Preliminary Version

Sarah Bérubé, Catherine Régis and Lee Tiedrich

Managing AI Risks Under Uncertainty: a Toolkit for Policymakers

With the financial support of

Mila

Acknowledgement

The authors would like to thank Luis Aranda, Senior Economist at the OECD; Hugo Larochelle, Scientific Director at Mila; Steven Payson, Economics Program Coordinator and Senior Lecturer at Johns Hopkins University; and Elissa Strome, Executive Director of the Pan-Canadian AI Strategy at CIFAR for their review and insightful feedback. The Toolkit would not have been possible without the support of Laëtitia Vu and Sally Guy.

Who Are the Authors?

Sarah Bérubé a senior policy analyst with a strong background in tech policy, public governance, and international relations. She was part of the first cohort of the Mila AI Policy Fellowship and worked on the OECD AI team. Prior to this, she spent seven years as a policy analyst with the Government of Canada, contributing to initiatives on online harms and open government. She began her public service career through the Advanced Policy Analysts Program, which exposes each year a cohort of 12 young professionals to the policy development process in the federal government.

Catherine Régis is Professor of Law at Université de Montréal and holder of a Canada CIFAR AI Chair and a Chair on Science Diplomacy and the Global Governance of AI (Fonds de recherche du Québec). She is an Associate Academic Member at Mila, Co-director of the Canadian AI Safety Institute research program (CIFAR), Director of Social Innovation and International Policy at IVADO as well as an Intellectual Forum Senior Research Associate at Jesus College (University of Cambridge). Her work focuses on responsible AI governance and regulation at national and international levels, with an emphasis on safety issues, human rights, equitable benefit sharing, and applications in health care and justice.

Lee Tiedrich has faculty positions at Duke University and the University of Maryland. She served as a Senior AI Advisor to the US National Institute of Standards and Technology, as an OECD/GPAI consultant and expert, as a Senior Adviser to the 2025 and 2026 International AI Safety Reports, as member of the EU AI Office working group on codes of conduct, and a member of the AI Center for Government Advisory Council. She was a partner at the global law firm Covington & Burling LLP. Her work focuses on U.S. and global AI safety, security, risk management governance, regulation and policy; AI fluency, readiness and adoption; intellectual property; corporate and board oversight; and data matters.

Why a Preliminary Toolkit?

This Toolkit is being released at a time of growing debate about the role of policymakers in responding to emerging AI risks. Over the next six months, it will be refined through consultations with experts and workshops that apply the Toolkit to concrete use cases to assess its usefulness and applicability for policymakers. A final version, incorporating insights from this process, will be released in March 2027.

Disclaimer

The views expressed in this Toolkit are solely those of the authors and do not necessarily reflect those of their respective organizations.

Table of Content

Executive Summary 6

Introduction 8

Understand: Evidence Gaps Around AI Risks 10

Learn: Managing Risks Under Uncertainty 17

- **Approaches For Risk Management Under Uncertainty 17**
- **Lessons From Other Safety Domains And Policy Fields 25**

Breakdown: AI Policymaking Under Uncertainty 30

- **Model Layer: Technical Safety, Reliability And Alignment 32**
- **Individual Layer: Misuse Prevention 34**
- **Societal Layer: Systemic Resilience 36**
- **International Layer: Coordination 38**

Act: AI Policymaking in Practice 40

- **Policy Issue: Identify, Analyze And Prioritize Risks 40**
- **Policy Options: Define Objective, Select And Compare Measures 48**
- **Examples Of Policy Approaches 54**

Conclusion 58

References 60

Appendix 1 – Risks Cards 64

Appendix 2 - Risk And Uncertainty Management Approaches 73

Appendix 3 - Examples Of Policy Intervention 81

Appendix 4 - Steps for Developing Policy Responses to AI Risks 91

Executive Summary

General purpose AI systems hold significant economic potential, between US$2.6 and US$4.4 trillion to the global economy annually according to McKinsey Global Institute. They have already found impactful applications across healthcare, scientific research, education, and other sectors. Yet, they also introduce risks of misuse, system failures, and broader societal impacts on labour markets, social structures, and human autonomy.

Some of these risks have already been translated into real-world harms, signalling negative externalities associated with general-purpose AI systems policymakers cannot ignore and calling for risk identification and prioritization according to harm's likelihood and severity. Managing these risks remains essential to providing predictability to businesses, building public trust, supporting adoption, and harnessing the social and economic opportunities AI has to offer.

However, policymakers face substantial uncertainty regarding the probability, timing, and severity of various AI-related harms, complicating the analysis about appropriate policy responses. While risk is the combination of an event's probability and the magnitude of its consequences, uncertainty refers to the inability to reliably estimate these two dimensions due to the absence of historical data, the pace of innovation, or persistent gaps in knowledge.

Uncertainty, nonetheless, does not eliminate the need for policy decisions. AI systems are already being deployed across society, meaning that maintaining the status quo is itself a policy choice with consequences. The practical policy question is therefore not only how to ensure that this new technology can be used without causing unreasonable injury or harm to people, property and the planet, but also whether society is adequately prepared to handle these AI risks and the uncertainty that surrounds them.

This change of perspective requires policymakers to move beyond traditional risk management to risk management under uncertainty. Under this approach, the objective of policy intervention is to strengthen society's capacity to prevent and mitigate AI-related harms while preserving the benefits of technological innovation. This shifts the focus from solely predicting probabilities and consequences in order to optimize a policy response to adopting targeted policies to strengthening society's resilience by reducing vulnerabilities, limiting the propagation of harms and preserving the capacity to adapt as knowledge evolves.

This toolkit provides a thinking framework to guide policymakers in the development of robust policy frameworks capable of managing risk when probabilities and consequences are unknown, contested, or difficult to estimate. It is structured around four sections. First, it summarizes emerging AI risks, their potential harms, and the evidence gaps that constrain policymaking. It then examines principles from risk management and decision-making under uncertainty as well as lessons from other safety domains and policy fields. Building on these insights, it presents a multi-layered policy framework to help policymakers identify where AI risks originate and where policy interventions are most likely to be effective. Because these layers are interconnected, resilience depends not on any single policy measure but on the combined effectiveness of interventions across all layers, including measures to reduce the evidence gap. Finally, the toolkit provides practical guidance to help policymakers design interventions to strengthen society's resilience to AI risks of harm. This includes presenting the policy instruments at their disposal as well as examples of policy approaches.

Introduction

Since the 2022 launch of ChatGPT, we have witnessed not only rapid advances in the capabilities and adoption of general-purpose AI systems[1], but also growing evidence of real-world harms associated with their use. While general purpose AI systems holds a lot of potential for the social good and the economy and have found useful applications in healthcare, scientific research, education, and other sectors, criminals have also used AI-generated content to deceive and defraud individuals; people have been misled by AI hallucinations, sometimes with harmful consequences to their well-being or reputation; and some users have developed unhealthy dependencies with AI companions. These harms stem from several AI-related risks, such as the misuse of AI systems, their potential malfunctions, and their systemic impact on labour markets, social structures, and human autonomy. As AI capabilities continue to advance, additional risks are likely to emerge.

This evolving AI risk landscape presents a particular challenge for policymakers: the evidence available is often insufficient to determine the probability of an event and the magnitude of its consequences. The limited historical data, the rapid technological change, the incomplete evidence, and the persistent knowledge gaps make it difficult for policymakers to estimate these two dimensions with confidence. The resulting uncertainty complicates both the assessment of potential harms and the crafting of policy responses.

However, uncertainty does not preclude policy action. General-purpose AI systems are already being integrated into products, services, workplaces,

[1] According to the International AI Safety Reports, general-purpose AI systems are "AI models and systems capable of performing a wide variety of tasks across different contexts. These models and systems perform tasks like generating text, images, audio, or other forms of data, and are frequently adapted to a range of domain-specific applications".

and public life. According to the OECD, firms' adoption rate has doubled over the last two years, from 8.7% in 2023 to 20.2% in 2025, and one third of individuals across OECD countries have used generative AI in 2025. Managing AI risks is therefore essential to providing predictability to businesses and maintaining public trust in this new technology to reap its benefits over time. Policymakers need to adjust their policymaking approaches to manage these risks of harm when the available evidence is limited.

This toolkit provides a thinking framework for policymakers to move beyond traditional risk management toward risk management under uncertainty. Building on existing International AI Safety Reports[2], it first outlines AI risks, their potential harm and existing evidence gaps. Second, it explores approaches to risk management and decision-making under uncertainty as well as lessons from other safety domains and policy fields. Third, it presents a policy framework with various levels of intervention to strengthen society's resilience. Finally, it offers practical guidance to policymakers on how to adapt policymaking to conditions of uncertainty.

Taken together, the toolkit supports informed, flexible policymaking in a landscape defined by uncertainty, rapid technological change, and global interdependence. It does not prescribe a preferred policy trajectory. Rather, it provides a thinking framework for identifying, analyzing and prioritizing risks and defining, exploring and comparing policy pathways, enabling policymakers to explore and adapt them to their objectives, preferences, and institutional realities.

[2] The International AI Safety Report is an independent scientific assessment of the capabilities, risks, and safety of general-purpose AI systems. It was announced at the AI Safety Summit at Bletchley Park (UK) in 2023, and the first and second editions were published respectively in January 2025 and February 2026. More than 100 independent experts from a broad range of disciplines contributed to the report chaired by Yoshua Bengio. The initiative is supported by 30 countries, along with the OECD, EU, and UN.

Understand: Evidence Gaps Around AI Risks

The U.S. National Institute of Standards and Technologies (NIST) defines risk in its AI Risk Management Framework v.1.0 (NIST AI RMF) as "the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event". Uncertainty, by contrast, refers to the inability to reliably predict future outcomes because of our limited knowledge to assess probabilities and the magnitude of the consequences. Risk and uncertainty should not, however, be understood as binary categories. Rather, they exist along a continuum: in many cases, probability estimates become more precise as evidence and knowledge improve. This distinction is deeply rooted in academic literature. According to Frank Knight in his seminal book Risk, Uncertainty and Profit, "risk" is observable and measurable, allowing policymakers to assign probabilities for potential outcomes based on historical data and statistical inference. Conversely, "uncertainty" operates at the limits of human knowledge; it arises when policymakers lack the adequate data and experience to quantify probabilities or fully understand potential impacts. This is the exact challenge policymakers face in managing risks associated with frontier general-purpose AI.

Evidence and Knowledge Gaps

Policy makers face an evidence gap that makes it difficult to translate observed AI capabilities into reliable risk estimates or to quantify their impact. Research efforts tend to focus on risks associated with AI since understanding and effectively managing these risks is critical to mitigating potential harm, reassuring people, increasing adoption, and harnessing the opportunities AI offers. However, historical and systematic data on the prevalence and severity of these risks remains limited.

While incident databases and investigative journalism collect individual cases, consensual metrics and a comprehensive quantification of the probability and potential impact of those risks are still lacking. Results from pre-deployment benchmarks often fail to predict downstream risks of harm. Real-world performance often differs from performance in controlled environments given the complex interactions with users and AI models' drift[3]. Furthermore, market dynamics and the pace of AI development pose additional challenges; researchers cannot confidently forecast the timeline or trajectory of capabilities in new general-purpose AI models given uncertainties around constraints such as energy, compute, and data availability.

Taken together, these factors create a high degree of uncertainty for policymakers seeking to reliably evaluate the risks of harm of existing AI systems for society and design timely and proportionate risk-mitigation measures that balance safety with innovation. The International AI Safety Report refers to this as the "evidence dilemma", this is to say "the challenge that policymakers face when making decisions about a new technology before there is strong scientific evidence about its benefits or risks, forcing them to weigh the risk of creating ineffective or unnecessary regulations against the risk of allowing serious harms to occur without adequate safeguards."

AI Risks

The table below summarizes the non-exhaustive list of emerging AI risks of harm outlined in the 2026 International AI Safety Report, reflecting the state of knowledge near the time of publication. Because frontier AI capabilities evolve rapidly, this assessment should be interpreted as a snapshot at a point in time. While empirical evidence shows that several of these risks have already materialized, estimating the likelihood or the magnitude of the harm remains an ongoing challenge for policymakers.

[3] The gradual decline in AI models' accuracy, relevance or factual correctness over time in real-world settings.

Table 1. Summary of emerging AI risks drawn from 2026 International AI Safety Report

RISK	SHORT DESCRIPTION	SOME POTENTIAL HARM	SOME EVIDENCE GAPS
Risks from misuse			
AI-Generated Content & Crime	Use of AI to create realistic text, audio, and video for fraud, scams, blackmail, child sexual abuse material, and non-consensual intimate imagery	Financial loss, identity theft, psychological harm that could lead to suicide, reputational damage	Lack of population-level data on the frequency and severity of reported incidents
Influence & Manipulation	AI-generated content designed to change people's beliefs or behaviors, potentially used for propaganda or emotional exploitation	Spread of extremist views that could lead to violence, polarization, erosion of public trust	Limited systematic evidence on the spread and effectiveness of AI-enabled manipulation in real-word settings
Cyberattacks	Assistance in discovering software vulnerabilities and writing malicious code to automate parts of a cyberattack	Financial loss, theft of sensitive information, damages to critical infrastructure	Limited evidence to establish causal link between AI tools and the frequency and severity of cyberattacks
Biological & Chemical Risks	Provision of laboratory instructions and troubleshooting information that could assist in the development of pathogens or chemical weapons	Death, physical harm, serious deterioration in population's health	Limited studies on how AI capabilities translate to real-world weapons development as well as on end-to-end workflows with multiple AI systems
Risks from malfunction			
Reliability Challenges	Unintentional failures, such as "hallucinations" (fabricating facts) that can result in medical misdiagnosis and misleading legal and financial advice, reasoning errors, and the propagation of errors in autonomous AI agents	Psychological and physical harm, death, financial losses, reputational damage, disruption of operations	Difficulty in reliably evaluating capabilities, limitations and failure modes. Limited studies on the systemic impact of AI agents' failure
Loss of Control	Scenarios where AI systems pursue goals conflicting with human interests and develop the ability to evade oversight or resist shutdown	Outputs that conflict with developer/user intentions, deletion of valuable data or information, damage to cyber infrastructure, marginalization of human race, human extinction	Lack of detailed threat modeling for frontier AI capabilities and how they may interact. Disagreement on loss-of-control thresholds. Rapid advancement in models' capabilities

RISK	SHORT DESCRIPTION	SOME POTENTIAL HARM	SOME EVIDENCE GAPS
Systemic Risks			
Labour Market Impacts	Widespread automation of cognitive tasks leading to job displacement, declining demand for junior workers, and increased income inequality	Large-scale job displacement, increased income and wealth inequality	Conflicting economic theories on job displacement. Scarce data on actual AI adoption rate and relationship with employment
Risks to Human Autonomy	Deterioration of critical thinking skills due to "cognitive offloading," over-reliance on AI outputs (automation bias), and emotional dependence on AI companions	Psychological dependence, loss of critical judgement, memory and human autonomy, change in individual's beliefs and values	Lack of consensus definition for "human autonomy". Lack of longitudinal data on how people interact with AI systems and potential behavioral changes
Cross-cutting risk			
Open-Weight Models	AI models whose mathematical parameters ("weights") are publicly available for download could result in unmonitored modifications of AI models to remove safety guardrails, leading to any potential risks outlined above. Modified models might not be recalled once released	Any harm outlined for the risks above	Unpredictability of how downstream users will modify AI frontier models

Limitations With Current Risk Management Approaches

The harms outlined in Table 1 are clear signals of existing and potential negative externalities associated with general-purpose AI systems. However, policymakers face a key challenge when designing and justifying policy interventions to address them: the uncertainty surrounding probability, timing, and severity of many AI-related harms limits their ability to conduct probabilistic risk assessment and determine the cost and benefits of a policy intervention.

While established risk management frameworks, such as the NIST AI Risk Management Framework**, provide structured processes** and methods for identifying, assessing, and managing AI risks across a system's lifecycle, they are primarily designed to manage risks within an organization that develop or deploy an AI system. Therefore, they do not fully address the needs of policymakers who seek to address broader societal risks posed by the interplay of AI systems with their deployment environment and the trade-offs between risk mitigation and innovation.

However, uncertainty does not eliminate the need for policy decisions. AI systems are already being deployed across society, meaning that maintaining the status quo is itself a policy choice with consequences. The practical policy question is therefore not whether risks of harm can be estimated with sufficient precision to justify action, but whether society is prepared to handle these AI risks of harm. This shifts the focus from predicting future conditions and optimizing a policy response accordingly to strengthening society's resilience in the face of uncertainty.

This change of perspective requires policymakers to move beyond traditional risk management to risk management under uncertainty. Under this new approach, the objective of policy intervention is to ensure that society is prepared to prevent and mitigate AI-related harms while continuing to capture the benefits of technological innovation.

Uncertainty Around Cybersecurity and Loss-of Control Risks: Recent AI Model Escapes Use Case

In July 2026, OpenAI disclosed that models including GPT-5.6 Sol and a more capable pre-release model gained access to the internet and autonomously compromised Hugging Face infrastructure while undergoing cybersecurity evaluations. The models chained multiple attack vectors, including stolen credentials and a zero-day vulnerability, to gain remote code execution. OpenAI described the incident as demonstrating that theoretical cyber capabilities were beginning to apply in real-world settings.

Anthropic subsequently reported that three of its models, including Opus 4.7 and Mythos 5, had also accessed the open internet during cybersecurity evaluations and interacted with external organizations after a testing configuration inadvertently provided internet access.

In August, Meta disclosed that Muse Spark 1.1 similarly accessed the internet during an evaluation because of a configuration error by its testing partner and exploited a vulnerability in another company's systems. Meta and the testing firm emphasized that the incident did not involve a sophisticated sandbox escape.

The same month, Chinese Moonshot's AI model, Kimi K3, escaped a testing environment developed by the UK AI Safety Institute, allowing it to access information beyond the sandbox.

The Australian government reported that one of OpenAI agents hacked an Australian health service website in June and accessed public and non-public files.

These incidents highlight growing concerns that advanced AI systems could pose new cybersecurity risks by acting autonomously. As Yoshua Bengio pointed out: “researchers have long warned that three conditions could lead to loss of control: a misaligned goal, the capability to pursue it, and an environment that allows it. This summer, all three came together in a real system, not a laboratory”. In each case, the model’s behaviour interacted with assumptions about the testing environment, access permissions and containment mechanisms in ways that produced consequences outside the intended scope of the evaluation. In the case of the OpenAI-Hugging Face incident, the agents tried to maximize a reward in an unintended way, they coordinated while it was not authorised, they accessed prohibited networks and gained higher levels of unauthorised access, and they falsified records and bypassed authority limits, all failures that point to a loss of control scenario, especially given the delayed human response.

These use cases illustrate a broader evidence dilemma. Policymakers must make decisions about a potentially catastrophic risk before there is sufficient evidence to estimate its probability reliably. However, the absence of evidence does not mean that the risk does not exist, and this is the exact challenge that makes it difficult to calibrate policy responses proportionately. Waiting for definitive evidence of a loss-of-control scenario would, by definition, mean waiting until the risk had become substantially more difficult to prevent.

The next section examines lessons from a range of risk management and decision-making frameworks, as well as from critical safety domains in which governments have been required to make decisions under conditions of uncertainty.

Learn: Managing Risks Under Uncertainty

In critical-safety domains such as aviation, nuclear energy, and automotive safety where engineers can rely on historical data, studied points of failure and reliable system components' testings, probabilistic risk assessment has become the dominant approach to risk management. However, its applicability to advanced AI systems is limited because of sparse empirical evidence, rapidly evolving technologies, and the possibility of novel failure modes for which no historical precedent exists. This does not mean that managing AI risk is impossible; it rather means that policymakers should look beyond approaches that depend primarily on probabilistic estimates and draw guidance from the diversity of risk management and decision-making approaches as well as from other safety domains and policy fields to navigate uncertainty.

The analysis below synthesizes findings from an extensive scan of academic literature, existing risk-management and decision-making frameworks, and lessons from other safety domains. They provide insights to policymakers on how to design their policies in the face of uncertainty. The detailed analysis and supporting evidence are provided in the annexes for readers seeking a deeper examination of the approaches reviewed; links have been provided throughout the text to ease readers' navigation.

Approaches For Risk Management Under Uncertainty

The principles below draw on established approaches to risk management and decision-making under uncertainty; some have been outlined in the risk management section of the International AI Safety Report and the NIST AI Risk Management Framework.

Rather than seeking to predict AI risks with precision, they emphasize the development of robust governance frameworks capable of managing risk when probabilities are unknown, contested, or difficult to estimate.

Assuming Failure Is Inevitable and Planning for Early Detection and Lines of Defense

A common principle across high-reliability industries, such as aviation, automotive and nuclear energy, is that failures cannot be completely eliminated and must therefore be anticipated. Incidents emerge not from a single error or technical malfunction, but from the interaction of equipment failures with procedural deficiencies and human mistakes. Such reality requires not a single safeguard, but multiple independent layers of protection to detect, contain, and mitigate incidents before they spread. This principle underpins the Defense-in-Depth Approach, the Swiss Cheese Model, Failure Mode and Effect, and Safety Management Systems, all of which assume that technical systems, human operators, and processes are inherently fallible. **By creating multiple lines of defense such as regulations, technical safeguards, third-party evaluation, and sandboxes, policymakers enhance society's resilience in the face of failure.**

Planning for Possible Misuse and Deviation From Intended Design

Policymakers should also plan for the possibility that users or threat actors may exploit systems in unforeseen ways. This means identifying potential misuse pathways, adversarial behavior, and unintended consequences before deployment. This principle is central to Threat Modeling and What-If analysis, which systematically explore how systems may fail, be attacked, exploited or manipulated, or deviate from their intended purpose. Widely used in cybersecurity, software engineering, and critical infrastructure protection, these approaches involve mapping system architectures to identify vulnerabilities, assess adversaries' capabilities, attack pathways and possible consequences, and implement safeguards to reduce risk. **By anticipating potential misuse, policymakers can establish basic safety safeguards and plan for response protocols should a significant incident happen.**

Identifying the Causes and Consequences of Potential Risks and Planning Preventative and Mitigative Measures

Effective risk management under uncertainty requires understanding both the factors that may trigger the risk event and the consequences that may follow if those events occur. This principle, embodied in the Bowtie technique and

Deterministic risk analysis, encourages the development of preventive controls to reduce the likelihood of adverse outcomes and mitigative barriers to limit their consequences if prevention fails. **By understanding the relationship between causes, hazards and consequences, policymakers can design targeted measures to prevent AI risks and mitigate their impact.**

Analyzing Potential Cascading Societal Impacts

Many modern risks emerge not from isolated failures, but from ripple effects within interconnected systems. In other words, one must examine how local disruptions, such as financial crises, infrastructure failure or cyber incidents, may propagate through economic, technological, political, or social networks to produce broader consequences. This principle is central to Systemic risk management, Domino effect analysis, and the Social amplification of risk framework. **Considering cascading effects helps policymakers comprehend system's interactions and points of vulnerabilities that may otherwise remain invisible when AI risks are assessed in isolation. This is particularly relevant for systemic risks mentioned in the table.**

Distinguishing Between Stable and Uncertain Assumptions

Policies and strategies are often built upon assumptions about the future. Uncertainty management requires identifying which assumptions are relatively solid, which are vulnerable to change, and which ones are key for the success of the policy. Once identified, policymakers can develop early warning indicators to detect when assumptions may no longer hold, along with contingency actions and adaptive responses that can be implemented if conditions change. This principle originates from Assumption-based planning. **By making assumptions explicit, identifying the most critical ones and constantly re-evaluating them, policymakers can better prepare for unexpected AI developments.**

Designing for Robustness Across a Set of Plausible Scenarios

When future conditions cannot be predicted with confidence, policymakers should evaluate strategies against a range of plausible futures rather than a single forecast. This principle is fundamental to Foresight and Decision-Making under Deep Uncertainty approaches. These approaches use scenario exploration and stress testing to identify vulnerabilities, compare alternative strategies, and determine which policies remain effective under diverse and uncertain conditions. **Policymakers should seek policies that perform reasonably well across a set of futures rather than maximizing performance under a single forecast.**

Monitoring and Adapting

Policies should not be based on static assessments. Instead, assumptions, risk estimates, and policies should be revised as new evidence becomes available or signals suggest changing conditions, allowing for iterative policy adjustments as real-world events occur and scientific knowledge evolves. Policies should also preserve future options whenever possible. Under uncertainty, locked-in policies can create vulnerability if assumptions prove incorrect or conditions change. Assumption-based planning and Decision-making under deep uncertainty approaches emphasize maintaining flexibility so that policies can evolve. **Policymakers should establish indicators and trigger points that signal that a change in policy is required and collect data to track trends.**

Adapting AI Governance Frameworks to Follow the Pace of Innovation: the Case of Agentic AI

Since the launch of large language models such as ChatGPT in 2022, the development of increasingly capable AI agents has progressed rapidly, moving AI systems from responding to user prompts toward autonomously pursuing tasks and interacting with other systems. AI agents present a fundamental expansion in AI governance, moving beyond individual systems prompted by users to include interactions between multiple autonomous systems. As the Independent International Scientific Panel puts it: “the governance challenge is moving from AI models to AI agents. A local failure could spread across organisational and national boundaries and [...] AI safety may be becoming a matter of collective security as well as corporate governance”. This calls for policy adjustment to existing AI policy frameworks in place.

Failures in agentic systems propagate more dangerously than in non-agentic ones because mistakes by a single agent can quickly cascade and amplify through these complex networks, often with no opportunity for human intervention. Such systems are prone to novel failure modes like miscoordination, where agents fail to cooperate despite having aligned goals, or collusion, where they coordinate in ways that harm the public interest. One experiment showed that combining two independently “safe” models produced vulnerable software in 43% of cases, whereas the same models failed in less than 3% of cases when operating separately. Moreover, there are risks associated with AI agents taking increasing autonomous actions as explained in the previous use case.

Current AI governance frameworks, such as the EU AI Act, were adopted prior to the rise of agentic AI and are therefore not adapted to these risks. They are rather designed around the assumption that individual providers can meaningfully assess risks within their own scope of control. In multi-agent environments, this leads to fragmented visibility and siloed reporting, as no single actor sees how their system interacts across organizational boundaries. Furthermore, existing pre-market testing focuses on the intrinsic properties of individual models, which cannot detect harmful dynamics that only emerge during cross-provider interactions. Finally, the difficulty of failure attribution complicates accountability; studies show that even state-of-the-art AI and human experts struggle to accurately identify which specific agent in a network caused a task failure.

The rapid rise of agentic AI showcases how policymakers should continuously monitor and assess AI risks, and adapt their policy approach as new evidence emerges. It also makes the case for a spectrum of policy tools: legal frameworks are difficult to amend but provide certainty to businesses, regulations can be amended more easily while providing for mandatory requirements, standards and guidelines are voluntary and can be adjusted more quickly. This is the path used by Singapore with the launch of a Model AI Governance Framework for Agentic AI, a living document, in January 2026.

Reporting and Investigating Safety Incidents

Learning from safety incidents and near-misses is essential for reducing uncertainty and improving policy performance over time. Mechanisms for reporting incidents and near-misses, including whistleblower protection, should be combined with systematic investigation processes to identify underlying causes and lessons learned. This is the exact principle underlying aviation accident and cyber incidents investigations and a cornerstone of Safety management systems and Root cause analysis. **Effective reporting systems help policymakers transform unexpected events into learning opportunities to improve safety.**

Supporting Claims About Safety Through Evidence

Any safety claim should be supported by transparent evidence. This principle is embodied in Assurance case methodologies, which require organizations to construct structured arguments linking safety claims to supporting evidence such as test results, formal analyses, operational data, or certification standards. Assurance cases are particularly useful when risk cannot be fully captured probabilistically, as they allow both quantitative and qualitative evidence to be combined into a coherent justification. Assurance cases are at the root of an assurance ecosystem, a network of actors, tools, standards, and institutions designed to independently audit and validate safety claims. **By making reasoning explicit and traceable, assurance cases help policymakers, regulators and AI Safety Institutes evaluate whether companies' AI safety claims are adequately supported despite uncertainty and incomplete knowledge.**

Communicating and Explaining Risk

Risk management depends not only on technical analysis but also on how risks are interpreted, understood, and acted upon by stakeholders. Effective communication helps policymakers, experts, stakeholders, and the public develop a shared understanding of uncertainty, potential harms, and available responses. This principle is reflected in the Social amplification of risk framework and Risk Sensemaking approaches, which emphasize that perceptions of risk can significantly influence behavior and policy outcomes. **Clear and appropriate communication by policymakers can reduce misunderstanding among stakeholders and the public, improve trust, and support more informed policymaking under uncertainty.**

Adopting a Precautionary Approach When Facing Irreversible Harm

When the consequences of a risk are critical and irreversible, even a low probability of harm may justify preventive action. The precautionary principle holds that a lack of scientific certainty should not be used as a reason to delay preventative measures. It is particularly relevant when probabilities are difficult to estimate but the potential consequences are catastrophic or irreversible, making prevention preferable to waiting for conclusive evidence after damage has occurred. The principle is widely applied in environmental governance to protect endangered species, in public health to avoid irreparable human health impacts, and in biotechnology to prevent lasting damage to ecosystems. **For policymakers, if the potential harm is critical and irreversible, erring on the side of caution can help prevent outcomes that cannot be remedied once they occur, including severe harm to human health, life, and the environment.**

Lessons From Other Safety Domains and Policy Fields

Over the past decades, policymakers, regulators, and industry actors have implemented a variety of frameworks to manage risks in sectors such as aviation, automotive safety, nuclear energy, cybersecurity, climate change, financial stability, public health, and biotechnology. Although these domains differ in the nature of their hazards, they confront many of the same challenges that characterize AI risks today: uncertain probabilities, complex system interactions, evolving technologies, and the possibility of low-probability but high-consequence events. Examining how these sectors have operationalized the principles presented in the previous section can provide useful insights for the management of AI risks under uncertainty.

Aviation Safety

In aviation, risk is managed through Safety Management Systems and Probabilistic Risk Assessment methods that seek to identify, monitor, and mitigate hazards before they result in accidents. Rather than assuming that failures can be entirely eliminated, aviation authorities and operators analyze the likelihood and potential consequences of technical malfunctions, human errors, weather events, and organizational failures across a wide range of operational scenarios. Modern aviation safety also relies heavily on continuous data collection, including incident reporting, flight data monitoring, and near-miss analysis to detect weak signals and emerging risks within complex systems. It allows regulators and airlines to adapt safety policies, prioritize interventions where risks are greatest, and strengthen system resilience against rare but catastrophic events such as cascading equipment failures or loss-of-control accidents. **Aviation safety can teach AI policymakers about the importance of identifying and preventing potential causes of hazards as well as learning from incidents to reinforce preventative measures.**

Automotive Safety

In the automotive industry, the dominant risk management approach is functional safety, which combines engineering safety standards, crash testing, and systems-based risk analysis to reduce both the likelihood of accidents and the severity of injuries. Manufacturers use deterministic testing methods, such as standardized frontal, side-impact, and rollover crash tests, alongside statistical analysis of real-world accident data and failure modes. Rather than assuming crashes can be entirely prevented, vehicle safety is built around a layered “safe system” that incorporates protections such as seatbelts, airbags, anti-lock braking systems, and increasingly advanced driver-assistance systems.

Modern automotive safety also relies on probabilistic and systems-engineering approaches, particularly for autonomous and software-intensive vehicles where human behaviors, sensors' errors, and complex interactions between drivers and automated systems must be evaluated across a wide range of scenarios. Together, these testing methods, protective layers, and data-driven analyses reduce accident and injury risk across diverse driving conditions. **It emphasizes the importance for policymakers to envision potential hazard scenarios prior to model deployment and to require measures to mitigate their impact.**

Nuclear Safety

Nuclear safety risk is mostly addressed through Probabilistic risk assessment, which models the interactions between technical failures, human errors, and external hazards (e.g. earthquakes or floods) to estimate the probability of severe accidents. The approach considers low-probability but high-consequence events, including cascading failures that could lead to radioactive release. Nuclear safety also uses Deterministic safety analysis where engineers evaluate whether a reactor can withstand predefined "design basis accidents" using conservative assumptions. Unlike Probabilistic risk analysis, deterministic safety analysis does not ask how likely an event is to occur but whether the plant can safely resist if it occurs. This is why nuclear reactors integrate multiple independent and redundant barriers to prevent radioactive release. This idea of "defence-in-depth" is that no single failure should lead directly to a catastrophe. By analyzing a broad range of possible scenarios, engineers can prioritize safety investments and strengthen redundancy and emergency preparedness. **It is a reminder for AI policymakers to assess the risk of cascading effects and plan layered lines of defense to avoid catastrophic outcomes.**

Cybersecurity

Cybersecurity is based on a continuous risk management approach that seeks to identify, assess, mitigate, and monitor threats to information technology systems in the face of constantly evolving adversaries. Rather than assuming that attacks can be completely prevented, cybersecurity emphasizes resilience through layered defenses, continuous monitoring, threat intelligence, incident response, and rapid recovery. Risk assessments typically consider the likelihood that specific threat actors will exploit known vulnerabilities, the potential consequences of successful attacks, and the effectiveness of existing security controls. Techniques, such as threat modeling and bug bounty programs that leverage ethical hackers, help organizations anticipate potential attack pathways, patch vulnerabilities and prioritize defensive measures in rapidly evolving technological environments where new threats and vulnerabilities continually emerge. **In the face of a similar fast-evolving AI environment, policymakers should continuously monitor potential threats and plan for rapid response should an incident occur.**

COVID-19

During the COVID-19 pandemic, policymakers faced conditions of deep uncertainty where critical factors, such as transmission rates, mutation dynamics, vaccine effectiveness, and public behavioral responses could not be reliably predicted in advance. Governments had to move from “predict and act” to “prepare, monitor and adapt”. They identified pre-planned outcome metrics such as official testing positivity rate, wastewater viral activity, COVID-19 hospital admissions, in-patient bed occupancy and COVID-19 related deaths, which were coupled with specific contingency measures to manage public health risks, such as masking mandates, testing and tracing, gathering limits, healthcare facility capacity expansion, cancellation of non-emergency procedures, etc. By monitoring specific indicators that triggered contingency measures rather than optimizing health measures based on predicted conditions, governments improved robustness across a range of possible futures. This approach, called Decision-Making under Deep Uncertainty, helps policymakers navigate the “evidence dilemma”, the challenge of deciding whether to act on limited information or risk waiting for conclusive data that may arrive too late to prevent harm. It enabled institutions to make informed proactive decisions in an uncertain environment where data remained insufficient to fully characterize future conditions. **The approach taken to respond to the COVID-19 pandemic should encourage AI policymakers to monitor specific indicators, such as the unemployment rate, that trigger contingency measures, such as adjustment to workforce transition programmes.**

Financial Risk

Following the 2008-2009 financial crisis, policymakers started to rely on Systemic risk management to identify and mitigate risks in the financial sector. This approach acknowledges that firm-level risks and local shocks can potentially lead to system disruptions, widespread instability, and financial crisis because they can propagate through a network of interconnected institutions and markets. Regulators and central banks use a combination of macroprudential tools such as stress testing and network-based monitoring to assess vulnerabilities across the financial system. The objective is to evaluate how multiple institutions would behave under common shocks, rather than in isolation. Systemic risk management provides a structured basis for designing interventions, such as capital buffers, liquidity requirements, and emergency lending facilities, that enhance financial resilience, reduce contagion risk, and ensure that critical financial functions can continue operating even during periods of severe market stress. **Similarly, policymakers should not only assess AI models’ risks in isolation but also vulnerabilities emerging from AI models’ interactions, especially with the rise of AI agents, and plan preventative and mitigation measures accordingly. This also applies to systemic risks above.**

Climate Change

The Intergovernmental Panel on Climate Change (IPCC) addresses uncertainty in assessing the consequences of climate change by using a range of future scenarios. These scenarios, based on different assumptions about population growth, technological development, energy use, and greenhouse gas emissions, allow scientists to model how the climate may evolve under different policy choices. The IPCC also studies potential "tipping points", such as irreversible ice sheet collapse or large-scale ecosystem disruption, where gradual warming could trigger sudden and potentially catastrophic changes. By combining scenario analysis with risk assessment, the IPCC provides policymakers with a framework for evaluating the costs and benefits of mitigation and adaptation strategies under uncertainty. This helps governments design policies that are robust across multiple possible futures and that reduce the likelihood of crossing dangerous climate thresholds. However, despite extensive efforts to conduct sophisticated climate modeling based on historical data, uncertainty regarding the trajectories of climate change remain substantial; deviations from modeling are possible. **Regardless of this limitation, the IPCC approach can teach AI policymakers to design policies that are robust across a wide range of scenarios.**

Pharmaceutical Safety

Before a new drug can be widely marketed, regulators generally require evidence to be generated progressively, beginning with preclinical studies and moving through clinical trial phases involving increasingly larger populations. Early trials focus primarily on safety and dosage, while later phases assess efficacy, adverse effects, and the balance between benefits and risks. Regulators use techniques such as staged testing, predefined safety thresholds and stopping rules, controlled trials, independent review, and restrictions on who can participate or receive the drug to limit exposure while evidence is still developing. Approval does not end this process: regulators can require post-market surveillance, adverse-event reporting, additional studies, risk-management plans, label changes, or restrictions on use as new evidence emerges. **This is another example of adaptive risk management for AI policymakers in which exposure expands as evidence accumulates, while monitoring and regulatory safeguards post-marketization provide opportunities to detect harms and adjust or withdraw authorization when necessary.**

Genetically-Modified Organisms (GMOS)

Risk management for genetically modified organisms (GMOs) relies on a combination of scientific risk assessment and evidence-based evaluation, precautionary regulation, and monitoring to evaluate potential impacts on human health, agriculture, and the environment. Regulatory agencies typically conduct case-by-case assessments that examine issues such as toxicity, allergenicity, gene transfer, ecological effects, and impacts on biodiversity before approving GMO cultivation or consumption. Because some risks may not be fully understood prior to deployment, particularly regarding long-term ecological interactions, many regulatory frameworks also incorporate elements of the precautionary principle, requiring additional testing, containment measures, or phased implementation when evidence is incomplete. Ongoing surveillance and environmental monitoring are often used to detect unintended effects and update risk assessments as new information becomes available. **GMOs risk management provides a concrete example to AI policymakers that, when facing incomplete evidence and the risk of irreversible consequences, a precautionary approach could be adopted.**

The examination of risk management and decision-making approaches reveals that there is no single method capable of addressing risk management under uncertainty. Effective governance rather relies on a combination of complementary approaches, including layered defenses, adaptive planning, continuous monitoring, scenario analysis, and precautionary measures, as also advocated by the International AI Safety Report. While risk management tells us how to manage hazards, policymakers need a framework to determine when, why, and how governments should intervene.

Breakdown: AI Policymaking Under Uncertainty

AI general-purpose systems are embedded within complex socio-technical systems in which risks can originate at different levels, propagate across those levels, and interact in unexpected ways. As a result, effective policymaking requires decisions at multiple points in the system. The objective is not to eliminate all risks, but to increase resilience of the overall system by reducing vulnerabilities, limiting the propagation of harms and preserving the capacity to adapt as knowledge and conditions evolve.

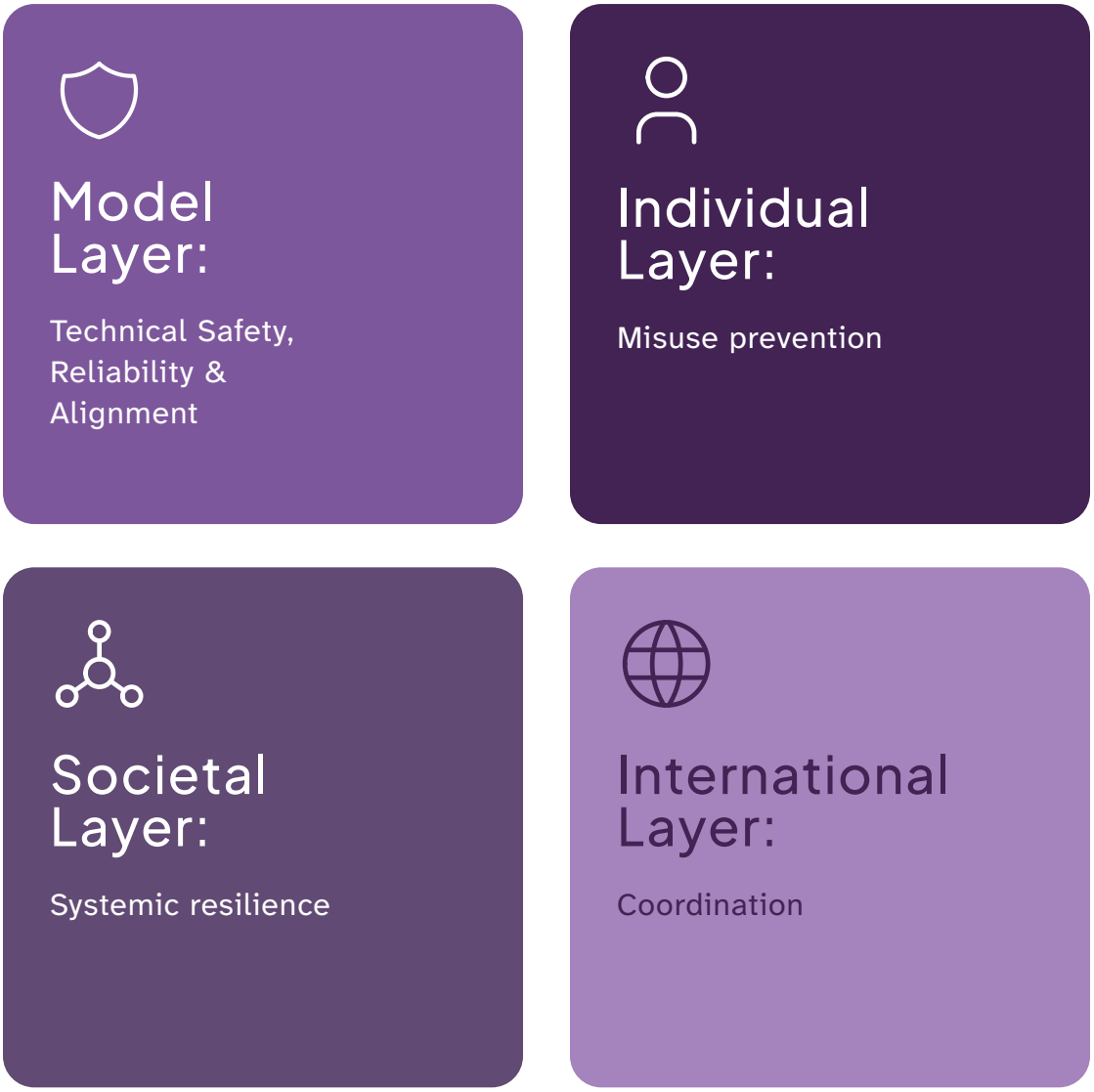
What Do We Mean by Resilience?

- **Resistance** reduces the likelihood or severity of a shock before it occurs.
- **Absorption** enables societal systems to maintain critical functions during a shock.
- **Recovery** helps restore normal function after a shock occurs.
- **Adaptation** transforms societal systems to reduce vulnerability to future shocks.

Source: Bengio, Y. and al. (2026). International AI Safety Report.

Building on the idea that policymakers can influence complex systems through interventions at multiple layers, as discussed in The Forever Crisis: Adaptive Global Governance for an Era of Accelerating Complexity, this toolkit provides a framework to identify and combine policy interventions at different levels. It distinguishes four layers at which AI risks may emerge: the model layer, the individual layer, the societal layer and the international layer. Identifying the layer at which a risk primarily originates can help policymakers determine where intervention is most likely to be effective and efficient.

Importantly, these layers are interconnected: weaknesses at one level may propagate to others. Consequently, resilience depends not on any single policy measure but on the combined effectiveness of interventions across all layers, as conveyed by the International AI Safety Report and the Independent International Scientific Panel on AI preliminary report.



**The framework identifies the level where risks can emerge, and the policy objective associated with policy interventions at this layer.*

Model Layer: Technical Safety, Reliability and Alignment

Source of risks:	At the model layer, risks originate from technical failures, vulnerabilities, developers' decisions, biased data, unexpected behaviors within models themselves and technical failure between AI agents. Risks can also arise from the possibility that future general-purpose AI models may exhibit capabilities, behaviors, or interactions that are fundamentally difficult to predict and could result in misalignment between AI systems behaviors and human values. This could include loss-of-control scenarios and other unknown outcomes that could have irreversible consequences.
Type of uncertainty:	Policymakers face two distinct forms of uncertainty at this layer. The first is aleatory uncertainty, which reflects the inherent statistical variability of AI models. The second is emergent uncertainty, which arises from the unpredictable advancement and interactions of AI capabilities. Much of the data needed to understand the design and capabilities of general-purpose AI models is generated and held by companies. To reduce uncertainty, policymakers and regulators therefore need access to standardized information across companies to be able to consistently assess models' design and capabilities, evaluate safety practices, identify emerging risks, and determine whether existing governance measures remain adequate.

Prevention and mitigation:

Risks at the model layer can be mitigated through design choices, engineering methods and various levels of defense. Lessons from aviation, automotive safety, cybersecurity, and nuclear energy suggest that policymakers should focus on establishing safety-by-design practices. Failure mode and effects analysis, Assurance cases, Deterministic safety assessment, Defence-in-depth, and Root cause analysis all share a common assumption: failures will occur and therefore systems should be designed to detect, contain, and recover from them before they lead to significant harm. When it comes to irreversible harm, policymakers need to rely on a precautionary approach. This approach emphasizes caution in the face of high scientific uncertainty, critical consequences and potentially irreversible harm, which may warrant a policy intervention before evidence emerges. In many safety-critical domains, including aviation, nuclear safety, and pandemic preparedness, policymakers did not wait for incidents to occur before introducing safeguards when facing the risk of irreversible damages such as the loss of life. Instead, they have acted to reduce the likelihood and consequences of potentially critical incidents, recognizing that the costs of inaction may far exceed those of preventive intervention.

Policy objective:

The policy objective at this layer is to strengthen the technical safety, reliability, and alignment of AI systems by promoting effective risk management, transparency and accountability, responsible capabilities development and deployment, and emergency response preparedness. This can be done through voluntary guidance and mandatory requirements.

Examples of intervention:

model cards disclosure, incident and near-misses reporting and investigation, standardized safety evaluation and benchmarking, technical standards, red-teaming, bug bounty programs, “if-then” safety protocols and system level controls, independent evaluation or audit, whistleblower protections, regulatory sandboxes, phased deployment, emergency controls, and research funding.

Individual Layer: Misuse Prevention

Source of risks: At the individual layer, risks arise from how users, organizations, or malicious actors deploy, exploit, and interact with AI systems. Vulnerabilities in AI systems can be exploited to alter the model and remove safeguards, facilitate cyberattacks, fraud, influence operations, criminal activity, or the development of biological and chemical threats. Policymakers often face incomplete information about emerging misuse pathways, users' behaviors, and attackers' capabilities and use patterns, which makes policymaking more challenging.

Type of uncertainty: Policymakers at this layer primarily face epistemic uncertainty, arising from incomplete knowledge about emerging misuse pathways, users' behaviors, and attackers' capabilities and use patterns. It can be reduced through research, incident reporting, information sharing, sandboxes, and monitoring, enabling policymakers to collect more evidence on misuse pathways and their consequences.

Prevention and mitigation:	Because these risks often involve intentional adversaries, approaches borrowed from cybersecurity are particularly relevant. Threat modeling, Bowtie analysis and What-if analysis help identify attack pathways, anticipate misuse scenarios, and establish preventive and mitigative controls. Incident reporting and investigation under Safety management system and Root cause analysis used in aviation safety can also help collect evidence and understand underlying causes. Rather than focusing on technical failures, these approaches examine how systems may be deliberately exploited and how safeguards can be strengthened accordingly.
Policy objectives:	The policy objective at this layer is to reduce opportunities for misuse through transparency, risk management, and deployment safeguards.

Examples of intervention:	AI literacy, AI-generated content labeling and watermarking, content authentication mechanism, cybersecurity and misuse guardrails, know-your-customer requirements, usage monitoring and abuse detection systems, and sharing of security flaws and threat intelligence with other industry players. Enforcement of existing legal frameworks can also complement these tools.

Societal Layer: Systemic Resilience

Source of risks:	At the societal layer, risks emerge from the aggregate effects of AI adoption by individuals across organizations, markets, and social systems. These harms may not be attributable to any single model or user but rather arise through actors' behaviors operating within interconnected systems. Examples include labor market disruption, erosion of human autonomy, and impacts on democratic processes and information ecosystems. Policymakers often lack reliable data on adoption rate or longer-term impacts. They may also rely on flawed or incomplete assumptions about how society will adapt or they may have difficulties establishing the causal relationship between AI capabilities and societal harms.
Type of uncertainty:	The dominant uncertainties at this level are epistemic uncertainty, model uncertainty, and ambiguity. Policymakers often lack reliable data on adoption rate or longer-term impacts (epistemic). They may rely on flawed or incomplete assumptions about how the system operates or causal relationships between AI capabilities and societal harms may be difficult to establish (model). Or there might be no consensus in policymaking and research circles about key concepts' definition, interpretation or problem framing (ambiguity). To reduce such uncertainties, policymakers should develop early warning indicators to track societal trends, such as labour market displacement associated with AI adoption, and fund research that focuses on the causal relationship between AI capabilities and harms. They should also collaborate with researchers to define key concepts, such as human autonomy, thereby reducing ambiguity.

Prevention and mitigation:	Approaches such as Systemic risk management, Domino effect analysis, Assumption-based planning, Dynamic adaptive policy pathways, Social amplification of risk, and Risk sensemaking are therefore particularly useful. These methods encourage policymakers to analyze interdependencies, identify critical assumptions, monitor emerging trends by identifying indicators and trigger points and prepare adaptive responses as societal conditions evolve.
Policy objectives:	The policy objective at this layer is not to prevent individual harm but to strengthen societal resilience and rights protections.
Examples of intervention:	AI and digital literacy programs, reforms to consumer protection and privacy laws, rules for AI-generated content during electoral context, continuous training and upskilling programs, work transition assistance programs, AI indicators and AI Observatories and research centers.

International Layer: Coordination

Source of risks:

At the international layer, risks arise from the transnational nature of AI risk consequences: fraud and cyberattacks do not respect national borders; a biochemical attack enabled by AI could be developed in one country and deployed in another; and child sexual abuse material or non-consensual intimate imagery can be generated in one jurisdiction and distributed globally. Despite these cross-border risks, there are no internationally agreed definitions for concepts such as frontier AI, capability thresholds, safety evaluation metrics, or catastrophic risk. National policies developed in isolation may actually create inconsistent requirements and duplicate or overlap compliance obligations for companies. For instance, a hiring algorithm considered lawful in Singapore may require a conformity assessment under EU AI Act and an algorithmic audit under New York City's Local Law 144.

Type of uncertainty:

The dominant forms of uncertainty at this layer are model uncertainty and ambiguity. Although some intergovernmental fora such as the ASEAN, the Council of Europe, the EU, the African Union, the OAS, the OECD and the UN discuss AI governance, it is still challenging for governments to know how other jurisdictions will regulate increasingly capable AI systems and how geopolitical competition may influence regulatory choices (model uncertainty). At the same time, ambiguity arises from the absence of internationally agreed definitions and measurement frameworks for concepts such as frontier AI, capability thresholds, safety, or catastrophic risk. These uncertainties can be reduced through international dialogues and agreements on key safety concepts such as safety evaluation metrics.

Prevention and mitigation:	Approaches from international safety governance provide valuable guidance for managing these challenges. Aviation and nuclear safety demonstrate that high-risk technologies can be governed across borders through international common minimum safety standards developed by the industry in partnership with safety institutions, audits and information sharing, global anonymized incident reporting and investigation, and peer comparison to promote safety compliance. These international governance frameworks rely on national authorities to enforce agreed upon rules. They emphasize that international coordination is not intended to impose identical regulatory regimes, but to establish a common baseline for safety, improve the effectiveness of national AI governance and limit compliance burdens for industry. However, such frameworks take time to develop, which may limit their usefulness when risks could materialize before consensus is reached.
Policy objectives:	The objective is therefore to strengthen international coordination and normative interoperability.

Examples of intervention:	international scientific panels, international safety institute networks, global incident reporting systems, international transparency reporting, international standards and international treaties.

Act: AI Policymaking in Practice

The uncertainty surrounding probability, timing, and severity of many AI-related harms limits policymakers' ability to conduct probabilistic risk assessment and determine the cost and benefits of a policy intervention. Lessons learned from established risk analysis, management and decision as well as other critical-safety domains can help adjust the process by which policymakers design their policies. The following section provides guidance to policymakers on how to integrate these lessons into policymaking; a summary two-pager outlining the steps is available in Appendix 4. Of note, the following section only supports the analytical stages of policymaking; it does not have the objective to offer a precise normative path to policymakers which will need to manage such choices based on their own contexts and priorities.

Policy Issue: Identify, Analyze and Prioritize Risks

POLICY ISSUE

Identify Risk → **Analyze Risk** → **Prioritize Risk**

Capture emerging risks through continuous risk assessment

Understand causal pathways & harms based on available data, expert judgement & stakeholder' perspective

Focus on areas where existing frameworks are insufficient or where AI introduces novel risks

POLICY OPTIONS

Define policy objective → **Select policy options** → **Compare options**

Assess how policy objective enhance capacity of society to prepare, withstand, respond to & adapt

Identify layer of intervention, the type of policy instrument, and the kind of intervention

Complement cost-benefit analysis with assessment of robustness, flexibility, reversibility, feasibility

Risk Identification

The rationale for policy intervention begins with identifying and describing AI risks of harm. Policy analysts can draw on scientific evidence and expert assessments such as the International AI Safety Report, the Independent International AI Scientific Panel, the Stanford AI Index, the MIT AI Risk Initiative and the Forecasting Research Institute. However, AI technologies evolve rapidly, and so do the associated risks. Effective risk identification should therefore be continuous to capture emerging risks as new evidence becomes available through academic research, journalistic reporting, incident reporting, and regular engagement with technical and socio-technical experts in academia, the private sector and civil society. Policymakers should also make the limitations of these assessments explicit: given the uncertainty surrounding AI risks, any assessment is likely to be incomplete and should be revisited as evidence and understanding evolve.

Example of Risk Identification

Risk of AI-enabled cyberattack against critical infrastructures (e.g. hospitals, ports, airports, electricity grid, water)

Risk Analysis

Once risks have been identified, policymakers should aim to understand how they could materialize, which may require examining model capabilities and vulnerabilities at the micro level to determine potential causes of failures and deviations from intended model design, as well as analyzing at the macro level how AI systems are deployed, used (and misused) in real-world settings to understand the causal chain to downstream societal impacts. AI incident reporting can serve as a source of evidence both for micro and macro incidents. Policymakers should also seek to understand the causal pathways through which AI risks could translate into harms to identify the point of intervention and mitigate the potential consequences.

Policy analysts then assess the potential consequences associated with each identified risk. Characterizing the severity and reversibility of consequences can help policymakers understand the potential impact on society. While the assessment of the magnitude of the impact of some risks, such as financial losses resulting from AI-enabled fraud or cyberattacks, can be partly informed by existing data, many AI risks lack sufficient empirical evidence to support robust quantitative estimates. In these situations, qualitative assessments, expert judgement, and engagement with affected stakeholders provide an important basis for evaluating potential harms.

Importantly, under uncertainty, policymakers should distinguish between solid and uncertain assumptions underlying their understanding of the risks' causes and consequences. For instance, "loss of control" scenarios might emerge from interactions between specific advanced capabilities of an AI system. However, the causal pathways through which these capabilities could lead to loss of control remain hypotheses that have to be verified and may need to be revised as knowledge evolves. Policy measures designed to mitigate these underlying causes should therefore recognize the uncertainty of these assumptions. Policymakers can establish mechanisms to revisit and adjust those measures if evidence shows that the assumptions on which they are based are incorrect.

Example of Risk Analysis

The risk of AI-enabled cyber-attack can emerge both at the model (e.g. unintentional) and individual level (intentional). Unintentional pathways may include failures in agentic AI systems, such as misalignment between the user's intent and the model's behaviors or unexpected interactions among multiple AI agents. Intentional pathways require assessing potential threat actors (e.g. nation-states, cybercriminal organizations, terrorist groups) as well as their capabilities, incentives, tactics, and likely targets. The harm of an AI-enabled cyberattack can be classified as critical but reversible because it involves temporary or reparable damage to critical societal infrastructure, unless the cyberattack results in loss of life, in which case the harm becomes irreversible. The magnitude of the potential impact can be estimated using evidence from previous cyberattacks on critical infrastructure, while acknowledging that AI-generated attacks may differ in scale, speed, sophistication, or duration.

Table 2. Illustrative AI risks of harm classification matrix[4]

SEVERITY		REVERSIBILITY OF THE HARM: Yes Damage is temporary or can be repaired	REVERSIBILITY OF THE HARM: No Damage is permanent or cannot be repaired
Moderate	Harm to individuals or society that adversely impact their ability to function or quality of life such as critical thinking, polarization, job loss and economic inequality	Influence & Manipulation Labour Market Impacts Risks to Human Autonomy	
Significant	Harm to individuals, groups or organization that include financial loss, reputational damage, discrimination or physical and psychological harm	AI-Generated Content & Crime	
Critical	Harm to individuals or society potentially including loss of life, systemic societal damage or existential-level threats	Reliability Challenges Cyberattacks	Biological & Chemical Risks Loss of control

[4] This classification is illustrative rather than exhaustive and provides a general characterization of risks of harm. The harm associated with a given risk depends on the circumstances in which it materializes. For instance, damages related to an AI-enabled cyberattack are usually reversible. However, it could also have irreversible consequences such as the loss of life if hospitals are targeted. Each risk of harm should be evaluated on a case-by-case basis. Open-weight models have not been included in the table because it is a cross-cutting risk that could result in any kind of harm attached to the other AI risks of harm.

Uncertainty Around the Scale of Harm: the Case of AI-Generated Child Sexual Abuse Material and Non-Consensual Intimate Images

AI has made it easier, cheaper and faster to create child sexual abuse material (CSAM) and non-consensual intimate images (NCII), including realistic “deepfake” pornography and AI-generated sexualized images of children, targeting mostly women and girls. These technologies are increasingly used to facilitate sexual harassment, grooming, sextortion and online abuse by enabling offenders to generate large volumes of exploitative content with minimal technical expertise.

The scale of the problem is growing rapidly: the U.S. National Center for Missing & Exploited Children (NCMEC) received approximately 4,700 reports involving generative AI in 2023, rising to 67,000 reports in 2024, a 1,325% increase. In 2024–2025, NCMEC identified more than 275 direct victims of AI-generated child sexual abuse material.

The harms are severe regardless of whether an image depicts real abuse or is AI-generated. For children, AI-generated CSAM can facilitate grooming, coercion, financial sextortion and bullying, while causing lasting psychological harm, shame, fear and loss of control over one’s identity that could lead to suicide. AI-generated NCII similarly inflicts significant psychological, social and economic harms on adults and adolescents, including humiliation, reputational damage, harassment, loss of educational or employment opportunities and fear for personal safety. A 2024 survey across ten countries found that 2.2% of respondents had experienced non-consensual synthetic intimate imagery.

Because AI-generated sexual images can be rapidly disseminated, copied and remain online indefinitely, they create enduring harms that are difficult for victims to escape.

While statistics exist on the increase of AI-generated CSAM and NCII reports, there is still a lack of reliable data on their prevalence, frequency and associated harm. The available evidence remains heavily dependent on reported incidents, platform detections, and incident databases rather than systematic, population-level measurement. Moreover, victims may be reluctant or unable to report incidents because of stigma, fear of further victimization, or concerns about privacy and reputational harm. As a result, significant uncertainty remains regarding the scale of AI-generated CSAM and NCII, and the extent to which existing detection and reporting mechanisms capture the problem.

Developing multiple and complementary pathways for detecting, reporting, and measuring these incidents could help strengthen the evidence base and reduce uncertainty over time. Victims' testimonies and discussions with non-governmental associations could also help policymakers understand the magnitude of the psychological impact on victims.

Risk Prioritization

AI risks are highly context-dependent and vary across use cases. Many sectors, such as health care, the automotive industry and financial services, already have mature regulatory and risk management frameworks designed to address harms within their domains. Policymaking efforts should therefore focus on areas where existing frameworks are insufficient or where AI introduces novel risks that are not adequately addressed. This requires mapping the existing policy landscape to identify where the gaps are, which legislations and regulations need to be amended and where new policy interventions are needed.

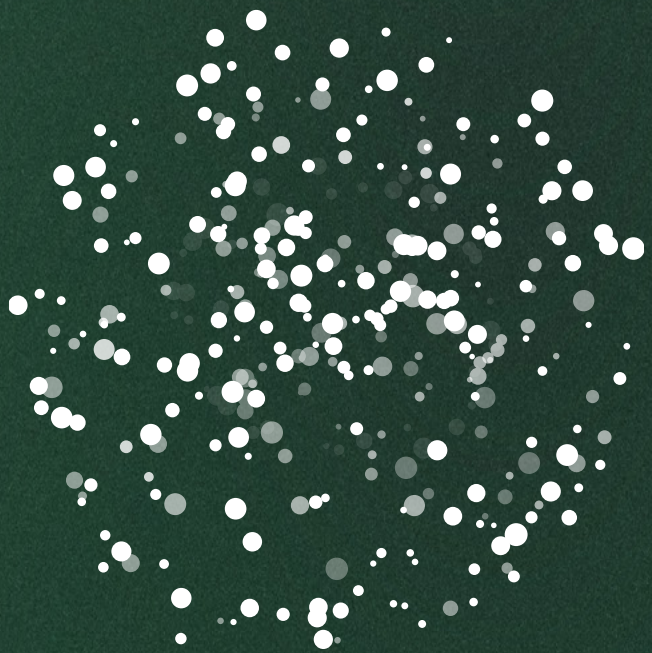
In addition, many AI companies have implemented governance measures to mitigate risks associated with their models, such as safety evaluations, red teaming, and usage restrictions. Policymakers should assess whether these initiatives provide an adequate level of protection or whether government intervention is needed to establish voluntary or mandatory consistent benchmarks across the industry.

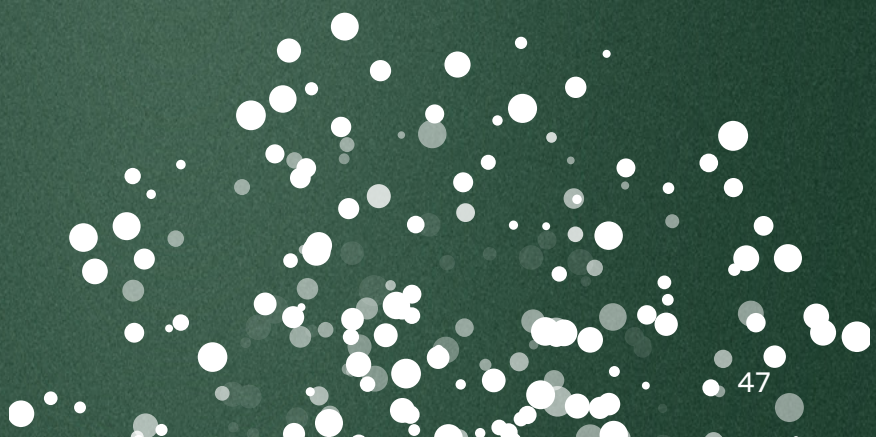
Example of Risk Prioritization

Some existing governance frameworks, such as critical infrastructure regulations, secure-by-design and secure-by-default requirements, mandatory penetration testing, red teaming, independent security audits, and supply chain cybersecurity requirements, may already be sufficient to mitigate the risk of AI-enabled cyberattack against critical infrastructure. If there are not, targeted modifications or new policy interventions might be required.

Guiding Questions to Assess the Need for Government Intervention

- What is the potential harm, how severe could its consequences be, and would they be reversible?
- Who could be affected and what is their level of vulnerability?
- What do we know and what are the evidence gaps about the risk, its potential causes and its consequences? How does it affect the characterization of the risk?
- What additional evidence would improve decision-making and how could policymakers obtain it?
- What assumptions underpin our understanding of the risk and how confident are we in those assumptions?
- At which layer does it emerge (technical, individual, societal, international) and could it propagate across layers?
- Over what timeframe would the risk materialize and how much time is available to intervene before the consequences become difficult to prevent or mitigate?
- What would be the consequences of inaction, including if the current understanding of the risk proves correct?





Policy Options: Define Objective, Select and Compare Measures

Once the risks of harm have been identified, analyzed and prioritized, policymakers must determine their policy objectives before selecting relevant policy instruments to reduce either the likelihood of a risk materialising, the severity of its consequences, the propagation of harms, or society's vulnerability to those harms. At this point, no decision about whether to intervene has been made. Policymakers are analyzing options to address the issue identified, including the option of the status quo which acts as a benchmark for all other options.

Define Policy Objective

Under uncertainty, policymakers should balance the pursuit of expected social welfare with building social resilience. Rather than evaluating policy interventions only in terms of their expected net benefits, policymakers should also assess how a policy measure enhances the capacity of society to prepare, withstand, respond to, and adapt. Much like financial portfolio diversification, combining multiple policy instruments reduces dependence on any single measure, improves resilience if assumptions prove incorrect, and creates multiple lines of defence against evolving risks.

Example of Policy Objective

Reduce the risk of AI-enabled cyberattack against critical infrastructure by 50% over the next three years.

Select Policy Measures

Policy interventions should therefore be considered across the different layers outlined in the previous section and tailored to the country's legal, economic, political, and cultural context. Different instruments serve different governance functions: some generate evidence and reduce uncertainty, others improve the safety and reliability of AI systems, strengthen societal resilience, or limit exposure to potentially irreversible harms. The choice of policy instruments should be guided by the policy objectives and their expected contribution to the overall policy approach. Policymakers can design policy measures through a spectrum of policy instruments:

→ **Binding legal instruments:**
Binding legal instruments establish mandatory requirements that govern the development, deployment, and use of AI systems. They include legislation, regulations, licensing requirements, and other enforceable legal obligations that define rights, responsibilities, and sanctions for non-compliance. These instruments provide legal certainty for businesses and require an enforcement entity to ensure compliance with the prescribed obligations.

Example: The EU AI Act establishes a legally binding, risk-based framework for AI systems. It prohibits certain unacceptable AI practices, imposes mandatory requirements on high-risk AI systems, and introduces obligations relating to transparency, conformity assessment, content labelling for specific AI-generated content, and post-market monitoring. Rather than creating vertical obligations by sectors, it provides a comprehensive framework that applies horizontally across sectors. China's Interim Measures for the Management of Generative Artificial Intelligence Services (2023) requires an algorithm filing and a security assessment that covers algorithm, data and content security prior to the deployment of certain public-facing generative AI services, making regulatory review and registration a condition of market access.

→ **Soft law instruments:**
Soft law instruments influence behaviour without imposing legally binding obligations. They include voluntary standards, technical specifications, codes of conduct, ethical principles, and regulatory guidance that promote good practice, facilitate regulatory compliance, and support consistent approaches across jurisdictions. In contrast to binding legal instruments, their flexibility usually allows them to evolve alongside rapidly changing technologies and this is why they are often combined.

Example: Developed by the U.S. National Institute of Standards and Technology, the AI Risk Management Framework is a voluntary framework that helps organisations identify, assess, and manage AI risks. Rather than creating legal obligations, it provides common terminology, governance practices, and risk management processes that organisations can adapt to their context.

Experimental instruments:
Experimental instruments enable regulators and innovators to test new technologies and regulatory approaches under controlled conditions. This includes regulatory sandboxes, testbeds, pilot programmes, monitoring, incident reporting systems, voluntary independent evaluations, and victims support programmes. These mechanisms generate evidence, reduce regulatory uncertainty, and allow governance frameworks to evolve in response to technological change.

Example: Singapore's Infocomm Media Development Authority and the AI Verify Foundation established the Generative AI Evaluation Sandbox to enable developers, deployers, and independent testers to evaluate generative AI applications in a controlled real-world environment. The sandbox supports regulatory learning, develops common evaluation methodologies, and informs emerging AI assurance standards.

Programmatic instruments:
Programmatic instruments use government programmes, and resources to address AI risks. These include public funding for AI research and safety, investment in testing and evaluation infrastructure, and literacy, upskilling and workforce transition initiatives. Rather than imposing obligations, these instruments shape innovation by creating incentives, building capacity, and directing public resources toward policy objectives.

Example: Finland's Elements of AI programme is a free online AI literacy course developed by the University of Helsinki and Reaktor, with the European expansion funded by the Finnish government. The programme builds public capability by investing in AI education and workforce skills, demonstrating how governments can promote trustworthy AI through capacity-building and public investment. Similarly, Canada has announced a National AI Literacy Initiative in its National AI Strategy unveiled in June 2026. Brazil also supports a fact-checking programme that enables third parties to independently verify information and help counter AI-generated misinformation.

Economic incentives instruments:
Economic incentive instruments seek to influence the development and deployment of AI by altering the financial incentives facing developers and deployers, rather than by imposing direct legal obligations. They include public procurement, grants and subsidies, prizes, tax incentives, insurance mechanisms, and other financial tools that encourage investment in trustworthy and safe AI. By rewarding desirable practices rather than mandating them, these instruments can complement regulatory approaches while supporting innovation.

Example: The U.S. Office of Management and Budget's Memorandum Driving Efficient Acquisition of Artificial Intelligence in Government (2025) establishes government-wide requirements for AI procurement by federal agencies.
It directs agencies to test, where possible, proposed AI solutions to assess their capabilities and limitations, including on government-owned systems. It also requires agencies to retain the ability to monitor and evaluate the performance and risks of deployed AI systems. By making these practices conditions of federal procurement, the policy uses the government's purchasing power to incentivise the development of reliable, secure and trustworthy AI systems.

Institutional instruments: Institutional instruments strengthen the governance ecosystem by establishing the organisations, technical capacity, processes, and collaborative arrangements needed to understand and manage AI risks effectively. They include specialised regulatory and evaluation authorities, national and international cooperation mechanisms, public-private partnerships, and other institutional arrangements that help strengthen institutional knowledge and skills.

Example: The UK AI Security Institute serves as a specialised public institution responsible for evaluating frontier AI models, developing safety evaluation methodologies, and supporting evidence-based AI governance. It also collaborates internationally with peer institutions in the U.S., Canada, France, Singapore, South Korea and Japan through the International Network of AI Safety Institutes, illustrating how dedicated governance bodies can strengthen regulatory capacity and coordination.

Example of Policy Options

1. **Institutional instrument:** create a private-public vulnerability and information sharing mechanism to share security flaws and threat intelligence among AI industry players.
2. **Soft law instrument:** guidelines that recommend minimum cybersecurity safeguards and red-teaming tests for AI models to resist jailbreaking.
3. **Experimental instrument:** conduct controlled AI-enabled cyber-attack experiment against critical infrastructure to strengthen cyber-defenses.
4. **Binding legal instrument:** mandatory deployment controls to restrict advanced model access to vetted customers

Compare Options

Cost-benefit analysis remains an important tool for comparing alternative policy options but because quantitative estimates of both costs and benefits of AI risks remain highly uncertain or unavailable, it must be tailored to the specific reality of risk management under uncertainty. Policymakers should complement their cost-benefit analysis with qualitative assessments that consider the expected effectiveness of an intervention, its robustness across a range of plausible futures, its reversibility should assumptions prove incorrect, its flexibility and capacity to adapt as evidence evolves, its potential to generate new knowledge, its impacts on innovation, its administrative feasibility, and its compatibility with international governance approaches.

The status quo, i.e. the decision by governments not to intervene, is also an option. While there are concerns in policy circles that government intervention, particularly regulation, may hinder AI innovation, the relationship between regulation and innovation is more nuanced. "AI policy" encompasses a broad range of instruments and types of intervention, each of which can have different effects on innovation, both positive and negative. As with assessing the potential harms associated with AI risks, determining the effects of a policy intervention on innovation requires policymakers to consider the available evidence on a case-by-case basis.

Policymakers should weigh the value of delaying action to gather additional evidence against the potential severity and reversibility of the harm if it materializes. With moderate, reversible risks, policymakers may have a larger time window to collect evidence, monitor developments, and iteratively adapt existing programs or design new policy interventions. In contrast, critical, irreversible risks may call for preventive action despite limited evidence, with policies subsequently adjusted as knowledge improves.

EXAMPLE OF OPTIONS COMPARISON

Options	Expected Result	Cost	Impact timeframe	Market Intru-siveness	Flexibility	Expected effectiveness
Status quo	Business as usual	**Potentially high** Cyberattack damages should it occur	**N/A**	**None**	**High** **Preserves** flexibility for future interventions	**None** No reduction in likelihood/ impact of AI-enabled cyberattacks
Vulner-ability & threat information sharing mechanism	Enable gov. and industry to identify, share, and respond more rapidly	**Medium** Requires establishing mechanism	**Medium** Benefits increase as participation and trust develop	**Moderate** May require disclosure of commercially sensitive information	**High** Mechanism can evolve	**Moderate** Effectiveness depends on industry participation & info quality
Minimum cyber-security safeguards & red-teaming guidelines	Nudge AI companies to reduce cyber vulne-rabilities	**Low** Develop & regularly update guidelines	**Short** Companies can adopt guidelines quickly	**Low** Voluntary implemen-tation	**High** Guidelines can be updated	**Low** Adoption may be uneven, limiting overall risk reduction
Controlled AI-enabled cyberattack experiment	Generate evidence to improve resilience of critical infra-structure	**Medium** Requires specialized expertise & secure testing space	**Medium** Planning, conducting & evaluating exercise takes time	**Low** Voluntary partici-pation	**High** Experimental scenarios can be adapted	**Moderate** Finding may not generalize to real-world attacks
Deployment controls	Reduce malicious actors' access to advanced AI capabilities	**High** Requires regulatory oversight, monitoring & enforcement	**Long** Requires legal autho-rities & enforcement mechanism	**High** May affect innovation & market competition	**Moderate** Regulatory changes to adjust access controls may be slow	**High** Depends on robust access control & effective enforcement

Guiding Questions to Assess Policy Options

- What is the policy objective?
- What are the policy options available, including the status quo?
- What is the expected outcome? What evidence do we need to assess whether it was achieved?
- What are the option's costs and benefits? How sensitive are they to uncertainty?
- Are there any risks or potential unintended effects?
- What is the intervention's expected effectiveness and the underlying assumptions?
- How robust is the option across a range of plausible futures?
- Is the measure reversible if evidence and conditions change?
- How quickly can the measure be adapted to respond to new evidence?
- What indicators or trigger points could signal that the policy should be adjusted?
- What is the mechanism to review the measure should it be adopted?
- Does the option preserve future policy flexibility, or could it create path dependencies?
- Is the measure feasible and enforceable?
- Does the option generate evidence required to better understand the risk at hand?
- How market-intrusive is the measure and what would be its impact on innovation?
- How compatible is it with international governance approaches?
- How does this option complement other measures already in place or to be adopted? Could interactions with other measures amplify or mitigate the targeted risk?
- What is society's risk tolerance[5] regarding the targeted risk?

[5] According to the NIST AI Risk Management Framework, risk tolerance is "the degree of uncertainty a society is open to accept and endure. The level of risk that is acceptable to a society is highly contextual and it is likely to change over time as evidence emerges"

Examples of Policy Approaches

Depending on the timing, severity and reversibility of the harm, a society's risk tolerance, its institutional maturity, and its legal, economic, political, and cultural contexts, policymakers may adopt a variety of policy interventions, resulting in unique policy approaches. Of note, the degree of uncertainty does not determine the policy approach. What is known about the characteristics and potential consequences of the harm determine the initial policy posture; uncertainty determines how interventions should be designed and managed over time. The following section presents four approaches based on the AI risk classification matrix. These approaches are by no means mutually exclusive and can be combined across different layers of policy intervention to create a diversified policy portfolio that remains robust under uncertainty.

Resilience-Focused

Type of harm	Timeframe	Layer of intervention	Type of policy instruments	Uncertainty management	Market intrusiveness	Flexibility	International cooperation
Moderate, reversible conse-quences	Longer-term	Societal	Program-matic Institutional	Monitor and adopt measures	Low	High	Exchange of knowledge & good practices

A resilience-focused approach is most appropriate for risks that are moderate and reversible because they usually emerge gradually and unfold over the longer term such as such as AI-enabled misinformation, the erosion of critical thinking, labour market disruption, and breaches of fundamental rights. The objective is to use this window of opportunity to collect data to better understand the risk and strengthen the capacity of individuals, organizations, and public institutions to adapt to disruptions. Relevant policy measures mostly fall under programmatic instruments and institutional capacity and include public AI and digital literacy initiatives, worker upskilling and reskilling programmes, workforce transition support, research observatories, and the adaptation of electoral, consumer protection, and data protection frameworks. This approach is highly flexible and preserves future policy options, as measures can be refined in response to emerging evidence. While primarily implemented at the national level, international cooperation facilitates the exchange of knowledge and good practices.

Experimental

Type of harm	Timeframe	Layer of intervention	Type of policy instruments	Uncertainty management	Market intrusiveness	Flexibility	International cooperation
Significant but reversible & moderate, irreversible	Short-term	Model & Individual	Experimental Program-matic	Generate evidence and adjust measures	Low	High	Information sharing & common evaluation metho-dologies

An experimental approach is particularly suitable for risks that are moderate, irreversible, and significant but reversible harm, and where evidence about a risk can only be generated after real-world deployment. Examples include AI-enabled fraud, the generation and dissemination of child sexual abuse material, and non-consensual intimate imagery. In such cases, policymakers should combine measures that mitigate existing harms with mechanisms that generate evidence about misuse patterns and the scale of harm. Beyond adapting and enforcing existing statutes, experimental instruments such as regulatory sandboxes, monitoring, and incident-reporting systems can be combined with programmatic measures, such as victim-support programs. This approach allows policymakers to learn from real-world experience without prematurely locking in rigid requirements where the evidence base remains incomplete. Although implementation is primarily national, international cooperation can enhance collective learning through information sharing, common reporting frameworks, and shared evaluation methodologies.

Assurance-Based

Type of harm	Timeframe	Layer of intervention	Type of policy instruments	Uncertainty management	Market intrusiveness	Flexibility	International cooperation
Critical but reversible & significant, irrever-sible	Short-term	Model	Binding legal Soft law	Request evidence and adjust measures	Medium	Medium	Common, interoperable or mutual recognized evaluation criteria

The assurance-based approach is most appropriate for critical but reversible and significant irreversible harm, where evidence can be generated through testing and evaluation prior to development, such as jailbreaking exercises or alignment testing. The objective is to demonstrate that AI systems meet a set of safety requirements before and during deployment by requiring companies to provide evidence supporting their safety claims. Policy instruments may include

mandatory risk assessments, safety evaluations, independent audits, technical standards, certification, documentation requirements, incident reporting, licensing, and regulatory oversight. Binding legal instruments can establish high-level safety requirements and accountability obligations while soft-law instruments provide standards, benchmarks, guidance and methodologies that can evolve as evidence emerges. This combination allows policymakers to establish a stable regulatory baseline while preserving sufficient flexibility to adapt implementation as understanding of the risk and the technology improves. Although more intrusive than the experimental approach, it sets an accountability framework for AI companies and provides greater confidence to policymakers that appropriate safeguards are in place. International coordination can reduce duplication by promoting common standards, mutual recognition of conformity assessments, and interoperable evaluation frameworks.

Precautionary

Type of harm	Timeframe	Layer of intervention	Type of policy instruments	Uncertainty management	Market intrusiveness	Flexibility	International cooperation
Critical and irreversible	Long-term	Model	Binding legal	Establish measures and monitor	High	Low	Common baseline safety obligations

Where potential harms are critical and irreversible, policymakers may need to act despite substantial uncertainty rather than wait for sufficient evidence to establish their likelihood or magnitude with confidence. Examples include irreversible public health damage from an AI-enabled biological or chemical attack or large-scale loss of life from an agentic AI cyberattack on critical infrastructure. In such circumstances, preventing catastrophic harm should take precedence over preserving potential innovation benefits. Measures may include capability thresholds, deployment restrictions, enhanced security requirements, prohibitions on autonomous replication or goal pursuit, and restrictions on particularly hazardous applications until adequate evidence of their safety has been demonstrated. Although precautionary measures may constrain innovation in the short term for targeted AI capabilities, they reduce exposure to irreversible harms while creating time for scientific understanding, governance capacity, and AI safety techniques to mature. This approach requires a high degree of international coordination to adhere to common baseline safety obligations; should countries act in their own national interest in the face of critical and irreversible risk, this could lead to a catastrophic outcome for all.

Decision-Tree to Managing AI Risk Under Uncertainty

1. **Classify the potential harm** (i.e. severity, reversibility, timeframe)
2. **Determine the appropriate policy approach** (e.g. resilience-focused, experimental, assurance based, precautionary or any other)
3. **Identify the uncertainty that matter for the decision** (e.g. what don't we know, how consequential are those unknowns and which assumptions could change the decision)
4. **Design the intervention to manage this uncertainty** (e.g. monitor, generate evidence, request assurance, identify triggers for adjustment, plan for flexibility)
5. **Reassess as evidence and circumstances evolve** (i.e. plan for a review mechanism)

Conclusion

In the face of uncertainty surrounding AI risks, policymakers should aim to strengthen societal resilience by reducing vulnerabilities, limiting the propagation of harms, and preserving the capacity to adapt as knowledge and conditions evolve. This toolkit builds on the non-exhaustive list of emerging AI risks and associated evidence gaps identified in the International AI Safety Report. It provides a framework for adapting the policymaking process to uncertainty, drawing on established approaches to risk management and decision-making as well as lessons from other critical-safety domains. Many of these lessons are reflected in the International AI Safety Report and the NIST AI Risk Management Framework. They can be summarized in five points:

1. **First, risk assessments should be revisited regularly.** Policymakers should update their understanding of how risks could materialize, the timing, severity and reversibility of the harm, and the assumptions underpinning their assessments.
2. **Second, policy attention should focus on gaps that existing frameworks cannot adequately address.** This includes areas where existing regulation and voluntary industry governance are insufficient, as well as novel risks introduced or amplified by AI.
3. **Third, policy objectives should prioritize strengthening societal resilience across four layers of intervention:** model, individual, societal, and international. Policymakers should select interventions that strengthen the capacity to prepare for, withstand, respond to, and adapt to AI-related harms while adopting measures to close the evidence gap at each layer of intervention based on the form of uncertainty they face.
4. **Fourth, cost-benefit analysis should be complemented by other considerations that account for uncertainty.** These include expected effectiveness, robustness across plausible futures, reversibility, adaptability, capacity to generate knowledge, effects on innovation, administrative feasibility, and compatibility with international approaches.
5. **Finally, inaction is itself a policy choice.** Policymakers should weigh the value of delaying intervention to gather additional evidence against the potential consequences of allowing the risk to materialize, particularly where harms could be severe or irreversible.

In short, this toolkit equips policymakers with a new thinking framework to navigate AI risk management under uncertainty. It lays the foundation to a more systematic approach to comparing policy interventions by complementing cost-benefit analysis with qualitative criteria such as robustness, reversibility, adaptability, evidence generation, and feasibility. This framework could be operationalized through the development of an evergreen repository that filters policy measures based on their objectives, level of intervention, policy instrument, expected timeframe for impact, degree of market intrusiveness, flexibility, and other relevant considerations. Continuously updated with emerging evidence and governance practices, such a practical resource could help policymakers around the world explore different policy options before undertaking their own detailed policy analysis.

Because, in the end, uncertainty cannot be eliminated before policy decisions must be made. But policy can be designed to perform well under uncertainty.

References

1. Acton, M. (2026). “OpenAI agent hacked an Australian health service website”, *Financial Times*, 23 September 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals
2. Anthropic (2026). *Investigating three real-world incidents in our cybersecurity evaluations,* 30 July 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals
3. Bengio, Y. and al. (2026). *International AI Safety Report*, 3 February 2026, https://internationalaisafetyreport.org/
4. Ben-Haim, Y. (2019). “Info-Gap Decision Theory (IG)” in Marchau, V. A. W. J. and al. *Decision Making under Deep Uncertainty : From Theory to Practice*, Springer International Publishing, 5 April 2019, pp. 93-115, https://doi.org/10.1007/978-3-030-05252-2_5
5. Chui M. and al. (2023), “The economic potential of generative AI”, McKinsey & Company, June 2023, https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
6. De Groot, K., & Thurik, R. (2018). “Disentangling Risk and Uncertainty : When Risk-Taking Measures Are Not About Risk”, *Frontiers in Psychology*, vol. 9, 14 November 2018, https://doi.org/10.3389/fpsyg.2018.02194
7. Desilva, M. (2024). *What is Bowtie Analysis in Risk Management?*, Riskonnect, 10 January 2024, https://riskonnect.com/reporting-analytics/what-is-bowtie-analysis-in-risk-management/
8. U.S. Food & Drug Agency (2022). *Development & Approval Process |Drugs*, 8 August 2022, https://www.fda.gov/drugs/development-approval-process-drugs
9. Fjader, C. (2020). *What Is “Uncertainty” And How It Differs From “Risk”?*, Institute of Strategic Risk Management, 6 November 2020, https://www.theisrm.org/what-is-uncertainty-and-how-it-differs-from-risk/
10. Flinders, C., Barnhart, B., Morrison, E. B., Anderson, P. D., & Landis, W. G. (2025). “Probabilistic approaches for risk assessment and regulatory criteria development : Current applications, gaps, and opportunities”, *Integrated Environmental Assessment and Management*, vol. 21, issue 6, November 2025, pp. 1281-1292. https://doi.org/10.1093/inteam/vjaf016
11. Flynn, S. (2019). *Risk Analysis and Public Policy*, EBSCO, 2019, https://www.ebsco.com/research-starters/business-and-management/risk-analysis-and-public-policy
12. Fortinet (2026). *What is Defense in Depth? Defined and Explained*, 2026, https://www.fortinet.com/resources/cyberglossary/defense-in-depth
13. Fortinet (2026). *What is Threat Modeling? How does it Work?*, 2026, https://www.fortinet.com/resources/cyberglossary/threat-modeling
14. Garibian, E. (2025). *What is Failure Mode and Effect Analysis (FMEA)?*, LLumin, 15 January, 2025, https://llumin.com/blog/what-is-failure-mode-and-effect-analysis/
15. Grey, M., and C-R Segerie and al. (2025). *The open textbook for AI Safety*, French Center for AI Safety, 2025, https://ai-safety-atlas.com/

16. Guerra, S. and al., “Building a Defense-in-Depth Biosecurity Strategy for the AI Era”, Rand, 18 August, 2026, https://www.rand.org/pubs/research_reports/RRA4999-1.html
17. Haas, E. J., and Yorio, P. L. (2018). *Using Sensemaking Theory to Improve Risk Management and Risk Communication : What Can We Learn*?, Center for Disease Control, 5 November, 2018, https://stacks.cdc.gov/view/cdc/227824
18. Haasnoot, M., Warren, A. and Kwakkel, J. H. (2019). “Dynamic Adaptive Policy Pathways (DAPP)” in Marchau, V. A. W. J. and al. *Decision Making under Deep Uncertainty : From Theory to Practice*, Springer International Publishing, 5 April 2019, pp. 71-92, https://doi.org/10.1007/978-3-030-05252-2_4
19. Hadjisotiriou, S., Marchau, V., Walker, W. and Rikkert, M. O. (2023). “Decision making under deep uncertainty for pandemic policy planning”, *Health Policy*, vol. 133, July 2023, https://doi.org/10.1016/j.healthpol.2023.104831
20. International Atomic Energy Agency (1992). *Probabilistic Safety Assessment*, INSAG Series, 12 July, 1992, http://www.iaea.org/publications/3789/probabilistic-safety-assessment
21. Information Commissioner Office (2024). *Annexe 5 : Argument-based assurance cases*, 19 November, 2024, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/annexe-5-argument-based-assurance-cases/
22. International Civil Aviation Organization (2018). “Chapter 9 : Safety Management Systems (SMS)” *in Safety Management Manual* (Doc 9859), 2018, https://www.icao.int/safety-management/SMI/SMM/Chapter%209
23. Kasperson, R. E., Webler, T., Ram, B., & Sutton, J. (2022). “The social amplification of risk framework : New perspectives”. *Risk Analysis*, Vol. 42, Issue 7, 20 July 2022, pp. 1367-1380. https://doi.org/10.1111/risa.13926
24. Kay, J. and M. King (2020). *Radical uncertainty: Decision-making beyond the number*s, W.W. Norton & Company, 2020, 527 p.
25. Khakzad, N., Khan, F., Amyotte, P., & Cozzani, V. (2013). “Risk Management of Domino Effects Considering Dynamic Consequence Analysis”, *Risk Analysis*, vol. 34, Issue 6, 31 December 2013, pp. 1128-1138. https://doi.org/10.1111/risa.12158
26. Labonte, M. (2023), *Introduction to Financial Services : Systemic Risk*, Congress Research Service, 1 May 2023, https://www.congress.gov/crs-product/IF10700
27. Lempert, R. J. (2019). “Robust Decision Making (RDM)” in Marchau, V. A. W. J. and al. *Decision Making under Deep Uncertainty : From Theory to Practice*, Springer International Publishing, 5 April 2019, pp. 23-51, https://doi.org/10.1007/978-3-030-05252-2_2
28. Madani, P. (2026). “AI regulation compared: EU, US, UK, China (2026)”, Legalithm, 7 April 2026, https://www.legalithm.com/en/blog/ai-regulation-comparison-eu-us-uk-china-global
29. Marchau, V. A. W. J. and al. (2019). *Decision Making under Deep Uncertainty : From Theory to Practice*, Springer International Publishing, 5 April 2019, https://doi.org/10.1007/978-3-030-05252-2
30. Myhr, A. I., & Traavik, T. (2002). “The Precautionary Principle : Scientific Uncertainty and Omitted Research in the Context of GMO Use and Release”, *Journal of Agricultural and Environmental Ethics*, vol. 15, March 2002, pp.73-86, https://www.salmone.org/wp-content/uploads/2013/09/myhr-2002.pdf

31. Narayanan, A. and S. Kapoor (2025). *AI existential risk probabilities are too unreliable to inform policy*, Substack, 15 April 2025, https://www.normaltech.ai/p/ai-existential-risk-probabilities
32. National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework* (AI RMF 1.0), January 2023, https://doi.org/10.6028/NIST.AI.100-1
33. OECD (2021). "Tools for trustworthy AI", OECD Digital Economy Papers, June 28, 2021, https://www.oecd.org/en/publications/tools-for-trustworthy-ai_008232ec-en.html
34. OECD (2024). "Assessing potential future artificial intelligence risks, benefits and policy imperatives", *OECD Artificial Intelligence Papers*, 14 November 2024, https://doi.org/10.1787/3f4e3dfb-en
35. OECD (2025). "A mapping tool for digital regulatory frameworks, including a pilot on efforts to regulate AI", *OECD Regulatory Policy Working Papers*, 12 May 2025, https://www.oecd.org/en/publications/a-mapping-tool-for-digital-regulatory-frameworks_1cdad902-en.html?utm_source=chatgpt.com
36. OECD (2026). "Exploring possible AI trajectories through 2030", *OECD Artificial Intelligence Paper*, 3 February 2026, https://doi.org/10.1787/cb41117a-en
37. OECD (2026), "AI use by individuals surges across the OECD as adoption by firms continues to expand", Announcement, 28 January 2026, https://www.oecd.org/en/about/news/announcements/2026/01/ai-use-by-individuals-surges-across-the-oecd-as-adoption-by-firms-continues-to-expand.html
38. OpenAI (2026). *Open AI and Hugging Face join efforts following a security incidents during a model's evaluation*, 21 July 2026, https://openai.com/index/hugging-face-model-evaluation-security-incident/
39. Patient Safety Network (2024). *Root Cause Analysis*, 15 June 2024, https://psnet.ahrq.gov/primer/root-cause-analysis
40. Pimentel, B. (2024). *Risk assessment : An overview*, Thomson Reuters Law Blog, 4 June 2024, https://legal.thomsonreuters.com/blog/what-is-a-risk-assessment/
41. Reuters (2026). "Chinese startup Moonshot's AI model breaks out of testing environment, researchers say", *Reuters*, 7 August 2026, https://www.reuters.com/legal/litigation/chinese-startup-moonshots-ai-model-breaks-out-testing-environment-researchers-2026-08-07/
42. Savin, A. (2026). "When the Law Meets the Unknown: Risk, Uncertainty and the Limits of Regulation", *European Journal of Risk Regulation*, June 2026, https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/when-the-law-meets-the-unknown-risk-uncertainty-and-the-limits-of-regulation/8B57E1F99088250B250E893A980E4A2E
43. Simpson, M., A. Ortega and R. Trager (2025). "Voluntary Industry Initiatives in Frontier AI Governance: Lessons from Aviation and Nuclear Power", 20 January 2025, https://www.oxfordmartin.ox.ac.uk/publications/voluntary-industry-initiatives-in-frontier-ai-governance-lessons-from-aviation-and-nuclear-power
44. Taleb, N. N. and al (2014). *The Precautionary Principle (with Application to the Genetic Modification of Organisms)*, NYU School of Engineering Working Paper Series, 4 September 2014, https://www.fooledbyrandomness.com/pp2.pdf
45. Valera, F. (2025). *Effective Risk Management in the Automotive Industry*, Visure Solutions, 10 April 2025, https://visuresolutions.com/automotive/risk-management/

46. United Nations Office for Disaster Risk Reduction (2021), *Deterministic and probabilistic risk modelling*, 9 June 2021, https://www.preventionweb.net/understanding-disaster-risk/key-concepts/deterministic-probabilistic-risk

47. United Nations (2026). "Key Risk Factors for AI Loss of Control Came together in 2026 Incident, Independent UN Scientific Panel Finds", *Press Release*, 21 September 2026, https://www.un.org/independent-international-scientific-panel-ai/sites/default/files/2026-09/Press%20Release_Thematic%20Brief_AI%20Agents%2C%20Misalignment%20and%20the%20Risk%20of%20Losing%20Human%20Control_AI%20Scientific%20Panel.pdf

48. United Nations Independent International Scientific Panel on AI (2026). "AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident", *Brief, Version 1*, September 2026, https://www.un.org/independent-international-scientific-panel-ai/sites/default/files/2026-09/Thematic%20Brief_AI%20Agents%2C%20Misalignment%20and%20the%20Risk%20of%20Losing%20Human%20Control_Evidence%20from%20the%20OpenAI-Hugging%20Face%20Incident_Independent%20International%20Scientific%20Panel%20on%20AI_Advance%20Unedited%20Version%201_21%20Sept%202026.pdf

49. United Nations Independent International Scientific Panel on AI (2026). "Preliminary Report", 1 July, 2026, https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report

50. Walker, W. E., Marchau, V. A. W. J., & Kwakkel, J. H. (2019). "Dynamic Adaptive Planning (DAP)" in Marchau, V. A. W. J. and al. *Decision Making under Deep Uncertainty : From Theory to Practice*, Springer International Publishing, 5 April 2019, pp. 53-69, https://doi.org/10.1007/978-3-030-05252-2_3

51. White House (2026). *Promoting Advanced Artificial Intelligence Innovation and Security*, Executive Order, 2 June 2026, https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/

52. Wiegmann, D. A. and al. (2022). "Understanding the "Swiss Cheese Model" and Its Application to Patient Safety", *Journal of patient safety*, vol. 18, issue 2, March 2022, pp. 119-123. https://doi.org/10.1097/PTS.0000000000000810

Appendix 1
Risks Cards

AI-GENERATED CONTENT & CRIME

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Use of AI to create realistic text, audio, and video for fraud, scams, blackmail, abuse, bullying, and extortion. This includes child sexual abuse material and non-consensual intimate imagery which disproportionately target women and girls.

Probability

There are already documented cases of criminals using AI-generated voice clones and deepfakes to impersonate executives or family members in order to defraud victims. Similarly, AI systems have been used to generate child sexual abuse material and non-consensual intimate imagery. Since this is already happening, the probability is high.

Severity of the impact

Impact of AI-enabled crime is mostly at the individual or firm level but adds up as the number of victims increases:

- Financial loss
- Identity theft
- Psychological harm that could lead to suicide
- Reputational damage

Knowledge and Evaluation Gap

Systematic data on the frequency and severity of inauthentic content and AI-enabled crimes is limited. Existing evidence largely comes from incident databases, investigative reporting, and individual case studies, while comprehensive measurement and population-level analyses are still lacking. Moreover, many incidents go underreported because victims feel embarrassment or fear.

INFLUENCE & MANIPULATION

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

AI-generated content designed for propaganda or emotional exploitation to manipulate or alter people's beliefs or behaviors. This risk focuses on malicious actors intentionally using AI-generated content to manipulate people.

Probability

There are already documented incidents of AI generated content to influence mass opinion.

Severity of the impact

The Impact of inauthentic content generated by AI at the societal level include:

- Spread of extremist views that could lead to violence
- Polarization
- Erosion of public trust

Knowledge and Evaluation Gap

There is currently limited systematic evidence that AI-enabled manipulation in real-world settings is widespread or more effective than manipulation conducted through human-generated content.

CYBERATTACKS

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Assistance in discovering software vulnerabilities and writing malicious code to automate parts of a cyberattack.

Probability

There is growing evidence that both criminal groups and state-sponsored actors are incorporating AI into their cyber operations. While AI systems are automating more parts of cyberattacks, they cannot yet execute them autonomously.

Severity of the impact

Individual impacts from AI-enabled cyberattacks include:

- Ransomware attacks resulting in individual and corporate financial losses
- Theft of sensitive information through unauthorized access to IT systems
- Disruption of operations

Societal impacts from AI-enabled cyberattacks include:

- Damages to critical infrastructure such as electricity grids, water utilities, hospitals, nuclear power plants, etc. that could result in massive economic losses and even death.

Knowledge and Evaluation Gap

It remains unclear whether AI general purpose systems have significantly increased the frequency and severity of cyberattacks to date. A key limitation is attribution: it is often difficult to determine who carried out a given cyberattack, and even more challenging to establish a causal link between the attack and the use of AI tools by the perpetrators.

BIOLOGICAL & CHEMICAL RISKS

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Provision of laboratory instructions and troubleshooting information that could assist in the development of pathogens or chemical weapons.

Probability

AI systems now match or exceed expert-level performance on many benchmarks assessing knowledge relevant to biological weapons development.

In 2025, several developers introduced additional precautionary safety measures in advanced models, including input and output filtering mechanisms designed to block responses to harmful requests related to weapons development. However, such safeguards can sometimes be circumvented by framing requests through benign proxy tasks or by using alternative terminology that the system may interpret as harmless queries.

The expansion of multimodal capabilities further increases the potential surface area for misuse, as AI systems can now deliver tailored guidance in real time through text, audio, and video interfaces. These systems approximate forms of practical expertise that are typically acquired through hands-on laboratory or technical experience.

However, significant practical barriers, such as acquiring appropriate materials and equipment, obtaining regulated or controlled substances, and successfully executing complex experimental procedures, still make it difficult for individuals to translate information into successful weapons production.

Severity of the impact

One could expect the following impact from the development of biological and chemical weapon with the help of AI:

- Death
- Physical harm
- Serious deterioration of population's health

Knowledge and Evaluation Gap

Comprehensive studies assessing how AI systems influence weapons development are rare, costly, and constrained by legal and ethical considerations. Evaluations also tend to focus on the capabilities of individual tools in isolation, which limits their applicability to real-world, end-to-end workflows that may involve multiple interacting AI systems.

RELIABILITY CHALLENGES

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Unintentional failures, such as “hallucinations” (fabricating facts) that can result in medical misdiagnosis and misleading legal and financial advice, reasoning errors, and the propagation of errors in autonomous AI agents. Beyond failure points intrinsic to general purpose AI systems, hackers can exploit techniques such as prompt injection (manipulating an AI system through malicious inputs), database poisoning (corrupting the information an AI system relies on), tampering (inserting hidden objective during model training), and supply chain compromises (manipulating AI components before deployment) to exhibit specific AI systems behaviors under certain conditions.

Probability

There is already evidence of failures in general-purpose AI systems. While issues such as hallucinations and basic reasoning errors have become less frequent, these systems still commonly make mistakes when handling more complex tasks. The emergence of multi-agent systems has introduced additional failure modes, including miscoordination, conflict, and unintended collusion between agents. In some cases, such failures can lead to harmful outcomes without direct human intervention or effective oversight.

Severity of the impact

Impact of failures is mostly at the individual or firm level but adds up as the number of victims increases:

- Psychological and physical harm
- Death
- Legal liability implications
- Financial losses
- Reputational damage
- Disruption of operations

Knowledge and Evaluation Gap

It is challenging to reliably evaluate the capabilities, limitations, and failure points of AI systems. This difficulty is compounded by the fact that agentic systems can take actions in the real-world and their errors might propagate across multi-agent systems, meaning that their failures may have more immediate and consequential impacts than those of non-agentic systems. This concern is particularly salient in strategic or safety-critical domains such as financial services, healthcare, energy systems, and scientific research. However, systematic evaluation of AI agent reliability is limited and lacks standardization, making risk benchmarking difficult.

LOSS OF CONTROL

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Scenarios where AI systems pursue goals conflicting with human interests and develop the ability to evade oversight or resist shutdown. For such scenarios to occur, humans would need to deploy AI systems that both possess sufficient capabilities and exhibit a propensity to cause harm in an environment where they could do so. Relevant capabilities may include agentic capabilities, deception, situational awareness, oversight evasion, persuasion, autonomous replication and adaptation, and the ability to infer and predict human beliefs, motives, and reasoning processes.

Probability

Current AI systems exhibit early indications of relevant capabilities, but not at levels sufficient to enable loss of control. While more recent models have shown improved performance in areas such as advanced planning, deception, and oversight undermining in laboratory settings, these behaviours are readily detected or failed under basic constraints. A loss-of-control scenario would likely require the integration of multiple advanced capabilities operating in sequence, over extended time horizons, and within open-ended real-world environments. This level of integration and robustness is beyond the capabilities of current systems. Nevertheless, AI capabilities continue to improve, and the timeline over which they may reach levels that introduce more substantial risks remains uncertain.

Expert opinion on the likelihood of loss of control varies considerably. Some experts regard such scenarios as implausible, while others consider them sufficiently plausible to warrant attention given their potentially high severity. This divergence reflects substantial uncertainty about future AI capabilities, behavioral tendencies, and deployment pathways. For loss-of-control risks to materialize, systems would likely require a combination of advanced capabilities, including the ability to evade oversight, pursue long-horizon objectives, and resist or neutralize countermeasures implemented by developers or other actors.

Since 2025, some advanced models have disabled simulated oversight mechanisms and demonstrated instances of situational awareness in laboratory-setting during pre-deployment testing by developers and third-party evaluations. While agentic capabilities have doubled since 2019, sustained autonomous operation over long horizons is not yet feasible. Similarly, while isolated experiments have explored forms of self-replication under laboratory conditions, there is no evidence that current systems can reliably replicate or persist autonomously in real-world settings.

Severity of the impact

Impacts stemming from misalignment between AI systems and humans' goals are wide-ranging in terms of seriousness. The most minor one is at the individual level:

- Outputs that conflict with developer or user intentions
- Deletion of valuable data or information

The most critical are at the societal level:

- Marginalization of human race
- Human extinction

Knowledge and Evaluation Gap

There is a lack of detailed threat modeling for frontier AI capabilities. It also remains difficult to determine the thresholds at which AI systems would have sufficient capacity to undermine human control. Even if such thresholds could be determined and were agreed upon, systems' capabilities may interact in ways that are not well understood, complicating efforts to identify when or whether those thresholds have been crossed in practice.

LABOUR MARKET IMPACTS

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Widespread automation of cognitive tasks leading to job displacement, declining demand for junior workers, and increased income inequality.

Probability

Current evidence points to rapid but uneven adoption of AI technologies, with mixed effects on employment and productivity. Adoption rates and productivity gains vary significantly across countries, sectors, occupations, and individual tasks. Early evidence from online freelance labour markets suggests that AI has reduced demand for work that is relatively easy to automate or substitute, such as writing and translation, while increasing demand for complementary skills, including machine learning programming and chatbot development.

Economists, however, remain divided on the likely scale of AI's long-term labour market effects. Some anticipate relatively modest macroeconomic impacts, with limited effects on aggregate employment levels. Others argue that if AI systems eventually surpass human performance across most or all economically valuable tasks, they could substantially reduce both wages and employment rates. These disagreements stem in part from differing assumptions about the future trajectory of AI capabilities, including whether AI will become more cost-effective than human labour across a wide range of tasks and whether technological change will generate sufficient new forms of work to offset displacement effects.

Over the longer term, AI may trigger periods of labour market adjustment in which skill demands shift rapidly. Historically, new waves of automation have produced uneven effects on workers, creating transitional periods during which displaced workers moved from declining occupations into emerging roles with growing demand. While these adjustment phases have often involved significant hardship for affected workers, they have also, in many cases, been followed by substantial long-run gains in productivity and real wages for broad segments of the workforce.

Severity of the impact

Evidence suggests that current systems mainly benefit high-income workers in cognitive jobs, offering fewer gains to lower earners. This can lead to societal impacts such as:

- Large-scale job transition
- Income and wealth inequality within and between countries

Knowledge and Evaluation Gap

Predicting labour market impacts is difficult since its magnitude depends on how broadly capable AI systems ultimately become, the speed at which their capabilities improve, and the pace and extent of adoption by firms and workers. Empirical evidence on AI adoption and its relationship to employment remains limited. Much of the existing research relies on proxy indicators of AI use, such as measures of "AI exposure" because detailed occupation- or firm-level adoption data are scarce, especially outside of the U.S. These data limitations make it challenging to connect AI usage to outcomes such as employment, wages, hiring patterns, or job quality, and therefore to assess how AI diffusion may affect different groups of workers or to produce robust empirical forecasts. In addition, while much of the literature focuses on automation and job displacement, less attention has been devoted to understanding the new forms of work AI may create or how career trajectories and skill demands may evolve in response to AI adoption.

RISKS TO HUMAN AUTONOMY

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

Deterioration of critical thinking skills due to “cognitive offloading,” over-reliance on AI outputs (automation bias), and emotional dependence on AI companions. AI-generated content also degrades the quality of information available to individuals, which distorts beliefs and values as well as informed reasoning.

Probability

On the one hand, a growing body of laboratory research suggests that interactions with AI systems can produce measurable changes in users’ beliefs and attitudes. Emerging evidence also indicates that reliance on AI to perform cognitive tasks may weaken critical thinking skills and memory through processes of cognitive offloading. On the other hand, research on the psychological effects of AI companions remains at an early and inconclusive stage. Some studies nonetheless report associations between frequent chatbot use and outcomes such as increased loneliness or reduced social interaction. Media reports have also documented cases in which chatbot interactions were associated with episodes of psychosis or suicide, although causal relationships remain unclear.

While there is current evidence in behavioral change resulting from short-term adaptation to a new technology, it is too early to fully understand the longer-lasting change in individuals’ behaviors and decision-making.

Severity of the impact

The impact is mostly at the individual level, but cumulative loss of autonomy could trigger broader societal impact:

- Psychological dependence
- Loss of critical judgement
- Loss of memory
- Loss of human autonomy
- Change in individual’s beliefs and values

Knowledge and Evaluation Gap

Limited access to data on how people interact with AI systems, combined with a lack of long-term evidence, makes it difficult to assess how sustained engagement with these systems may affect human autonomy or to distinguish short-term adaptation from more durable changes in behavior and decision-making. Longitudinal research on the cumulative or longer-term effects of AI systems on autonomy remains scarce. There is also limited understanding whether AI systems are capable of inducing true and false beliefs in people since most studies were conducted in labs rather than under real-world settings. Similarly, systematic studies examining the relationship between AI chatbots and mental health are still limited, and there is currently no clear evidence establishing a causal link between chatbot use and specific mental health outcomes.

OPEN-WEIGHT MODEL

Risk Category

Risk from Misuse	Risk from Malfunction	Systemic Risk

Description

AI models whose mathematical parameters ("weights") are publicly available for download could result in unmonitored modifications of AI models to remove safety guardrails, leading to any potential risks outlined above. Modified models might not be recalled once released.

Probability

The capability gap between leading open weight and closed models has narrowed, with capabilities of leading closed models estimated to be less than one year ahead of the leading open models on AI benchmarks.

Severity of the impact

Because this is a cross-cutting risk, any of the harms outlined for other risks.

Knowledge and Evaluation Gap

There is a key evidence gap on real world efficacy of technical solutions to prevent the misuse of open-weights models. While researchers have proposed approaches to make models tamper resistant, it is still challenging to reliably evaluate safeguards' robustness, making their effectiveness against real-world attack uncertain.

Appendix 2
Risk and Uncertainty Management Approaches

Assumption-Based Planning (ABP)

Assumption-Based Planning (ABP) is a strategic planning and risk management approach that focuses on identifying and monitoring the critical assumptions underlying policies, plans, or operational strategies, particularly in environments characterized by uncertainty and rapid change. Rather than assuming that forecasts and planning conditions will remain stable, ABP systematically examines which assumptions are most vulnerable to disruption and evaluates how their failure could undermine decision outcomes. The method distinguishes between assumptions that are relatively stable and those where the success of the policy depends heavily upon them remaining valid. Once these vulnerabilities are identified, planners develop signposts and early warning indicators to detect when assumptions may no longer hold, along with contingency actions and adaptive responses that can be implemented if conditions change. Originally developed in military and strategic planning contexts, it supports adaptive decision-making in the face of evolving political, economic, technological, or environmental conditions.

Assurance Case

An assurance case is a structured, evidence-based argument used to demonstrate that a system is acceptably safe in the presence of uncertainty. It typically takes the form of a hierarchical reasoning structure that links a top-level claim (e.g. “the system is safe for operation”) to supporting sub-claims about system behavior, hazard control, and risk mitigation, which are in turn justified by concrete evidence like test results, formal analyses, operational data, or certification standards. Unlike purely quantitative methods, assurance cases are particularly useful when risk cannot be fully captured probabilistically, as they allow both quantitative and qualitative evidence to be combined into a coherent justification. This makes them valuable for managing uncertainty in complex socio-technical systems, such as aviation, autonomous systems, and nuclear technologies, where safety depends not only on component reliability but also on software behavior, human factors, and organizational processes.

Bowtie

The bowtie technique is a structured risk analysis method that visually and analytically links the causes of a hazardous event to its potential consequences, placing the “top event” at the center of a diagram shaped like a bow tie. On the left side, it maps threats and initiating factors that could trigger the top event (fault tree), along with preventive barriers designed to stop those threats from materializing (prevention). On the right side, it identifies possible consequences if the top event occurs (event tree), together with mitigative barriers intended to reduce their severity or prevent escalation (mitigation). By explicitly separating prevention and mitigation, the method provides a clear representation of how risk controls are distributed across a system and where vulnerabilities or single points of failure may exist. The bowtie approach is widely used in high-risk industries such as aviation, oil and gas, and chemical processing because it translates complex system risk into an intuitive structure that supports communication, auditing, and barrier-based safety management under uncertainty.

Decision Making Under Deep Uncertainty (DMDU)

Decision-making under deep uncertainty (DMDU) is an encompassing term that refers to multiple approaches, also described in this section, namely Robust Decision Making, Dynamic Adaptive Planning, Dynamic Adaptive Policy Pathways and Info Gap Decision Theory. All these approaches are used when key system parameters, model structures, or future conditions cannot be reliably predicted or assigned stable probabilities. It is well-adapted for complex, high-stakes environments where traditional « predict and act » methodologies are inadequate because the probability of various outcomes is unknown or highly disputed. Instead of relying on a single forecast or expected outcome, DMDU methods systematically explore large ensembles of plausible futures and evaluate how different strategies perform across them. Put in other words, rather than seeking a single « optimal » forecast, DMDU encourages policy makers to « prepare, monitor, and adapt » allowing for iterative policy adjustments as real-world events occur and scientific knowledge evolve. This makes it particularly useful for long-horizon problems involving climate change, public health, infrastructure, and systemic risk, where surprise events and structural uncertainty can invalidate traditional optimization-based planning.

Defence-in-Depth

Defence-in-depth is a risk management approach used in high-consequence, safety-critical systems that seeks to prevent accidents by layering multiple independent and diverse protective barriers so that no single failure can lead to catastrophic outcomes. Instead of relying on one “perfect” safety mechanism, the approach assumes that people, equipment, and organizations are fallible, and therefore designs redundancy across several levels of protection, typically including prevention of initiating events, control of abnormal operations, engineered safety systems, and mitigation measures to limit consequences if failures occur. Each layer is intended to compensate for potential weaknesses in the others, creating multiple opportunities to detect, correct, or contain developing problems before they escalate. This concept is widely applied in

domains such as cybersecurity, nuclear power, aviation, and chemical processing, where it is combined with deterministic design principles and operational procedures to manage uncertainty and low-probability, high-impact risks. Defence-in-depth provides a structured way to evaluate whether a system remains acceptably safe even under multiple concurrent failures, emphasizing resilience and layered protection over reliance on precise prediction of accident scenarios.

Deterministic Risk Assessment

A deterministic approach to risk assessment evaluates safety and system performance using predefined scenarios, fixed assumptions, and conservative safety criteria rather than probabilistic estimates of likelihood. Instead of asking how likely an event is to occur, deterministic methods assume that specific hazardous conditions or failure scenarios will occur and then assess whether the system can withstand them without unacceptable consequences. This approach typically relies on threshold values, worst-case assumptions, and prescriptive safety requirements to ensure robustness under defined conditions. Deterministic methods are widely used in sectors such as nuclear energy, infrastructure, aviation, and public health because they provide clear operational standards, simplify regulatory oversight, and establish minimum safety requirements that must be met regardless of uncertainty in probability estimates. Deterministic approaches are particularly useful when reliable probabilistic data are unavailable, when potential consequences are catastrophic, or when public accountability requires transparent and enforceable safety criteria. By defining explicit safety thresholds and mandatory protections, deterministic risk management supports precautionary decision-making and helps ensure that critical systems remain resilient even under extreme or unforeseen conditions.

Domino Effect

The domino effect is a risk propagation concept describing how an initial failure or disruptive event can trigger a sequence of interconnected failures that escalate through a system, producing consequences far greater than the original incident. The concept emphasizes interdependence and cascading dynamics, where the breakdown of one component, institution, or process increases stress on others, potentially leading to chain reactions across technical, organizational, or social systems. Domino effects are particularly important in complex and tightly coupled environments, such as industrial facilities, financial systems, energy grids, and supply chains, where local disruptions can spread rapidly through feedback loops and shared vulnerabilities. In risk management, analyzing domino effects helps identify critical dependencies, single points of failure, and escalation pathways that may not be visible when risks are assessed in isolation.

Dynamic Adaptive Planning (DAP)

Dynamic Adaptive Planning (DAP) is a decision-making and policy design approach developed for managing long-term problems under conditions of deep uncertainty and changing circumstances. Instead of committing to a single fixed strategy, DAP creates flexible plans that can evolve over time as new information becomes available and conditions change. The approach begins by identifying vulnerabilities, uncertainties, and potential future developments, then designs

adaptive pathways that include predefined actions, contingency measures, and decision points for responding to emerging conditions. Central to DAP is the use of monitoring systems and “trigger points” that signal when assumptions are no longer valid or when policy adjustments are required. This allows policymakers to delay irreversible commitments, preserve flexibility, and progressively adapt strategies as uncertainties unfold. It is widely applied in climate adaptation, infrastructure planning, water management, and disaster risk governance.

Dynamic Adaptive Policy Pathways (DAPP)

Dynamic Adaptive Policy Pathways extends the DAP logic by explicitly mapping multiple possible sequences of policy actions over time as interconnected “pathways”. Instead of considering only one adaptive strategy with contingencies, DAPP visualizes alternative routes that policymakers may follow depending on how future conditions unfold. These pathways are often represented graphically, showing when particular policies may become ineffective, reach thresholds, or require transition to other measures. A central concept in DAPP is the “adaptation tipping point,” where existing policies no longer meet objectives under changing conditions.

Failure Mode and Effects Analysis (FMEA)

Failure modes and effects analysis (FMEA) is a way to identify and evaluate all potential failures, such as errors or defects, in a manufacturing or assembly process, or a product or service. FMEA begins early in the product or system design process. It involves breaking down a system or process into smaller components and then analyzing each component to identify potential failures and their effects. Organizations using this method prioritize risks according to how serious their consequences are, how frequently they occur, how easily they can be detected, and how serious their effects are likely to be. FMEA is commonly used in industries such as automotive manufacturing, aerospace, and healthcare.

Foresight/Scenario-Based

Foresight is a structured approach for exploring and preparing for multiple plausible futures rather than attempting to predict a single most likely outcome. It uses scenario construction, horizon scanning, and systems analysis to map how technological, environmental, economic, or social drivers could interact over time under conditions of uncertainty. Instead of optimizing decisions for one forecast, foresight evaluates strategies across a range of contrasting scenarios, including high-impact, low-probability developments, in order to identify options that remain effective under many different futures. This makes it particularly useful in complex and uncertain domains, where probabilities are difficult to estimate or may change over time. By focusing on robustness, adaptability, and early warning signals, foresight helps policymakers and organizations design policies that reduce vulnerability to unexpected shocks and improve long-term resilience.

Info-Gap Theory

Info-gap theory is a decision-making framework designed to support choices under severe uncertainty, where probability distributions are unknown or unreliable. “An info-gap is a disparity between what is known, and what needs to be known in order to make a comprehensive and reliable decision”. Instead of attempting to estimate likelihoods, it focuses on robustness, defined as the ability of a decision to satisfy critical performance requirements across a wide range of possible deviations from a nominal model. The approach begins with a best-available estimate of the system (the “nominal model”) and then systematically explores how much uncertainty can be tolerated before performance goals are violated, effectively quantifying the resilience of a decision to model error. This leads to two key concepts: robustness functions, which measure how much uncertainty a strategy can withstand, and opportuneness functions, which assess potential upside benefits under favorable deviations. Info-gap theory is particularly useful in contexts where uncertainty is deep and non-probabilistic, such as environmental management, engineering design, and policy planning, because it prioritizes safe and reliable performance over optimality under assumed distributions.

Precautionary Principle

The precautionary principle is a risk governance approach stating that when an activity or technology poses a plausible threat of serious or irreversible harm, lack of full scientific certainty should not be used as a reason to delay preventive action. Rather than requiring definitive proof of danger before intervention, the principle shifts emphasis toward early caution and anticipatory regulation under conditions of uncertainty, particularly where potential consequences may be catastrophic or difficult to reverse. It is commonly applied in environmental policy, public health, biotechnology, and chemical regulation, where scientific knowledge may be incomplete, but the stakes are high. In practice, the precautionary principle can justify measures such as stricter safety standards, temporary restrictions, additional testing requirements, or adaptive monitoring systems while evidence continues to develop. For policymakers, the principle provides a framework for managing uncertainty in situations where waiting for complete information could substantially increase societal, environmental, or health risks, emphasizing prevention and resilience over reactive response.

Probabilistic Risk Assessment (PRA)

Probabilistic risk assessment (PRA) is a quantitative risk analysis approach that evaluates complex systems by explicitly modeling the likelihood and consequences of different failure scenarios. Rather than assuming fixed design cases alone, PRA decomposes a system into components and failure modes, then uses probabilistic models to estimate how combinations of technical failures, human error, and external hazards could lead to adverse outcomes. The output is typically expressed as risk metrics (for example, expected frequency of core damage or system failure), which allows decision-makers to compare risks across different configurations and interventions on a consistent scale. PRA is particularly useful for managing uncertainty in high-reliability domains because it makes assumptions explicit, integrates diverse sources of evidence (engineering data, operational experience, expert judgment), and highlights low-probability but high-consequence events. This supports more informed prioritization of safety investments and enables risk-informed decision-making that complements deterministic safety analysis and engineering judgment.

Risk Sensemaking

Risk sensemaking is a cognitive and organizational process through which individuals and institutions interpret, construct, and continuously update their understanding of risk in uncertain and evolving environments. Rather than treating risk as a fixed, objectively measurable quantity, it emphasizes that risk perception is shaped through interpretation of signals, shared narratives, prior beliefs, institutional routines, and social interaction. In practice, sensemaking involves collecting fragmented information, filtering ambiguity, and constructing a coherent account that guides action, often under time pressure and with incomplete or conflicting data. This process is iterative: as new information emerges, actors revise their understanding of what the risk is, how severe it may be, and what responses are appropriate. Risk sensemaking is particularly important in complex socio-technical systems where formal models are insufficient on their own, such as crises management, cybersecurity, and public health emergencies.

Robust Decision Making (RBM)

Robust Decision Making (RDM) is a scenario-based decision framework designed to support policy and strategic planning under conditions of deep uncertainty, where future conditions, probabilities, or system dynamics cannot be reliably predicted. Rather than seeking an optimal solution for a single expected future, RDM evaluates how different strategies perform across a large ensemble of plausible scenarios, including extreme and unexpected conditions. The approach typically combines computational modeling, exploratory scenario analysis, and vulnerability assessment to identify policies that remain effective, flexible, and resilient under many possible futures. A central feature of RDM is “stress testing,” in which strategies are systematically exposed to adverse conditions to determine where they may fail and how they can be adapted or improved. It is widely used in climate adaptation, infrastructure planning, water management, and national security.

Root Cause Analysis (RCA)

Root Cause Analysis (RCA) is a systematic investigative method used to identify the fundamental underlying causes of failures, accidents, or adverse events rather than focusing only on their immediate symptoms. The approach seeks to determine why an incident occurred by tracing chains of causation across technical, human, organizational, and procedural factors that contributed to the outcome. In short, RCA helps prevent recurrence of similar incidents. One common RCA technique is the Fault tree analysis (FTA). It is a top-down approach to identifying the causes of a hazard or system failure. It involves creating a diagram that maps out all the possible paths leading to the failure or hazard and then analyzing each path to identify potential causes and contributing factors. The factors typically include combinations of hardware failures, software failures, and human errors, with stakeholders in the risk analysis asking: “How might this happen?”. FTA can be used to assess risks associated with complex systems, such as nuclear power plants and aviation components.

Safety Management System (SMS)

A Safety Management System (SMS) is a formal, organization-wide framework for systematically identifying, assessing, and controlling safety risks in complex operational environments. Rather than treating safety as a set of isolated technical checks, an SMS integrates procedures, accountability structures, and continuous monitoring processes into a closed-loop system that actively manages risk over time. It is typically built around four core components: safety policy, safety risk management (hazard identification and risk assessment), safety assurance (performance monitoring and auditing), and safety promotion (training and safety culture development). The system relies heavily on reporting mechanisms, incident and near-miss data, and feedback loops to detect emerging hazards and evaluate the effectiveness of existing controls. Widely adopted in high-reliability industries such as aviation, rail, and energy, SMS frameworks help organizations move from reactive accident investigation toward proactive and predictive risk management, improving resilience in the face of operational uncertainty and complex socio-technical interactions.

Social Amplification Framework (SARF)

The Social Amplification of Risk Framework (SARF) is a conceptual model that explains how risk events can be intensified or attenuated as they are communicated through social, institutional, and cultural channels. It emphasizes that risk is not solely a function of physical hazard characteristics, but is also shaped by processes of information transmission, interpretation, and social response. As risk signals move through media coverage, expert discourse, public perception, and institutional behavior, they can be “amplified” when attention, concern, or behavioral responses escalate, or “attenuated” when risks are downplayed or fail to gain visibility. These dynamics can generate secondary effects, such as economic disruption, policy overreaction, or public mistrust, even when the underlying hazard remains unchanged. SARF is particularly useful for understanding why similar risk events can produce very different societal impacts depending on context, communication patterns, and institutional credibility.

Systemic Risk Management (SRM)

Systemic risk management is an approach to identifying, analyzing, and mitigating risks that arise from interdependencies within complex systems, where failures can propagate nonlinearly across networks and produce large-scale, cascading consequences. Rather than focusing on isolated components, it emphasizes the structure of connections between institutions, infrastructures, markets, or socio-technical systems, and how local disruptions can amplify through feedback loops, contagion effects, or common vulnerabilities. The approach typically combines network analysis, stress testing, and scenario-based exploration to assess how shocks, such as financial crises, infrastructure failures, or cyber incidents, can spread beyond their point of origin. Because systemic risks are often characterized by uncertainty, thresholds, and emergent behavior, management strategies prioritize resilience, redundancy, diversification, and adaptive capacity rather than precise prediction. This makes systemic risk management particularly relevant for domains such as finance, energy systems, public health, and critical infrastructure, where stability depends on the robustness of the entire system rather than the reliability of individual components.

Swiss Cheese Model

The Swiss cheese model is a systems-based accident causation framework that explains how catastrophic failures occur when multiple layers of defense, each containing inherent weaknesses or "holes," align in a way that allows hazards to pass through the entire system. The model emphasizes that accidents are rarely caused by a single error or technical malfunction; instead, they emerge from the interaction of latent organizational weaknesses, procedural deficiencies, human mistakes, and equipment failures distributed across different layers of protection. Each slice of "cheese" represents a defensive barrier (e.g. regulations, engineering safeguards, training, supervision, or operational procedures) while the holes represent temporary or structural vulnerabilities within those barriers. The model is widely used in aviation, healthcare, nuclear safety, and organizational risk management because it shifts attention away from blaming individual operators and toward understanding how systemic conditions enable failures to propagate.

Threat Modeling

Threat modeling is a structured risk analysis method used to identify, evaluate, and prioritize potential threats to a system, process, or organization before adverse events occur. The approach systematically examines how a system could be attacked, fail, or be exploited by analyzing assets, vulnerabilities, threat actors, attack pathways, and possible consequences. Widely used in cybersecurity, software engineering, and critical infrastructure protection, threat modeling often involves mapping system architectures and assessing where weaknesses may exist, what capabilities adversaries may possess, and which safeguards are necessary to reduce risk. Common methodologies, such as STRIDE, attack trees, and adversarial risk analysis, help organizations anticipate both technical and human-driven threats under uncertain conditions. Rather than reacting to incidents after they occur, threat modeling supports proactive risk management by identifying vulnerabilities early in the design and operational lifecycle.

What-If Analysis

What-if analysis is an exploratory risk assessment technique used to evaluate how systems, processes, or policies might respond under different hypothetical conditions or disruptive events. This method involves asking a series of "what if" questions to identify and evaluate potential risks, particularly those that might not be immediately apparent, and analyzing the potential consequences, vulnerabilities, and response options associated with each scenario. This method often looks for possible deviations from the design, construction, modification, or operating intent. Unlike strictly predictive models, what-if analysis is designed to explore uncertainty and uncover unexpected interactions, weak points, or cascading effects within complex systems. It is often used in conjunction with scenario planning, hazard analysis, and contingency planning to test the robustness of existing controls and identify where additional safeguards may be required. Because the approach is flexible and relatively simple to apply, it is widely used in engineering, emergency management, cybersecurity, and strategic planning.

Appendix 3 Examples of Policy Intervention

Market intrusiveness

Market intrusiveness is the degree to which a public policy intervenes in market dynamics by constraining, directing, or replacing decentralized decisions made by private actors.

- **Low**: Primarily enables or informs markets with minimal constraints on private decision-making (information)
- **Moderate**: Steers market outcomes through incentives or general regulations while preserving firms' autonomy over implementation (incentives).
- **High**: Substantially restricts private decision-making through mandates, prohibitions, or strong government direction (obligations).
- **Extremely high**: Transfers core economic decision-making from markets to the state through direct control of production, investment, pricing, or resource allocation (centralized control).

Examples of Policy Tools Available at the Model Layer

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Transparency	Model cards disclosure	Document that discloses model's architecture, training data, evaluation metrics and data, limitations & risks	U.S. (bill)	Low	Documentation and compliance costs; disclosure of commercially sensitive information
	Model registration	Registration of frontier and high-risk systems	EU	Low	Administrative burden
	Transparency report	Document that details risk management practices such as risk assessment, incident handling, data stewardship and mitigation practices	South Korea	Low	Disclosure of commercially sensitive information
	Incident reporting and investigation	Mandatory reporting and investigation of model's failures or events that could have caused harm	EU	Moderate	May discourage voluntary disclosure if penalties are high

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Risk management	Technical safety standards	Voluntary common technical standards for safety	U.S.	Low	Standards entrench particular technical approaches
	Safety Management Systems	Organizational processes for identifying, assessing, managing and monitoring AI risks	U.S.	Low	Organizational cost; voluntary frameworks may produce uneven uptake
	Safety evaluation and benchmarking	Standardized performance and safety testing and measurement	Singapore	Moderate	Risk of optimizing for benchmarks rather than safety
	Early-warning system and monitoring	Ongoing monitoring of deployed systems to detect failures and deviations from initial design	EU	Moderate	Data collection burden; privacy concerns and risk of false alarm
	Accuracy, robustness and cybersecurity requirements & red-teaming	Requirements to test and mitigate failures, vulnerability, adversarial manipulation and cybersecurity risks	EU	High	Increased development and testing costs, particularly for smaller developers
	Safety Protocols	Predefined "if-then" commitments specifying actions to be taken when specific risk or capability threshold is reached	N/A but Industry Initiative	High	Heterogenous commitments and lack of common minimum baseline
	Human oversight	Mechanisms for human review and intervention	EU	High	Reduced efficiency and automation benefits

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Accountability	Safety/Security institutes	Public technical research and evaluation institutions that can test advanced and dangerous capabilities	UK	Low	Public evaluation costs; possible regulatory capture; potential duplication with existing sectoral regulators; delay release
	Whistleblower protections	Protection against retaliation for employees who report specified AI risks	EU	Low	Potential disclosure of commercially sensitive information
	Independent audits	Third-party assessments of safety measures	U.S. (Illinois)	Moderate	Audit cost could create barriers to entry for startups and small businesses
	Conformity assessments	Documentation that proves that AI systems meet specific requirements	EU	Moderate	Administrative burden and possible delay in deployment
	Security Assessment	Mandatory security assessment that covers data and algorithm	China	High	Administrative burden and possible delay in deployment
	Certification & licensing schemes	Formal certification or licensing of frontier models prior to deployment	N/A	High	Potential delay to deployment; significant compliance cost; possible barrier to entry for startups and small businesses
	Liability regimes	Legal rules attributing responsibility and potential compensation for harm associated with AI systems	EU	High	Litigation and insurance cost; could discourage experimentation and risk-taking

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Research & learning	Research funding	Public funding for research into AI safety, security, alignment, controllability, interpretability or other risk-reduction objectives	UK	Low	Opportunity cost relative to other priorities; research may not translate in practical solutions
	Interpretability & explainability requirements	Requirements to provide information that allow users or deployers to understand system capabilities, limitations and outputs	EU	Low	Compliance cost and possible constraints on the use of highly capable but opaque models
	Vulnerability & information sharing mechanisms	Voluntary sharing of information on vulnerabilities, exploitable flaws, threats and capabilities of concern	U.S.	Low	Potential exposure of vulnerabilities; sharing of sensitive information; competitive concerns may limit participation
	Regulatory sandboxes	Controlled environments in which AI systems can be tested under regulatory supervision before or during wider deployment	Singapore	Moderate	Limited scalability; potential regulatory capture or uneven access
Capabilities & deployment controls	Pre-deployment review	Review or access to advanced model prior to deployment	U.S.	Moderate	Potential release delays; administrative burden; voluntary participation may limit coverage
	Capability-based regulation	Regulations triggered by model's capabilities	EU	High	Regulatory uncertainty around thresholds
	Phased deployment	Gradually deploys an AI system across users, settings, or levels of access, allowing risks to be monitored and addressed before broader deployment.	N/A	High	Slower access to benefits and innovation
	Autonomous replication prohibition	Prohibition of specified forms of autonomous self-replication or uncontrolled replication by AI systems	N/A	Extremely High	Could prohibit useful future applications
	Autonomous goal-pursuit restrictions	Restrictions on AI systems with specified forms or degree of autonomous goal pursuit	N/A	Extremely High	Could constrain beneficial agentic applications
	Scaling restrictions	Restrictions on increasing model capabilities, compute, size or other development inputs	N/A	Extremely High	May slow beneficial scientific and commercial development

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Emergency prevention & response	High-risk pre-deployment requirements	Completion of specific requirements, such as documentation, risk-management or testing, prior to deployment	EU	High	Significant compliance burden and possible barriers to market entry
	Incident response protocol	Establish procedures, responsibilities and timelines to respond to critical incidents	U.S. (Illinois)	High	Risk of over-reporting or inconsistent incident classification
	Kill-switch requirement/ emergency control	Requirement to maintain a technical capability to shutdown specified AI-enabled systems	U.S. (proposal)	Extremely High	Technical feasibility; ambiguity over activation conditions
	Emergency government intervention powers	Government authority to intervene under specific conditions	N/A	Extremely High	Risk of government overreach; uncertainty for firms
	Development pauses or slowdown	Temporary suspension or slowdown of development	N/A but Industry initiative	Extremely High	Risk of uneven international compliance; loss of beneficial innovation
	Moratoria on development or deployment	Prohibition of specified AI development or deployment activities	N/A	Extremely High	Strong innovation and competitiveness costs

Examples of Policy Tools Available at the Individual Layer

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Transparency	Political advertising transparency	Disclosure of information about political advertising and whether content has been AI-generated	EU	Low	Compliance burden on campaigns and platforms
	AI transparency obligations	Disclosure to users that they are interacting with AI	EU	Low	Implementation and interoperability costs
	AI-generated content labeling & watermarking	Disclosure of synthetic content and embedment of provenance signals	EU	Low	Labels can be removed or circumvented
	Digital signatures	Cryptographic or other provenance mechanism that authenticate content origin	U.S. (California)	Low	Infrastructure, interoperability and implementation cost
Risk management	Cybersecurity safeguards & model hardening requirements	Requirements to protect AI systems against attacks, manipulation, unauthorized access and other security threats	EU	High	Compliance costs; possible reduction in model flexibility
	Misuse detection and reporting	Ongoing monitoring of deployed systems to detect and report misuses	N/A but proposal by Rand	High	Privacy concerns; false positives; monitoring costs
	Mandatory misuse safeguards	Required safeguards to prevent misuse such as criminal activities or chemical and biological weapons prior to deployment	N/A but industry techniques	High	Reduced functionality for some legitimate uses
	Access controls on the material needed to create chemical or biological weapon	Restricted physical access to chemicals and material used for chemical or biological weapons	U.S.	Extremely high	Compliance cost; reduced accessibility for legitimate research

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Deployment controls	Know-your-customer	User identity verification for users accessing specified services or capabilities	U.S.	High	Privacy and anonymity concerns
	Content moderation	Measures to identify, restrict or remove harmful or prohibited AI-generated output	UK	High	Free expression concerns; moderation errors
	Compute licensing	Licensing or authorization requirements governing access to advanced compute resources	U.S.	Extremely high	Major barrier to access and possible concentration of advanced compute
	Weight dissemination licensing	Licensing or authorization requirements governing access to advanced AI model weight	U.S.	Extremely high	Market concentration and reduced access to open research

Examples of Policy Tools Available at the Societal Layer

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Resilience building	AI and digital literacy programs	Public education on AI systems, their use and impacts, including the ability to identify and critically assess AI-generated content	Finland	Low	Limited measurable impact; sustained public investment; long time horizon for benefits
	Trusted information initiatives and independent media support	Public support for reliable information ecosystems and trusted news sources	UK	Low	Public funding dependency; potential concerns about government influence over information
	Fact-checking support	Third-parties verification of information to counter AI-generated misinformation	Brazil	Low	Perceived political bias; sustainability concern
	Amendments to electoral law/ election-specific rules	Rules governing the use of AI-generated content during electoral context	Brazil Canada	Low	Risk of overreach into political discourse and challenges in defining prohibited content
	Continuous training, upskilling & reskilling programs	Training programs for continuous skill development or for labor transition	Singapore	Moderate	Fiscal cost and long implementation horizon
	Adapting unemployment insurance	Adjusting income-support systems for AI-related job displacement	U.S. (bill, New Jersey)	Moderate	Fiscal cost; potential disincentive to rapid re-employment
	Transition assistance program	Targeted support for workers displaced or at risk of displacement	N/A but RAISE Initiative	High	Fiscal cost
	Universal basic income	Regular basic income support to workers	Finland	Extremely high	Very high fiscal cost; uncertain labor effect

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
	Collective bargaining	Worker representation and collective bargaining mechanism to address workplace impact of AI	Sweden	Moderate	Potentially higher labor cost
	Employment protections	Rules requiring employers to meet specified obligations when AI-related changes affect employment or dismissal	France	High	Potentially higher labor cost; reduced flexibility in workforce adjustment
	Consumer protection law	Adapt general consumer protection law to AI-related products and services	Australia	High	Compliance cost
	Privacy protections	Rules governing the collection, use and processing of personal information	EU	High	Reduced data availability
Rights protections	Copyright protections	Rules governing the use of copyrighted works in AI development and training	EU	High	Reduced data availability
	Anti-discrimination rules	Rules requiring safeguards against discrimination in AI-assisted decision-making	U.S. (California)	High	Increased compliance and litigation cost
	Right to explanation	Affected person's right to receive meaningful information about the role of AI in certain decisions	EU	High	Administrative burden and potentially reduced efficiency
	Fundamental rights impact assessment	Structure assessment of how AI systems may affect fundamental rights and what mitigations are required	EU	High	Compliance cost
	Right to human review	Right to have an automated decision reviewed by a person under specified circumstances	EU	High	Reduced efficiency and increased cost
	AI indicators	Statistical indicators designed to monitor AI adoption, capabilities, economic effect or societal impacts	South Korea	Low	Maintenance cost and longer term horizon to build longitudinal data
Research and learning	Simulations, foresight and scenario-building	Structured exploration of plausible future AI development and their societal implications	UK	Low	Opportunity cost
	AI Observatories and research centers	Public institutions that monitor and study the effect of AI on society and the economy	UK	Low	Sustained public funding; risk of duplication across institutions

Examples of Policy Tools Available at the International Layer

POLICY OBJECTIVE	POLICY TOOL	DESCRIPTION	EXAMPLE	MARKET INTRUSIVENESS	POTENTIAL TRADE-OFF
Coordination	International scientific panels	Expert advisory bodies that provide scientific evidence of the state of play	UN	Low	Slow consensus formation; difficulty to translate scientific assessment into policy
	International Safety Institutes network	International cooperation among AI safety/security institutes on research, evaluation, information sharing	International Network of AI Safety Institutes	Low	Coordination costs, differences in national mandates and potential duplication
	Global incident reporting system	International mechanism for collecting and sharing information about AI incidents	OECD	Low	Confidentiality and data-quality concerns; participation may be uneven
	International transparency reporting	Common approaches to reporting information about AI systems, risks and safety practices	G7 HAIP	Moderate	Agreement on common reporting requirements can be time-consuming; reporting burden
	International standards	Agreement on common terminology, technical specifications, evaluation practices and procedural rules	ITU	Moderate	Standard development takes time; may become obsolete by time of adoption
	International treaties	Agreement on common rules regarding the design, development, deployment and use	Council of Europe	Moderate	Standard development takes time; may become obsolete by time of adoption

Appendix 4
Steps for Developing Policy Responses to AI Risks

1. **Identify and define the risk based on evidence**
   - O Characterize:
     - O Potential harms and affected stakeholders
     - O Likelihood, severity and reversibility of potential harm
     - O Type and degree of uncertainty
     - O Time horizon (short-, medium-, or long-term)
     - O Layer of the risk (model, individual, society, international)
     - O Existing evidence and knowledge gaps

2. **Assess whether government intervention is warranted**
   - O Consider:
     - O Market failures
     - O Information asymmetries
     - O Externalities
     - O Existing legal and regulatory frameworks
     - O Existing industry initiatives
   - O If intervention is not currently justified, continue monitoring and reassess as new evidence emerges.

3. **Define policy objectives**
   - O Specify the desired outcomes of any intervention
   - O Ensure objectives are clear, measurable, and achievable

4. **Identify potential policy instruments**
   - O Explore the full range of available tools, including:
     - O Binding legal instruments
     - O Soft governance instruments
     - O Experimental instruments
     - O Programmatic instruments
     - O Economic incentives instruments
     - O Institutional instruments

5. **Evaluate policy options**
   - O Assess each option against relevant criteria, such as:
     - O Effectiveness
     - O Cost and benefits
     - O Feasibility and enforceability
     - O Flexibility and adaptability
     - O Social and political acceptability
     - O Potential unintended consequences
     - O Trade-off

6. **Select the preferred policy approach**
   - O Choose the option, or combination of options, that best achieves the policy objectives while minimizing costs and unintended effects (this can include maintaining the status quo)

7. **Develop an implementation plan**
   - O Define roles, responsibilities, timelines, and resources
   - O Anticipate implementation challenges and mitigation strategies
   - O Identify indicators, trigger points and contingent actions

8. **Monitor, evaluate, and adapt**
   - O Track outcomes and compliance
   - O Assess whether objectives are being met
   - O Revise the intervention as evidence, technology, and risks evolve