unexpected harms, regulators should consider establishing an insurance fund or funds and associated governance—potentially at an international level—to enable compensation when inevitable harms arise. Such a fund, contributed to by all stakeholders—developers, users, and governments—would spread the risks across all participants and ensure compensation is paid when harm arises.

With appropriate adjudication, developers and deployers can ensure they are not held liable if they have taken all possible measures given the state of knowledge at the time of deployment to avoid known risks. Such an arrangement would contribute to a more vibrant and competitive development environment, as opposed to arrangements in which developers alone are held responsible for outcomes in which it may prove difficult to assign responsibility. It would also provide assurances to end users and society not possible under regulatory arrangements alone.

Just as when the motor vehicle was first developed, we are on the cusp of a range of new technologies that will be equally or even more transformative. We can look to the past for guidance, but ultimately, we are going on a journey into the unknown. We need to know that we will face the unexpected. We must become more comfortable about knowing that human advancement comes from facing the unexpected when it occurs and learning from it. Not taking a journey because we cannot be assured that no harm will occur is to guarantee no progress is made.

About the Author

Bronwyn Howell is a nonresident senior fellow at the American Enterprise Institute, where she focuses on the regulation, development, and deployment of new technologies and the use of technology in the health sector.

Notes

1. Kristi Hines, "History of ChatGPT: A Timeline of the Meteoric Rise of Generative AI Chatbots," Search Engine Journal, June 4, 2023, https://www.searchenginejournal.com/history-of-chatgpt-timeline/488370.

2. Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note," Reuters, February 2, 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01; and Oskar Mortensen, "How Many Users Does ChatGPT Have? Statistics & Facts (2024)," Seo.ai, April 24, 2024, https://seo.ai/blog/how-many-users-does-chatgpt-have.

3. Felix M. Simon, *Artificial Intelligence in the News: How AI Retools, Rationalizes, and Reshapes Journalism and the Public Arena*, Columbia University, Tow Center for Digital Journalism, February 6, 2024, https://towcenter.columbia.edu/sites/default/files/content/Tow%20Report_Felix-Simon-AI-in-the-News.pdf.

4. Luisa Stracqualursi and Patrizia Agati, "Twitter Users Perceptions of AI-Based E-Learning Technologies," *Scientific Reports* 14, no. 5927 (2024), https://www.nature.com/articles/s41598-024-56284-y.

5. Chen Zhi, "China and US Should Join Their Strengths in AI Revolution," *China Daily*, February 21, 2024, https://www.chinadaily.com.cn/a/202402/21/WS65d5382ba31082fc043b8335.html.

6. Stracqualursi and Agati, "Twitter Users Perceptions of AI-Based E-Learning Technologies."

7. Amanda Hetler, "Pros and Cons of AI-Generated Content," TechTarget, July 8, 2024, https://www.techtarget.com/whatis/feature/Pros-and-cons-of-AI-generated-content.

8. Stracqualursi and Agati, "Twitter Users Perceptions of AI-Based E-Learning Technologies."

9. Simon, *Artificial Intelligence in the News*.

10. Chris Vallance, "Artificial Intelligence Could Lead to Extinction, Experts Warn," BBC, May 30, 2023, https://www.bbc.com/news/uk-65746524.

11. Abhishek Khajuria, "The EU AI Act: A Landmark in AI Regulation," Observer Research Foundation, February 13, 2024, https://www.orfonline.org/expert-speak/the-eu-ai-act-a-landmark-in-ai-regulation.

12. Alex Engler, "The EU AI Act Will Have a Global Impact, but a Limited Brussels Effect," Brookings Institution, June 8, 2022, https://www.brookings.edu/articles/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect.

13. Tim Hickman et al., "Long Awaited EU AI Act Becomes Law After Publication in the EU's Official Journal," White & Case, July 16, 2024, https://www.whitecase.com/insight-alert/long-awaited-eu-ai-act-becomes-law-after-publication-eus-official-journal.

14. Government of Canada, "Bill C-27: An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to Make Consequential and Related Amendments to Other Acts," November 4, 2022, https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html.

15. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, January 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf; and US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Framework: Generative Artificial Intelligence Profile*, July 2024, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

16. White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

17. White House, Office of Management and Budget, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence," March 28, 2024, https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.

18. Consumer Protections for Artificial Intelligence Act, SB24-205, Colorado Legislature (2024), https://leg.colorado.gov/bills/sb24-205.

19. Generative Artificial Intelligence Accountability Act, SB-896, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB896; and David Shepardson and Anna Tong, "California Governor Vetoes Contentious AI Safety Bill," Reuters, September 30, 2024, https://www.reuters.com/technology/artificial-intelligence/california-governor-vetoes-contentious-ai-safety-bill-2024-09-29.

20. Generative Artificial Intelligence: Training Data Transparency, AB-2013, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013.

21. California Consumer Privacy Act of 2018: Opt-Out Preference Signal, AB-3048, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3048; and Data Brokers: Accessible Deletion Mechanism, SB-1076, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1076.

22. Automated Decision Systems, AB-2930, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2930.

23. California Digital Content Provenance Standards, AB-3211, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3211.

24. Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, SB-1047, California Legislature (2023–24), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047.

25. *Donoghue v. Stevenson* [1932] UKHL 100, [1932] AC 562 (1932), House of Lords (UK).

26. Kenneth Ross, "Product Liability Law and Its Effect on Product Safety," *InCompliance*, February 1, 2023, https://incompliancemag.com/product-liability-law-and-its-effect-on-product-safety.

27. International Standards Organization and UN Industrial Development Organization, *ISO 31000:2018—Risk Management: A Practical Guide*, 2021, https://www.iso.org/publication/PUB100464.html.

28. Nigel Shadbolt, "'From So Simple a Beginning': Species of Artificial Intelligence," *Daedalus* 151, no. 2 (2022): 28–42, https://direct.mit.edu/daed/article/151/2/28/110616/From-So-Simple-a-Beginning-Species-of-Artificial.

29. Frank H. Knight, *Risk, Uncertainty and Profit* (Boston, MA: Houghton Mifflin Company, 1921), 20.

30. Cass R. Sunstein, "Knightian Uncertainty" (working paper, Social Sciences Research Network, December 12, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4662711.

31. Giandomenico Majone, "The Precautionary Principle and Its Policy Implications," *Journal of Common Market Studies* 40, no.1 (March 2002): 89–109, https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-5965.00345. Giandomenico Majone draws on UN General Assembly, "Report of the United Nations Conference on Environment and Development," August 12, 1992, Principle 15, https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_CONF.151_26_Vol.I_Declaration.pdf. He also draws on the Wingspread Conference on the Precautionary Principle, which took place on January 26, 1998, and is reported in Science & Environmental Health Network, "Wingspread Conference on the Precautionary Principle," August 5, 2013, https://www.sehn.org/sehn/wingspread-conference-on-the-precautionary-principle.

32. Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica* 47, no. 2 (March 1979): 263–92, https://www.jstor.org/stable/1914185.

33. Daniel Kahneman, *Thinking, Fast and Slow* (London: Penguin Books, 2011).

34. See, for example, Nassim Nicholas Taleb, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets* (London: Penguin Books, 2007); Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (New York: Penguin Books, 2009); John Kay and Mervyn King, *Radical Uncertainty: Decision-Making Beyond the Numbers* (London: W. W. Norton & Company, 2020); and Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (New York: Little, Brown Spark, 2021).

35. Martha C. Nussbaum, *Anger and Forgiveness: Resentment, Generosity, Justice* (New York: Oxford University Press, 2016).

36. EU Artificial Intelligence Act, article 3 (2024).

37. International Standards Organization and UN Industrial Development Organization, *ISO 31000:2018—Risk Management*, 3.1.

38. Didier Bourguignon, *The Precautionary Principle: Definition, Applications and Governance*, European Parliamentary Research Service, December 2015, 7, https://www.europarl.europa.eu/thinktank/en/document/EPRS_IDA(2015)573876.

39. Bourguignon, *The Precautionary Principle*, Table 1.

40. Alexander S. Gillis, "ISO 31000 Risk Management," TechTarget, October 2023, https://www.techtarget.com/searchsecurity/definition/ISO-31000-Risk-Management; and André Hammer, "Iso 31000," Readynez, April 5, 2024, https://www.readynez.com/en/blog/unveiling-the-8-principles-of-iso-31000.

41. Maureen K. Ohlhausen, "The Procrustean Problem with Prescriptive Regulation," *ComLaw Conspectus: Journal of Communications Law and Technology Policy* 23, no.1 (2014), https://scholarship.law.edu/cgi/viewcontent.cgi?article=1548&context=commlaw.

42. Terje Aven, "On How to Deal with Deep Uncertainties in a Risk Assessment and Management Context," *Risk Analysis* 33, no. 12 (2013): 2082–91, https://pubmed.ncbi.nlm.nih.gov/23656628; and Julie Shortridge, Terje Aven, and Seth Guikema, "Risk Assessment Under Deep Uncertainty: A Methodological Comparison," *Reliability Engineering & System Safety* 159 (2017): 12–23, https://www.sciencedirect.com/science/article/abs/pii/S095183201630713X.

43. Philip E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton, NJ: Princeton University Press, 2005); and Dan Pilat and Sekoul Krastev, "Gaining Insight into Foresight," Decision Lab, https://thedecisionlab.com/thinkers/political-science/philip-tetlock.

44. Bernard Forest de Bélidor, *La Science des Ingénieurs, Dans la Conduite des Travaux de Fortification et d'Architecture Civile* [The Science of Engineers, in the Conduct of Fortification and Civil Architecture Works] (Paris, France: Chez Claude Jombert, 1729); and Isaac Elishakoff, *Safety Factors and Reliability: Friends or Foes?* (Dordrecht, Netherlands: Kluwer Academic Publishers, 2004).

45. Charles Wert, "The Evolution of Risk Management: From Ancient Civilizations to a Holistic Approach," April 26, 2003, https://www.linkedin.com/pulse/evolution-risk-management-from-ancient-civilizations-holistic-wert.

46. US Department of Commerce, National Institute of Standards and Technology, "Enhancing Risk Management," September 26, 2022, https://csrc.nist.gov/nist-cyber-history/risk-management/chapter.

47. Shadbolt, "From So Simple a Beginning."

48. Nicky Case and Hack Club, "AI Safety for Fleshy Humans, Part I," May 2024, https://aisafety.dance/p1.

49. Kahneman, Sibony, and Sunstein, *Noise*.

50. Bronwyn Howell, "The Precautionary Principle, Safety Regulation, and AI: This Time, It *Really Is* Different," American Enterprise Institute, September 4, 2024, https://www.aei.org/research-products/report/the-precautionary-principle-safety-regulation-and-ai-this-time-it-really-is-different.

51. In the Matter of *Use of the Carterphone Device in Message Toll Telephone Service; in the Matter of Thomas F. Carter and Carter Electronics Corp., Dallas, Tex* (*Complainants*), *v. American Telephone and Telegraph Co., and Associated Bell System Companies, Southwestern Bell Telephone Co., and General Telephone Co. of the Southwest* (*Defendants*), Federal Communications Commission, 13 F.C.C.2d 420 (1968); 13 Rad. Reg. 2d (P & F) 597 Release-Number: FCC 68-661, June 26, 1968, https://web.archive.org/web/20150120021035/http://www.uiowa.edu/~cyberlaw/FCCOps/1968/13F2-420.html; and David Brodwin, "Carterfone Case Showed How Regulations Promote Competition," *US News & World Report*, June 28, 2012, https://www.usnews.com/opinion/blogs/economic-intelligence/2012/06/28/carterfone-case-showed-how-regulations-promote-competition.

52. Protection for Private Blocking and Screening of Offensive Material, 47 U.S.C. § 230(c)(1) (2012).

53. Bronwyn Howell and Petrus H. Potgieter, "Industry Self-Regulation of Cryptocurrency Exchanges," in *Workshop on the Ostrom Workshop 6* (Bloomington, IN: Digital Library of the Commons, 2019), https://dlc.dlib.indiana.edu/dlc/items/cdb713b2-a5c6-4739-a5b7-17023a755e3d.

54. EU Parliament and EU Council, Regulation (EU) 2016/679, April 27, 2016, in *Official Journal of the European Union*, L 199/1 (May 4, 2016), https://gdpr-info.eu.

55. EU Parliament and EU Council, Regulation (EU) 2022/1925, September 14, 2022, in *Official Journal of the European Union*, L 265/1 (October 12, 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2022.265.01.0001.01.ENG&toc=OJ%3AL%3A2022%3A265%3ATOC.

56. Alessia S. D'Amico, "The DMA's Consent Moment and Its Relationship with the GDPR," *European Journal of Risk Regulation* (2024), https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/dmas-consent-moment-and-its-relationship-with-the-gdpr/EC81E79B3839BDDD441CF6EC2BFB6BA6.
57. European Commission, "AI Act," August 8, 2024, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
58. EU Artificial Intelligence Act, article 3(2) (2024).
59. EU Artificial Intelligence Act, article 3(1) (2024).
60. EU Artificial Intelligence Act, article 5 (2024).
61. EU Artificial Intelligence Act, article 50 (2024).
62. EU Artificial Intelligence Act, article 6 and annex III (2024).
63. EU Artificial Intelligence Act, chap. V (2024).
64. EU Artificial Intelligence Act, article 55 (2024).
65. EU Artificial Intelligence Act, recital 102 and articles 25(4), 52(2), and 54(6) (2024).
66. European Commission, "AI Act."
67. European Commission, "AI Act."
68. European Commission, "AI Act."
69. European Commission, "AI Act."
70. US Department of Commerce, National Telecommunications and Information Administration, "Artificial Intelligence Accountability Policy Report," March 2024, https://www.ntia.doc.gov/sites/default/files/publications/ntia-ai-report-final.pdf.
71. US Department of Commerce, National Institute of Standards and Technology, "AI RMF Development," January 2, 2024, https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development.
72. Microsoft, *Governing AI: A Blueprint for the Future*, May 25, 2023, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw.
73. Claire O'Brien et al., "The Role of Harmonised Standards as Tools for AI Act Compliance," DLA Piper, January 11, 2024, 12, https://www.dlapiper.com/es-pr/insights/publications/2024/01/the-role-of-harmonised-standards-as-tools-for-ai-act-compliance.
74. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework* (*AI RMF 1.0*), 2.
75. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework* (*AI RMF 1.0*), 1.
76. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework* (*AI RMF 1.0*).
77. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework* (*AI RMF 1.0*).
78. White House, Office of Management and Budget, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence."
79. Stanford University, Institute for Human-Centered Artificial Intelligence, "About," https://hai.stanford.edu/about; and Stanford University, Institute for Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report 2024*, April 2024, https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.
80. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Framework*.
81. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Framework*, 3.
82. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Framework*.
83. US Department of Commerce, National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (*AI 600-1*) *Initial Public Draft*, April 2024, https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.
84. Bronwyn Howell, "Who Will Monitor the AI Monitors? And What Should They Watch?," AEIdeas, April 11, 2024, https://www.aei.org/technology-and-innovation/who-will-monitor-the-ai-monitors-and-what-should-they-watch.

85. Bronwyn Howell, "Competition in AI Regulation: Essential for an Emerging Industry," AEIdeas, July 1, 2024, https://www.aei.org/technology-and-innovation/competition-in-ai-regulation-essential-for-an-emerging-industry.

86. California Office of the Governor, letter to the California State Senate, September 29, 2024, https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf.

87. Dominique Shelton Leipzig et al., "Colorado Governor Signs Comprehensive AI Bill," Mayer Brown, June 4, 2024. https://www.mayerbrown.com/en/insights/publications/2024/06/colorado-governor-signs-comprehensive-ai-bill.

88. California Office of the Governor, letter to the California State Senate.

89. Wendy Lee, "AI Safety Bill Passes California Legislature," *Los Angeles Times*, August 29, 2024, https://www.latimes.com/entertainment-arts/business/story/2024-08-29/newsom-scott-wiener-sb1047-ai-bill.

90. Bronwyn Howell, "Regulating AI, Hypothetically and in Reality," AEIdeas, August 28, 2024, https://www.aei.org/technology-and-innovation/regulating-ai-hypothetically-and-in-reality.