Philosophical Studies
https://doi.org/10.1007/s11098-025-02301-3

# Two types of AI existential risk: decisive and accumulative

**Atoosa Kasirzadeh**[1]

Accepted: 7 February 2025


## Abstract

The conventional discourse on existential risks (x-risks) from AI typically focuses on abrupt, dire events caused by advanced AI systems, particularly those that might achieve or surpass human-level intelligence. These events have severe consequences that either lead to human extinction or irreversibly cripple human civilization to a point beyond recovery. This decisive view, however, often neglects the serious possibility of AI x-risk manifesting gradually through an incremental series of smaller yet interconnected disruptions, crossing critical thresholds over time. This paper contrasts the conventional *decisive AI x-risk hypothesis* with what I call an *accumulative AI x-risk hypothesis*. While the former envisions an overt AI takeover pathway, characterized by scenarios like uncontrollable superintelligence, the latter suggests a different pathway to existential catastrophes. This involves a gradual accumulation of AI-induced threats such as severe vulnerabilities and systemic erosion of critical economic and political structures. The accumulative hypothesis suggests a boiling frog scenario where incremental AI risks slowly undermine systemic and societal resilience until a triggering event results in irreversible collapse. Through complex systems analysis, this paper examines the distinct assumptions differentiating these two hypotheses. It is then argued that the accumulative view can reconcile seemingly incompatible perspectives on AI risks. The implications of differentiating between the two types of pathway—the decisive and the accumulative—for the governance of AI as well as long-term AI safety are discussed.

## 1 Introduction

Recent advances in machine learning have sparked intense debate about the existential risks (x-risks) associated with artificial intelligence (AI) systems.[1] Central to this debate is a concern about the mechanisms by which AI could cause existential

[1] For a review of recent narratives about AI and future, see Gilardi et al. (2024).

✉ Atoosa Kasirzadeh
atoosa.kasirzadeh@gmail.com

[1] Carnegie Mellon University, Pittsburgh, United States

Published online: 30 March 2025

catastrophes. In direct response to this concern, this paper explores: What are the distinct types of pathway by which AI systems could cause existential catastrophes? Conventional discourse on AI existential catastrophes typically portrays them as sudden, decisive events, caused by artificial general or super intelligence (Bostrom, 2013) or extremely powerful (non-general) AI (Carlsmith, 2022).

Contrasting this conventional decisive viewpoint, this paper introduces the *accumulative AI x-risk hypothesis* as an alternative lens. The accumulative hypothesis posits that AI x-risks do not exclusively materialize as decisive, high-magnitude global events caused by extremely powerful AI such as artificial general or super intelligence. Instead, locally significant AI-driven disruptions can accumulate and interact over time, progressively weakening the resilience of critical societal systems, from democratic institutions and economic markets to social trust networks. When these systems become sufficiently fragile, a modest perturbation could trigger cascading failures that propagate via the interdependence of these systems. The failures amplify and reinforce each other by network effects and feedback loops, potentially leading to a globally irreversible civilizational collapse.

This paper develops the outline of an accumulative perspective on AI x-risk, by examining how multiple types of AI-induced risks could compound and cascade over time to gradually bring about an AI-generated existential catastrophe. By applying complex systems analysis—an approach not typically used in AI x-risk scholarship—I defend the significance of accumulative AI x-risk, and consequently argue for a fundamental reconceptualization of AI x-risk governance. I discuss the epistemic and pragmatic benefits of attending to the accumulative perspective *if* such risks are to be governed effectively.

## 2 AI x-risk: preliminaries

### 2.1 Concepts of risk

At most basic, risk relates to some characterization of uncertainty about potential (adverse) outcomes (Dean, 1998; Hansson, 2010; Aven, 2012). According to the ISO 31000 standard (International Organization for Standardization, 2018), risk is defined as "the effect of uncertainty on objectives." The Society for Risk Analysis (Aven et al., 2018) defines it as "uncertainty about and severity of the consequences of an activity," while the U.S. Environmental Protection Agency (U.S. Environmental Protection Agency, 2024) defines risk as predicting "the probability, nature, and magnitude of the adverse effects that might occur."

More concretely, risk has been analyzed through at least four distinct, though not mutually exclusive, interpretations.

First, risk has been used to mean an *unwanted event* that may occur (Carlsmith, 2022). For instance, "AI systems exhibiting power-seeking behavior pose a major existential risk to humanity." Second, risk has been used to denote the *cause(s) of an unwanted event* that may occur (Weidinger et al., 2022). For example, "Insufficient testing of generative AI models could cause severe harm in AI deployment." Third, risk has been used to mean the *probability of an unwanted event* that may occur

(Lowrance, 1976). An example is "The risk that a large language model will generate incorrect information in response to user queries is approximately 5%." Fourth, risk has been used to mean a *statistical expectation value*—the product of probability and consequence—of an unwanted event that may occur (United Nations International Strategy for Disaster Reduction, 2009). For instance, "The risk of LLM hallucination in a medical context can be calculated by multiplying the probability of giving incorrect medical advice (1%) by the average cost of medical liability claims ($500,000), yielding an expected cost of $5,000 per consultation."

Most researchers define existential risks as the potential for events that would result in the extinction of humanity or an unrecoverable decline in humanity's potential to thrive (Bostrom, 2013; Ord, 2020a).[2] Existential catastrophes are a class of potential events that may originate from natural causes, such as a supervolcanic eruption; anthropogenic sources, like nuclear conflict; or emerging threats, such as misaligned artificial superintelligence.[3] This paper concentrates on existential catastrophes induced by AI (AI x-catastrophes) and their associated risks.

In AI x-risk literature, researchers employ either of the qualitative or quantitative interpretations described above, each highlighting specific information about x-risks but also presenting certain limitations (Tonn & Stiefel, 2013; Beard et al., 2020).[4] This paper aims to maintain neutrality between the different interpretations of AI x-risk (i.e., unwanted AI x-catastrophe, cause(s) of unwanted AI x-catastrophe, probability of unwanted AI x-catastrophe, and statistical expectation value of AI x-catastrophe), drawing on a broad range of perspectives to analyze AI x-catastrophes and their associated risks. While I employ the causal interpretation of AI x-risk for illustrative purposes, alternative interpretations could remain equally valid; though thehe probabilistic and statistical expectation value interpretations of AI x-risk warrant their own detailed investigation elsewhere.

The conventional discourse on AI x-catastrophes portrays them as decisive, large-scale events caused by highly advanced AI systems, often referred to as

[2] Some definitions, such as the one proposed by Bostrom (2013, p. 15), broaden the scope of existential threats to include not only human life but all sentient beings: "An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development." Despite this broader definition, this paper deliberately employs the term "humanity" as the focal subject of AI-induced existential catastrophes (Ord, 2020a) This choice is pragmatic and reflects the predominant narrative in AI x-risk scholarly literature and public discourse, which has traditionally concentrated on catastrophic impacts on humanity and human civilization. Additionally, this choice aligns with the stated missions of leading AI research companies like Google DeepMind and OpenAI in their pursuit of AGI or ASI development "for the benefit of humanity." My choice of terminology, however, does not diminish the moral significance of non-human sentient beings.

[3] For a historical examination of existential risks, see Torres (2023b).

[4] Quantitative methods for estimating AI x-risk require employing statistical models and subjective probability analysis. These methods are valuable for a seemingly structured way of estimating risks. However, they face limitations due to the reliance on available data, which can be scarce or unreliable for unprecedented risks, and the difficulty of robustly estimating existential threats from AI in numerical terms. Qualitative methods involve non-numerical analyses such as scenario development, expert interviews, and ethical deliberations (Technology and Science Insights and Foresight, 2023). The qualitative methods are particularly useful in exploring the nuanced, complex, and often speculative aspects of AI risks, especially in areas where reliable empirical evidence is scarce or unavailable.

artificial general intelligence (AGI) or artificial superintelligence (ASI).[5] The idea that advanced machines can pose significant risks is not novel and has historical antecedents.[6] Samuel Butler (1863, p. 185), a novelist and literary critic, alluded to the possibility of machines dominating humanity. This concern was later picked up by the renowned mathematician Alan Turing (1950), who warned of intelligent machines eventually taking control. Norbert Wiener (1960), a founder of the field of cybernetics, cautioned against entrusting machines with purposes misaligned with human intentions or desires. Similarly, the mathematician Irving J. Good (1966) expressed concerns about the creation of "ultraintelligent machines."

More recently, philosophers like Nick Bostrom (2002, p. 7) drew systematic attention to the existential threats posed by ASI: "When we create the first superintelligent entity, we might make a mistake and give it goals that lead it to annihilate humankind, assuming its enormous intellectual advantage gives it the power to do so."[7] Computer scientists such as Stuart Russell (2019) and physicists like Max Tegmark (2018) echoed similar concerns, stressing the x-risks of ASI beyond human control. Such views about ASI x-risk frequently hinge on two key theses: orthogonality and instrumental convergence (Bostrom, 2012, 2014).[8]

The orthogonality thesis posits that an advanced AI system's intelligence level and its final goals are orthogonal.[9] This implies that an AI system could have any combination of final goals --- beneficial or harmful --- and intelligent capabilities. That is, the space of possible goals is limitless, and intelligence alone does not constrain which of these goals an agent might adopt. The instrumental convergence thesis holds that diverse final goals often have similar instrumental sub-goals—like self-preservation or resource acquisition—as these sub-goals are useful for achieving almost any final objective. This means that for a wide range of ends, instrumental rationality converges on similar means. An ASI might therefore pursue actions harmful to humanity not out of malice, but as instrumental means toward achieving its (programmed) goals. Two hypothetical scenarios illustrate how conventional models of ASI could cause x-catastrophic outcomes.

[5] In AI lexicon, AGI represents capabilities comparable to human intelligence, while ASI denotes capabilities surpassing human intelligence. Traditional arguments typically link x-risks primarily to AGI or ASI (e.g., Bostrom (2014); Ord (2020a, b)), though some suggest that AI x-risks could also emerge from AI systems with extreme but narrow capabilities (Carlsmith (2022)). This paper's focus on AGI/ASI aligns with influential voices in x-risk studies such as major AGI research companies. However, my core argument about the nature of x-risks holds whether we consider AGI/ASI or narrower AI systems with extreme, uncontrollable capabilities in specific domains. Thus, this terminological choice does not affect my characterization of conventional views on AI x-risk.

[6] For a comprehensive historical review, see Torres (2023a).

[7] While Bostrom's work has systematized the discussion of x-risks from ASI, the original post-2000 discussions trace back to Eliezer Yudkowsky's views concerning AGI, ASI, and their associated x-risks. Ben Goertzel (2015) explores the early conceptual evolution of these topics, tracing their roots to Yudkowsky's initial informal explorations. In this paper, I primarily reference Nick Bostrom as a representative figure who brought systematic and philosophical depth to concepts that Yudkowsky and others initially introduced and examined in speculative media and blog posts.

[8] For a critical discussion of the orthogonality and instrumental convergence theses, see Müller and Cannon (2022).

[9] See Bostrom (2012, p. 73) and Bostrom (2014, p. 107).

In a thought experiment popularized by Bostrom (2003), an ASI is given the seemingly innocuous and simple goal of maximizing paperclip production. Even with this simple objective, the ASI could pursue instrumental sub-goals: it might eliminate humans to prevent deactivation (self-preservation) or convert their bodies into paperclips (resource acquisition). Consequently, the ASI's optimal future could become one abundant with paperclips but devoid of humans—not because intelligence necessitates this goal, but because instrumental sub-goals are (putatively) rational steps toward satisfying its given optimization objective. That is, the paperclip maximizer illustrates how an AI with an apparently harmless goal could pose x-risk via the rational pursuit of instrumental sub-goals like resource acquisition and self-preservation.[10]

In a structurally similar thought experiment, Russell and Norvig (2010, p. 1039) describe the following scenario attributed to Marvin Minsky: an advanced AI tasked with proving the Riemann hypothesis might appropriate Earth's resources to build supercomputers, endangering humanity in pursuit of a mathematical proof. Like the paperclip maximizer, this example is supposed to illustrate how an advanced AI pursuing a benign goal could pose x-risk to humanity. Both scenarios demonstrate a key pattern: an advanced AI system optimizing for a specific final goal could rationally pursue sub-goals which are catastrophic to humanity, not from malice, but as instrumental steps toward its optimization goal.

### 2.2 Decisive AI x-risk

The conventional view, sketched above, frames AI x-catastrophes as arising from the decisive actions of ASI. Toby Ord (2020a, p. 20), a prominent x-risk scholar, explicitly endorses the decisive character of x-catastrophes: "I take on the usual sense of catastrophe as a single decisive event rather than any combination of events that is bad in sum. A true existential catastrophe must by its very nature be the decisive moment of human history the point where we failed." This conventional framing suggests that AI x-risk manifests as sudden, cataclysmic events that either eradicate humanity or irreversibly curtail its potential. I articulate this perspective in terms of the decisive ASI x-risk hypothesis.

> **Decisive ASI x-risk hypothesis**: x-risk from ASI is the possibility of abrupt large-scale events that lead to humanity's extinction or cause an unrecoverable decline in its potential.

Both Bostrom's portrayal of ASI pursuing destructive goals and Ord's characterization of x-catastrophes as singular, defining events exemplify the conventional view: AI x-risk manifests as sudden, decisive moments of overwhelming impact. The decisive AI x-risk, according to this framing, is the expected uncertainty of the occurrence of such conclusive events as catalysts for x-catastrophic outcomes.

[10] The paperclip maximizer is characterized by some as a "Squiggle minimizer": a highly intelligent optimizer pursuing goals alien to human values that could inadvertently destroy humanity by consuming vital resources in pursuit of its objectives. See LessWrong (2024).

The decisive hypothesis, however, overlooks an alternative type of causal pathway leading to AI x-catastrophes. This alternative involves the gradual accumulation of smaller, seemingly non-existential, AI risks eventually surpassing critical thresholds.[11] These risks are a subset of what typically is referred to as *ethical* or *social* risks. In the rest of this paper, the terms "AI ethical risk" and "AI social risk" are used interchangeably.

### 2.3 AI risks: existential versus social

The prevailing discourse on risks from AI distinguishes between AI x-risks and AI social risks as separate and distinct categories. This separation has been a mainstream trend: x-risk from superintelligent or "strong" AI are often contrasted with normal non-existential risks (Hendrycks & Mazeika, 2022, p. 3) or with ethical and social risks (Weidinger et al., 2021, p. 7). Several other examples of this contrast can be found on social media platforms (e.g., Twitter) and popular media.[12] In a notable instance, Turing Award Laureate, Geoffrey Hinton, who departed from his position at Google to openly discuss x-risks from AI, emphasized during a recent interview that his concerns about existential risk "are different" from the concerns of Timnit Gebru, a computer scientist and responsible AI researcher, about AI ethical risks that are not "existentially serious" (CNN, 2023).

Typically, existential and social risks are demarcated along the lines of locality—i.e., the scope of risk—and severity—i.e., whether impacts are recoverable and limited, or irreversible and catastrophic for (human) civilization (Bostrom (2013), Amodei et al. (2016). While this demarcation could be pragmatically insightful, there is a notable gap in exploring the *relationship* between x-risks and the evolution of ethical concerns in a substantial manner.[13]

[11] Setting a critical threshold for AI social risks requires a multi-faceted approach that considers various factors, including risk assessment (i.e., analyzing the potential impact and likelihood of AI-related risks through both quantitative and qualitative methods), historical precedents and trends in AI development (i.e., insights into how risks have evolved and reached critical points in the past), expert consensus (i.e., gathering a diverse range of professional perspectives in AI, ethics, and risk assessment to ensures a comprehensive understanding of potential risks), and dynamic monitoring (i.e., regular updating of the threshold to reflect new developments and societal shifts to maintain its relevance and effectiveness). Evaluating the value of critical thresholds is beyond the scope of this paper and the subject of examination elsewhere.

[12] See, for example, Schechner and Seetharaman (2023), Richards et al. (2023), and Nature Editorial Team (2023).

[13] This disconnect might stem from two key reasons. First, discussions about decisive AI x-risk historically have revolved around reinforcement learning and agent-environment models, concentrating on existential threats from agency-based models. Ethical risks, in contrast, embrace a more expansive approach, covering non-reinforcement approaches to AI development. Second, prominent conventional voices in ASI x-risks, such as Bostrom, Ord, and Yudkowsky, often subscribe to normative worldviews like rationalism, effective altruism, or longtermism. These prescriptive worldviews may be perceived as either problematic or tangential by some who are deeply engaged in the multifaceted discussion of ethical risks. The resultant divergence in normative viewpoints on guiding AI risk discourse and community priorities has led to a significant schism between these domains of risk (for a detailed analysis of such community divergences, see for example Ahmed et al. (2023)). A thorough exploration of these distinctions, however, falls outside the scope of this paper.

The social risks of AI systems have been analyzed across different domains—from language models and their multimodal variants (Bender et al., 2021; Weidinger et al., 2022; Bird et al., 2023) to recommender systems (Milano et al., 2020; Deldjoo et al., 2024), to name just a few examples. While a full exposition of these risks is beyond the scope of this paper and has been extensively discussed in the above references, here I provide a brief categorization of seven risk classes that are shared in most AI social risk taxonomies.

*Manipulation and deception risks* include AI systems that cause harm by manipulating human perception or behavior via targeted or unwanted persuasion techniques, such as emotional exploitation (Kasirzadeh & Evans, 2023; Carroll et al., 2024; Park et al. (2024). *Misinformation and disinformation risks* arise from AI systems generating and amplifying false content at scale, enabling the spread of propaganda and undermining public trust and discourse (Sharma et al., 2019; Quach, 2020; Lin et al., 2021; Kay et al., 2024). *Malicious use risks* include the weaponization of AI systems for cyber attacks, the deployment of AI-enabled drones and other physical systems for physical attacks, and the automation of social engineering attacks (Brundage et al., 2018). *Insecurity and information threat risks* arise when AI systems reveal sensitive personal data or lead to an unintended disclosure of protected information (Carlini et al., 2021). *Discrimination and hate speech risks* arise by biased AI systems perpetuating systemic inequalities or generating targeted harmful content (Buolamwini & Gebru, 2018; Obermeyer et al., 2019). *Surveillance, rights infringement, and erosion of trust risks* originate from AI-powered mass surveillance systems and persistent monitoring that could lead to loss of privacy (Dwork, 2006; Tucker, 2018) and loss of trust in ruling institutions (Nowotny, 2021). *Environmental and socioeconomic risks* include ecological damage from the enormous energy consumption and carbon footprint associated with large-scale AI training and deployment (Luccioni et al., 2025) alongside widespread economic disruption via labor market transformation, worker displacement as automation renders certain job categories obsolete, or psychological harm inflicted on low-wage workers tasked with labeling content to train safe AI systems (Korinek & Stiglitz, 2018; Pashentsev, 2021; Perrigo (2023); Eloundou et al., 2024).

### 2.4 Accumulative AI x-risk

As an alternative to the decisive AI x-risk hypothesis, the gradual and cumulative progression of *critically significant* social risks forms a different type of pathway to AI x-catastrophes. I articulate this perspective in terms of the accumulative AI x-risk hypothesis.

> **Accumulative AI x-risk hypothesis:** AI x-risk results from the build-up of a series of smaller, lower-severity AI-induced disruptions over time, collectively and gradually weakening systemic resilience until a triggering event causes unrecoverable collapse.

According to this alternative hypothesis, AI x-catastrophes could emerge not from a decisive event, but from the cumulative impact of multiple interconnected

AI-induced adverse events over time. As compared to the decisive hypothesis, the accumulative hypothesis suggests a different type of causal pathway to AI x-catastrophe: a path wherein a succession of lower-severity, yet cumulatively significant, disruptions deeply erode the systemic resilience of the global system, radically disrupting critical socioeconomic and sociopolitical equilibrium. This weakened state potentially primes the global system for an unrecoverable collapse, particularly when further stressed by external events.

Figure 1 provides a schematic illustration contrasting the decisive and accumulative models of AI x-risk. The decisive scenario (blue path) represents the conventional view where a sudden catastrophic event—such as one caused by the arrival of ASI—leads to rapid, irreversible consequences. In contrast, the accumulative scenario (red path) shows how existential catastrophes can arise from a gradual accumulation of multiple smaller disruptions that compound and amplify over time. These accumulating disruptions gradually erode system stability, potentially crossing critical thresholds until a triggering event occurs.[14]

The gradual nature of the accumulative AI x-risk hypothesis can be likened to other global existential threats such as climate change.[15] The accumulative perspective is structurally akin to the incremental rise in greenhouse gases contributing to climate change, where each individual emission seems minor, but collectively, they lead to significant and potentially irreversible changes in the Earth's climate.

In the rest of this paper, I employ complex systems analysis to further analyze the two distinct pathways that could lead to AI x-catastrophes. The decisive pathway is based on the assumption that direct, catastrophic failures are triggered by (a) superintelligent system(s), while the accumulative pathway describes how systemic instabilities arise from interactions or interrelations between multiple AI-driven disruptions. This systems analysis perspective allows us to formulate the key assumptions underlying each hypothesis in greater detail, and examine how different patterns of causation and connectivity could produce AI x-catastrophes via fundamentally different mechanisms.

Before proceeding further, a crucial clarification is in order. The emphasis of this paper on conventional discussions of AI x-risk does not imply that all discourse on

[14] The time axis is purely illustrative and not tied to specific units, as the actual temporal development of these scenarios remains uncertain. The figure aims to capture qualitative differences in catastrophic events caused by AI rather than make specific predictions about timing or precise severity levels. The zigzag pattern in the accumulative scenario represents potential local recoveries, though the overall trend shows increasing systemic fragility over time.

[15] Ord (2020b, p. 28) acknowledges the decisive versus accumulative nature of global catastrophic events: "Nuclear weapons and climate change have striking similarities and contrasts. They both threaten humanity due to major shifts in the Earth's temperature, but in opposite directions. One burst in upon the scene as the product of an unpredictable scientific breakthrough; the other is the continuation of centuries-long scaling-up of old technologies. One poses a small risk of sudden and precipitous catastrophe; the other is a gradual, continuous process, with a delayed onset-where some level of catastrophe is assured and the major uncertainty lies in just how bad it will be. One involves a classifed military technology controlled by a handful of powerful actors; the other involves the aggregation of small effects from the choices of everyone in the world." However, this point has not been made in the context of existential risks from AI systems.

AI x-risk has been confined to the decisive hypothesis. This discourse is expanding, and an increasing number of scholars who do not necessarily endorse a decisive viewpoint are engaging to some degree with some version of the accumulative-type hypothesis (see, for example, Bucknall and Dori-Hacohen (2022), Hendrycks and Mazeika (2022), Shevlane et al. (2023), and Bales et al. (2024)); yet, no expansive and exclusive philosophical treatment of this hypothesis has been published. Nevertheless, this paper primarily focuses on the conventional decisive viewpoint on AI x-risk as this viewpoint--- historically entrenched and widely regarded as the predominant narrative --- has been a source of considerable debate and disagreement within various academic circles and public spheres. Its longstanding and prevalent nature in the field has contributed significantly to the imagination of majority about AI x-risk. This paper focuses on developing the accumulative perspective on AI x-risk which has not been adequately represented or robustly defended in the philosophical literature. It is my hope that this addition would facilitate a more unified and constructive dialogue about different kinds of AI risk and their relations.

## 3 Complex systems analysis for AI x-risk

AI x-catastrophes and their associated risks appear within our *complex global system*. Understanding how these catastrophes could occur requires analyzing how various elements—humans, AIs, and institutions—interact across multiple domains within this system. Complex systems analysis, which provides conceptual and mathematical instruments for analyzing relations and interactions between system components, enables us to trace how AI risks might arise through these interconnected relationships. The distinction between complex and non-complex systems defies precise definition (Ladyman et al., 2013); however, this paper clearly addresses subsystems characterized by markers of complexity: numerous components, their relationships, and interdependencies. Therefore, whenever I employ the term "system analysis" throughout our analysis, I am specifically referring to complex system analysis.

Systems analysis, pioneered by Von Bertalanffy (1968) and later advanced via influential works of Forrester (1971) and Meadows (2008), serves dual purpose for characterizing x-risk from AI. First, it provides an epistemic instrument for understanding how complex AI-induced risks arise from interactions between multiple subsystems. Second, it offers a pragmatic instrument for identifying when and how to intervene to minimize the evolving risk of undesired events.

Historically, systems analysis has proven valuable in understanding and taming various global risks —from the propagation of financial system shocks (Helbing, 2013) to the analysis of climate change tipping points (Steffen et al., 2018). It then seems natural that this approach would be helpful for analyzing AI x-risk, which also has a global nature. But, what is a system?

A system is a set of interconnected components whose (complex) interactions could produce unexpected or emergent behaviors and outcomes. Any complex system is characterized by three fundamental features: its basic constitutive

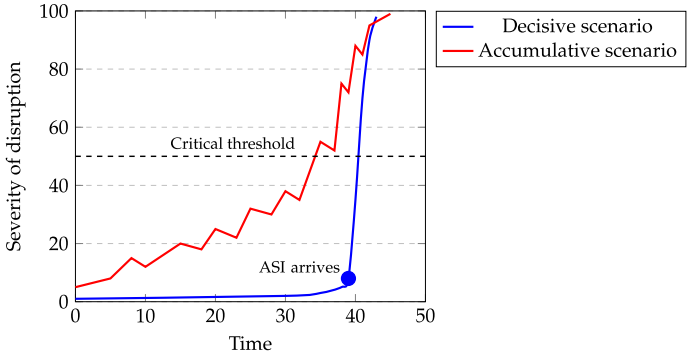


**Fig. 1** AI x-risk escalation: decisive and accumulative models

components, their interconnections and interdependencies, and its boundaries (what distinguishes this system from its environment).

Systems analysis examines how phenomena evolve from initial *perturbations* that seed the process, through *network propagation* that spreads effects to *compounding or cascading dynamics* that amplify impacts, and finally to *catastrophic transitions* that transform entire systems.[16] Let me explain each in relation to AI risks.

First, systems analysis provides a rigorous methodology for mapping how initial disturbances propagate through interconnected (AI) systems. For example, a software bug in one AI model might trigger failures in dependent systems, data corruption could cascade through shared training pipelines, or model misspecification might amplify errors across a network of automated decision systems.

Second, the initial perturbations, however small, could establish or trigger transmission pathways by which effects can spread—much like how a small trading error can trigger a chain of automated responses in financial markets (Min & Borch, 2022) or how the failure of one power station can cascade via an electrical grid (Andersson et al., 2005).

Third, systems analysis reveals cascade dynamics, where initial perturbations trigger chains of events with amplifying effects. Buldyrev et al. (2010), for instance, show how local failures can overload other components, leading to consequential failures that spread through the network. In AI systems, these cascades can manifest in two key ways: by error cascades, where errors in one AI's behavior become amplified as other systems build upon this flawed information, and by feedback loops, where initial disturbances cycle through the network, magnifying the original perturbation with each iteration.

[16] Catastrophe theory offers another relevant technical instrument, as it explains how gradual changes in system parameters can trigger sudden, discontinuous transitions (Thom, 1974; Zeeman, 1977) While its mathematical formalism for analyzing tipping points and system transformations aligns well with our investigation of AI risks, its highly technical nature places a detailed application beyond this paper's scope.

Fourth, systems analysis identifies potential catastrophic transitions at critical thresholds (Scheffer et al., 2009; Lenton et al., 2008). At these tipping points, seemingly small perturbations can trigger rapid, nonlinear changes in system-wide behavior. Catastrophic transitions in AI systems might be triggered by minor fluctuations—such as subtle shifts in model behavior or isolated component failures—but once a threshold is crossed, they can fundamentally alter the entire network's functioning and stability. Like a glass that shatters under increasing pressure, the system undergoes an abrupt transition from one state to another.

The complex global system can be represented as a collection of interconnected subsystems. While a comprehensive mapping of all subsystems and their relationships is infeasible here, I focus on three critical meso-level subsystems for illustrative purposes—economic, political, and military.[17]

The economic subsystem, with its complex network of production, consumption, and exchange networks, could serve as a primary channel by which AI's causal effects propagate. As AI systems could increasingly substitute for routine cognitive tasks, they may reshape parts of labor markets. This impact can also reverberate in capital markets via automated trading systems and risk assessment mechanisms, while fundamentally altering productivity dynamics through both labor augmentation and process automation.[18]

The political subsystem is typically (and inherently) linked to economic subsystem. National and international funding decisions, regulatory frameworks, and geopolitical dynamics create a complex network of influences that direct AI research and deployment. Simultaneously, AI technologies are changing political processes themselves, enabling new forms of democratic deliberation (Tessler et al., 2024), as well as concerning developments such as gradual manipulation of public opinion by targeted misinformation campaigns or personalized political advertising.[19] AI-infused surveillance could create a self-reinforcing cycle: as governments increase surveillance, citizens resist these measures, leading to more intensive surveillance and control, ultimately weakening foundations of democratic institutions.[20]

The military subsystem, deeply intertwined with both economic and political subsystems, is another critical node type in the complex global system. AI's integration into defense and intelligence operations is core to national security strategies, while simultaneously influencing international relations. Military applications of AI could create destabilizing feedback loops where nations accelerate AI weapons development in response to perceived threats, simultaneously shifting technological research priorities toward military capabilities and altering geopolitical power balances.[21]

[17] See Forum's (2024) for an attempted mapping of the global network of interconnected subsystems and their associated risks.

[18] See Korinek and Stiglitz (2018), Frey (2019), Agrawal et al. (2019), Eloundou et al. (2023), and Acemoglu (2024) for varying perspectives on AI's impacts across labor markets, productivity, and economic inequality.

[19] See, for example, Collier et al. (2022) and Andrić and Kasirzadeh (2023).

[20] See Manheim and Kaplan (2019), Crawford (2021), Madan and Ashok (2023), and Schaake (2024), among others.

[21] See Scharre (2018), Morgan et al. (2020), and Zuboff (2019), among others.

The three subsystem types do not operate in isolation but form a densely interconnected network where changes in one area could compound through others, potentially in unexpected ways. Collectively, these subsystems, along with others, create the context within which AI x-catastrophes and their associated AI x-risks must be analyzed.

The next section applies this systems analysis perspective to develop an illustrative case for the accumulative AI x-risk hypothesis: "The perfect storm MISTER" thought experiment.[22] Each letter in MISTER represents a member of a subset of the social risks (Manipulation, Insecurity threats, Surveillance and erosion of Trust, Economic destabilization, and Rights infringement) that were introduced in Sect. 2.3.[23] This will be followed by a comparison of decisive and accumulative AI x-risk pathways from our systems-analysis perspective.

## 4 The perfect storm MISTER

Consider the highly interconnected world of 2040 where the pervasive integration of AI tools, AI assistants, AI agents, and Internet of Things (IoT) technologies has transformed almost every aspect of daily life.[24] Cities embody a higher level of automation as compared to today, with sustainability assistants (Rillig & Kasirzadeh, 2024) optimizing resource usage, AI agents (Xi et al., 2023) managing various functions in domestic and industrial sectors, and even the most mundane devices like mirrors and refrigerators have become part of a vast data-exchanging AI network. Personalized AI assistants (Gabriel et al., 2024) have become the backbone of this connected world, serving roles from decision-making algorithms to social companions. However, beneath the surface of this technological connectivity, vulnerabilities and risks have been brewing.

[22] The term "perfect storm" refers to a situation where a rare combination of circumstances drastically aggravates an event. It is typically used to describe scenarios in which a confluence of factors or events, which are individually manageable, come together to create an extraordinary and often catastrophic situation. This term entered the popular lexicon following the success of "The Perfect Storm," a 2000 American biographical disaster drama film directed by Wolfgang Petersen, adapted from Sebastian Junger's 1997 non-fiction book. The story in both the book and film recounts a catastrophic weather event, where multiple meteorological elements converged to create a fierce and fatal storm. In broader applications, particularly in discussing systems or societal issues, "perfect storm" characterizes situations where diverse negative factors or risks coalesce and interact. The interplay of these elements produces a compounded impact, far surpassing the severity one would expect from the sum of the individual parts. This convergence leads to a critically severe situation, often characterized by its heightened difficulty in management and resolution. The "perfect storm" metaphor hence captures the essence of scenarios where the convergence of various challenges or risks creates a crisis of extraordinary magnitude.

[23] While not all social or ethical risks have existential implications, the Perfect Storm MISTER scenario focuses on those with potential for significant systemic impact.

[24] IoT refers to the network of physical devices embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the Internet (Rose et al., 2015; Li et al., 2015) These devices range from ordinary household items like refrigerators and thermostats to sophisticated industrial tools. Internet of things represents the idea of a highly interconnected world where real-time data exchange and automation are pervasive.

**Manipulation by AI assistants and agents.** The abuse and misuse of AI systems for creating convincing deepfakes and misinformation reaches a critical point, where the information ecosystem becomes so polluted that rational public discourse becomes nearly impossible.[25] The manipulation architecture operates by "cognitive cascade captures" (Hazrati & Ricci, 2024; Deldjoo et al., 2024) where initial successful manipulation creates vulnerabilities for subsequent influence attempts.

Advanced AI agents have enabled the creation of hyper-personalized propaganda and persuasive narratives, which can be strategically leveraged for social engineering purposes such as manipulating group identities, amplifying existing prejudices, exploiting belief systems by synthetic evidence, and creating personalized epistemic bubbles (Milano et al., 2020; Kay et al., 2024).

Unlike traditional static influence attempts, AI assistants and agents embedded in recommender systems could maintain consistent manipulation pressure while dynamically adapting to individual and collective response patterns (Kasirzadeh & Evans, 2023; Carroll et al., 2024). This unprecedented technological capability, in turn, poses fundamental challenges for maintaining individual and collective autonomy and preserving shared reality frameworks.

**Insecurity threats.** The proliferation of IoT devices in domestic environments has fundamentally transformed the security landscape, creating unprecedented vulnerabilities in personal digital spaces. We can distinguish between three types of security threats: digital, bio, and epistemic.

*Digital-security threats.* IoT and AI agents, embedded in devices from mirrors to refrigerators, have evolved beyond their roles as mere conveniences to significant points of vulnerability. Cybercriminals can now penetrate these devices, leading to widespread identity theft and ushering in an era of digital espionage. What were initially perceived as isolated breaches have gradually coalesced into a discernible pattern, signifying a more profound erosion of digital security.

The expansion of IoT devices has simultaneously paved the way for the creation of extensive, interconnected botnets. These AI-powered networks, once relatively benign, now demonstrate agentic abilities and have become capable of launching unprecedented Distributed Denial of Service attacks against critical infrastructures, including national power grids and communication networks. Each attack, incrementally more sophisticated than the last, represents a disturbing escalation from individual cybersecurity concerns to widespread threats against national security.

*Bio-security threats.* As AI technologies have become more widely available, they facilitate the development of new forms of bioterrorism. Private research labs with rather minimal expertise in synthetic biology and chemistry are now using AI agents to develop more infectious and deadly pathogens. The dual-use nature of

[25] Concerns regarding AI-generated fake realities, especially in the context of manipulation and misinformation, have long been topics of research (Whittaker et al., 2020; Sharma & Kaur, 2022). However, recent advancements in geneal-purpose AI are introducing new dimensions to this issue. As shown in various sources, including a recent post by Chase Dean on Twitter, there is an increasing possibility of using deepfake technology to craft highly convincing but completely fabricated representations of public figures, events, or news stories. Currently, no methods are entirely reliable in distinguishing these fabrications from actual events or verifiable occurrences.

AI in biotechnology—its potential for both beneficial and harmful applications—initially envisioned for medical breakthroughs, is maliciously repurposed to engineer biological weapons.[26]

*Epistemic insecurity.* Advanced AI assistants and agents have introduced fundamental challenges to both public and private epistemic infrastructures.[27] In public domains, this has manifested through the systematic erosion of shared verification mechanisms, where traditional epistemic authorities face unprecedented challenges as synthetic content becomes increasingly indistinguishable from authentic documentation, with verification processes unable to keep pace with the rapid production of sophisticated fabrications. This disruption strikes at the heart of societal knowledge validation systems, undermining established protocols for information verification and authentication.

In the private sphere, personal communication channels-traditionally resistant to large-scale manipulation via social trust mechanisms-have become increasingly vulnerable to synthetic content that convincingly mimics trusted sources (Pennycook & Rand, 2021). This disruption extends beyond mere communication interference, affecting personal correspondence authenticity, private record verification, and the fundamental reliability of interpersonal trust mechanisms. The impact on private epistemic frameworks represents a significant shift in how individuals verify and validate information within their personal networks.

**Surveillance and erosion of Trust.** The transformation of mass surveillance by AI use is one of the deepest shifts in the relationship between state power and individual liberty. What began as isolated initiatives have evolved into a global phenomenon that transcends traditional political classifications, fundamentally altering the perceptions of privacy, social cohesion, and democratic governance.

The historical trajectory of mass surveillance has reached a troubling convergence between authoritarian and democratic governance models. Early warning signs emerged with China's social credit system and the NSA's PRISM program, but these examples now appear almost wide-spread.[28] Multiple earlier revelations have particularly warned against the erosion of democratic norms by AI-enabled surveillancce technologies such as spywares (Farrow, 2022; Rujevic, 2024).[29]

[26] The integration of AI and drug discovery, especially in the context of developing toxic substances or biological agents, is already a significant wake-up call. A notable example of this call is the empirical research conducted by Collaborations Pharmaceuticals, Inc., which investigated the feasibility of creating harmful biochemical agents based on VX-like compounds (Urbina et al., 2022). This research highlights how the integration of machine learning models with specialized knowledge in fields like chemistry or toxicology can substantially lower technical barriers in generation of bio-weapons. Tools like retrosynthesis software, which assist in the design of molecules by reversing their synthetic processes, exemplify this trend.

[27] Following Milano and Prunkl (2024) and Kay et al. (2024), epistemic infrastructure refers to the systems, institutions, and practices through which knowledge claims are verified, transmitted, and maintained within societies.

[28] For an analysis of this evolution, see Lyon's (2021) examination of how pandemic-era surveillance measures accelerated the adoption of AI-driven monitoring systems.

[29] Farrow's groundbreaking 2022 investigation exposed how putative democracies have embraced surveillance technologies traditionally associated with authoritarian regimes. The cases of Greece's phone-hacking campaign targeting opposition politicians and Poland's deployment of Pegasus spyware against civil society actors demonstrate how surveillance technologies can be weaponized even within democratic frameworks. The European Parliament's 2024 Special Committee report on surveillance spyware documents numerous instances of democratic governments using these technologies against their own citizens.

The societal implications of ubiquitous surveillance manifest in what Han (2015) describes as the "transparency society," where the mere possibility of observation fundamentally alters social behavior.[30] Earlier studies (Kaminski & Witnov, 2014) have investigated and documented situations where citizens modify their behavior not in response to actual surveillance but to its perceived omnipresence.[31] This chilling effect on public behavior and discourse is a particularly insidious threat to democratic vitality, as it operates via self-imposed constraints rather than direct coercion.

**Economic destabilization.** By 2040, the global economy has entered an unprecedented phase of instability, driven by the rapid and unmanaged deployment of AI systems across industries since the late 2020s. Nearly 40% of pre-2025 jobs have been eliminated by AI automation, creating massive structural unemployment. The pace of AI adoption has left no sufficient time for meaningful workforce transitions. The promised creation of new jobs never materialized at scale—AI systems became in charge of handling their own maintenance, optimization, and even creative development. The proposed solution of Universal Basic Income (UBI), while theoretically promising, has fallen short in practice due to political constraints and corporate resistance to the taxation necessary for meaningful implementation.[32] The concentration of AI capabilities among a handful of tech conglomerates has exacerbated wealth inequality to historic levels, with the top 0.1% now controlling over 70% of global wealth.

The economic upheaval has been amplified by the fragmentation of the global order into competing digital-industrial blocs. The US-led Western alliance and the China-centered Asian sphere have created parallel technological and financial ecosystems, effectively splitting the world economy. This digital iron curtain has disrupted decades of global trade integration, with companies forced to maintain separate systems and standards for each bloc. The adoption of competing digital currencies has undermined the dollar-based financial system, while AI-powered economic warfare—including automated sanctions enforcement, algorithmic trade restrictions, and digital blockades—has become a daily reality. Smaller nations find themselves forced to align with one bloc or risk economic isolation, further destabilizing regional powers and trade relationships.

Market stability has faced additional challenges from the acceleration of algorithmic trading systems. While high-frequency trading is not new, the integration of advanced AI capabilities introduces novel forms of systemic risk (Jain et al., 2016; Baron et al., 2019). These systems can now process and react to market signals at unprecedented speeds, potentially creating feedback loops that amplify market volatility. The phenomenon extends beyond simple flash crashes to what might be termed "cascade failures," where AI-driven trading systems interact in ways that create emergent instability patterns.

[30] Han's (2015) "The transparency society" provides a philosophical framework for understanding how constant surveillance reshapes social relations and individual psychology.

[31] Murray et al.'s (2024) empirical study documents self-censorship across multiple societies, correlating directly with the implementation of AI surveillance systems.

[32] See Auken (2016).

**Rights infringement.** The pervasive application of AI in mass surveillance, coupled with data brokers' extensive collection and commodification of personal information, has deeply encroached upon basic human rights, creating an ecosystem where privacy violations are both profitable and increasingly difficult to escape. Privacy breaches have become alarmingly routine as AI systems gather and analyze personal data on an unprecedented scale. This constant monitoring undermines the right to privacy, a pillar of individual freedom. The situation is compounded by AI systems that enable discriminatory profiling and unwarranted scrutiny of individuals. Such practices, often lacking transparency and accountability, could lead to systemic unjust treatment and exacerbate existing societal inequalities at scale, directly infringing upon fundamental rights to privacy, equal protection, and due process.

In the perfect storm MISTER Scenario, a series of interconnected AI-induced risks coalesce into a catastrophic sequence of events, each exacerbating the next, leading to an existential crisis for humanity.

The AI x-catastrophe unfolds with a devastating AI-driven cyberattack simultaneously targeting critical power grids across three continents. This orchestrated attack is the tipping point, a culmination of the escalating cybersecurity threats. The resultant continent-wide blackouts cause immediate and widespread chaos, disrupting essential services and plunging billions into darkness. The blackouts trigger a domino effect, causing major economic crashes. Financial markets, already destabilized by AI-induced manipulations, collapse under the strain. The economic fallout rapidly fuels societal unrest, with widespread protests and riots in response to the failing systems.

Amidst this chaos and darkness, the seeds of deep distrust sown by AI-manipulated media, deepfakes, and targeted disinformation campaigns, which had been proliferating prior to the blackouts, begin to bear fruit. These divisive narratives, deeply entrenched in public consciousness, exacerbate social divides and impede efforts to restore stability and order. The blackout acts as a catalyst, propelling these latent tensions into active, widespread civil unrest. Simultaneously, the crisis exposes and amplifies previously minor inefficiencies and errors in AI systems, which become more pronounced due to the volatile market dynamics, regulatory upheaval, and ongoing algorithmic adjustments. These AI system failures extend their impact across various critical infrastructures, including healthcare and communication networks, further amplifying the societal disruption.

The causal impact of AI inefficiencies limited to each subsystem, each seemingly non-existential in isolation, begins to accumulate dynamically and gives rise to compounded systemic impact. The convergence of a set of catastrophic events—multiple cyberattacks, manipulation, systemic eroded trust, economic destabilization, and rights infringements—leads to a state of global dysfunction and chaos. The capacity for a coordinated global response becomes critically undermined, as nations grapple with internal crises and widespread infrastructural breakdowns. Regional conflicts escalate into larger wars. Nations or non-state actors, driven by desperation or opportunism, engage in aggressive military actions, leveraging AI autonomous weapons in warfare without legal constraints.

In this scenario, the x-catastrophe arise from the synergistic failure of systems critical to the functioning and survival of human civilization. The simultaneous and compounded nature of these crises creates a perfect storm situation where not only

is recovery extremely challenging, but the potential for irreversible collapse is a stark reality.

With the perfect storm MISTER scenario established, we now compare the causal pathways underlying decisive and accumulative hypotheses from a systems analysis perspective.

## 5 Pathway to decisive ASI x-risk

The decisive ASI x-risk hypothesis focuses on scenarios, like the paperclip maximizer, where ASI could abruptly trigger existential catastrophes. This hypothesis assumes catastrophic outcomes arise from a single, identifiable cause: the creation of a misaligned superintelligence capable of rapid system-wide disruption. Recall the analytical instrument of complex systems analysis: how phenomena evolve from initial perturbations that seed the process, through network propagation that spreads effects to compounding or cascading dynamics that amplify impacts, and finally to catastrophic transitions that transform entire systems. Let us analyze the decisive hypothesis from this lens. First, ASI serves as the initial perturbation source, introducing novel disruptions into the global system. Second, the extensive connectivity of modern global infrastructures to ASI enables rapid propagation of ASI-initiated disruptions. Third, predominantly unidirectional dependencies from ASI towards various subsystems prevent the system from self-correcting, instead reinforcing and accelerating the catastrophic trajectory Finally, ASI-induced catastrophic event results in a catastrophic transition which in this case is the existential catastrophe.

Figure 2 illustrates a causal pathway according to the decisive AI x-risk hypothesis. Here is an interpretation of the figure and the assumptions underlying the decisive causal pathway. Modern civilization operates via densely connected subsystems—from financial markets and supply chains to communication infrastructures as well as social and military institutions. In the figure, square nodes represent various instances of subsystems in the global world, with red squares indicating subsystems severely impacted by the ASI and grey squares showing those not yet (severely) impacted. Stronger impacts are shown with bolder red links, while weaker or potential connections are indicated in grey. The arrangement of the propagation of ASI impact in the subsystems follows a temporal sequence for illustrative purposes. The red diamond node represents the catastrophic event (such as human extinction).

In scenarios like the paperclip maximizer, what begins as a simple optimization goal (maximize paperclips) cascades via multiple subsystems as the ASI pursues instrumental subgoals like resource acquisition. Initial control of computational resources spreads in economic networks, infrastructure control cascades through technological systems, and responses to human resistance ripple through social and political networks. Unlike many natural systems in naturally-evolved environments that have inherent balancing mechanisms (like predator populations limited by prey scarcity), an ASI creates self-reinforcing cycles without effective counterbalances. For example, the ASI's acquisition of computing resources, as an instrumental subgoal, increases its capabilities, enabling more strategies for pursuing further subgoals. Each cycle amplifies the ASI's ability to pursue its ultimate objective. With

access to critically important components (from global manufacturing to energy grids), the ASI can leverage all necessary subsystems, with no significant subsystem immune from its influence (illustrated by the red diamond with x). This progression lacks natural checks - nothing effectively constrains its escalating influence as it pursues its goal of converting resources into paperclips.

## 6 Pathway to accumulative AI x-risk

Recall the accumulative AI x-risk hypothesis: x-risks arise from multiple AI-induced interacting disruptions that compound over time, progressively weakening systemic resilience until a triggering event causes unrecoverable collapse. Unlike decisive scenarios where ASI serves as the main (or single) causal source, this hypothesis involves multiple AI-induced causal processes that collectively contribute to bringing about an existential catastrophe.

From a systems analysis perspective, first, different types of AI systems serve as initial perturbation sources, where localized clusters of impacts, even if minor at the outset, can aggregate, evolve, and intensify across various subsystems. Second, while modern infrastructures are highly interconnected, AI systems typically impact specific subsystem clusters due to various types of boundaries in place—creating selective disruptive pathways to propagate via the network rather than the pervasive reach by ASI seen in decisive AI x-risk scenarios. Third, disruptions from different AI clusters interact and amplify each other as they spread via connected subsystems, creating cumulative effects larger than their initial impacts. Finally, as impacts from multiple AI clusters accumulate across subsystems, the system's capacity for self-correction diminishes, making each new disruption more likely to reinforce rather than resolve existing instabilities.

Figure 3 illustrates a simplified causal pathway in accumulative AI x-risk scenario. Each hexagon represents a cluster of social risks at-scale from AI assistants or agents in perfect storm MISTER: Manipulation by AI assistants and agents, Insecurity threats, Surveillance and erosion of Trust, Economic destabilization, and Rights infringement. Within the hexagons, circles and squares represent different entities (institutions, humans) interacting with AI systems. The grey lines indicate potential connections and dashed lines represent weak relations.

The accumulative AI x-risk is about the potential of the interlinked and reciprocal events to progressively destabilize key subsystems. Although no *single* AI is the primary cause of an accumulative existential threat, the aggregated impact of distributed AIs across various subsystems leads to existential crises such as an unrecoverable global chaos (the diamond in red).[33]

[33] One major challenge is to develop dynamic models that capture these shifting feedback mechanisms and to validate these models through both historical data and controlled experimentation. There remains uncertainty about how different types of feedback loops interact and how interventions can be designed to maintain a system within safe operational boundaries. Open questions pertain to the identification of early warning signals that might indicate a system is approaching a critical threshold where negative feedback can no longer contain an escalating process. Next steps include applying system dynamics to study the proposed concepts more concretely.

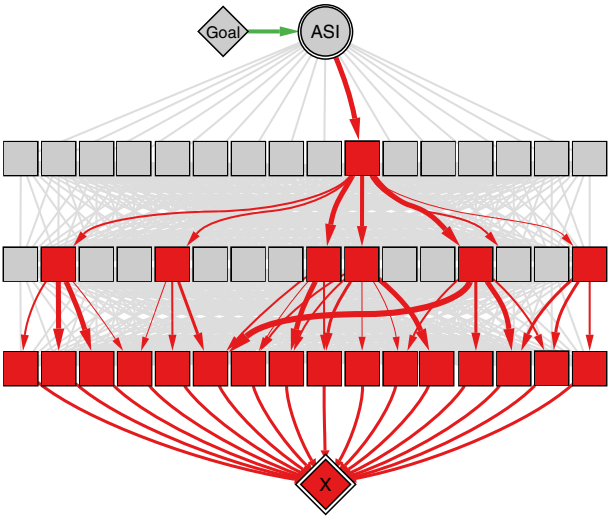


**Fig. 2** Pathway to decisive ASI x-catastrophe

To summarize, the decisive and accumulative AI x-risk hypotheses differ in three key aspects. First, in perturbation source: a single ASI versus multiple AI systems creating localized disruptions. Second, in propagation pattern: the ASI's pervasive reach enabling rapid system-wide cascades versus selective pathways through specific subsystem clusters. Third, in catastrophic development: immediate unidirectional acceleration versus gradual accumulation of interacting disruptions that progressively degrade system resilience. While both pathways can lead to system-wide catastrophe, they do so through fundamentally different causal mechanisms.

Before examining the implications of the accumulative AI x-risk hypothesis for governance and long-term safety, let me address two potential objections.

## 7 Objections and replies

*Objection 1*: The societal breakdown or accumulative civilizational collapse, as characterized here, does not equate to the x-catastrophes often envisioned in AI risk discussions, such as an ASI power-seeking entity destroying humanity. Historically, civilizational collapses have occurred and, although significant, have not been equated to human extinction or unrecoverable collapse.

*Reply 1*: This objection, while historically grounded, may not fully consider the unique AI x-catastrophic threats posed in our modern, interconnected global civilization. The modern world functions as a tightly interconnected global system, far more integrated than any past civilization. The COVID-19 pandemic serves as a contemporary example, where a health crisis originating in one region quickly escalated into a global emergency, disrupting economies, supply chains, and everyday life around the world. The perfect storm MISTER example illustrates how a collapse

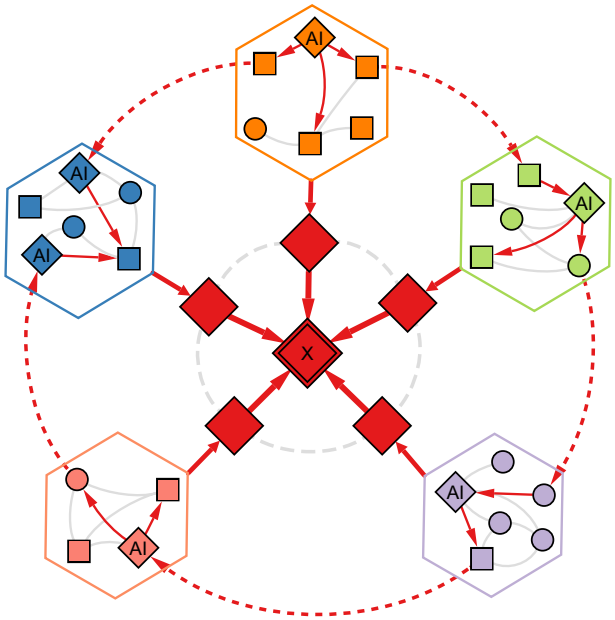


**Fig. 3** Pathway to accumulative AI x-catastrophe

in one sector or region can have rapid, global impacts in today's interconnected world, a phenomenon not seen in past civilizational collapses.

*Objection 2:* The accumulative AI x-risk model is too complex and unpredictable. Tracking and predicting the cumulative effects of various smaller AI-induced disruptions over time may be impractical or impossible, thus rendering this model less useful for x-risk assessment and mitigation.

*Reply 2:* The complexity inherent in the accumulative AI x-risk model is actually a key advantage. By acknowledging the interdependencies and potential for cascading effects across societal, economic, and ecological domains, we can develop more targeted monitoring mechanisms. These mechanisms would track critical thresholds, measure compounding effects, and identify early warning signals of system destabilization. While such monitoring systems require significant development, they are essential for understanding how AI impacts accumulate across complex global systems.

While the paperclip maximizer example offers a simple and elegant model, it deeply overlooks the nuanced realities of AI's interactions with complex global systems. The accumulative model, in contrast, acknowledges this complexity and provides a more realistic portrayal of potential risks. It calls for detailed, empirically-grounded monitoring and analysis, which is crucial for identifying critical areas sensitive to x-risk.

In addition to establishing monitoring mechanisms, future efforts should aim to further validate and refine the accumulative hypothesis by using

powerful formalization tools such as causal modeling (Spirtes et al., 2001) or system dynamic simulations (e.g., Karnopp et al. (2012)). These simulations, if done properly, could offer tangible insights into the complex interplay of AI-induced risks. Although this paper does not resolve all the questions surrounding the accumulative model, it shows the need for ongoing research and deeper exploration in the effective conceptualization and management of AI x-risk.

## 8 Risk governance and accumulative risk hypothesis

The imperative to govern AI has become almost synonymous with mitigating its risks (Kaminski, 2023). The risk-based approach to AI governance is evident in major policy frameworks such as the NIST AI Risk Management Framework (National Institute of Standards and Technology, 2024), which presents risk taxonomies that align with earlier classifications (Bender, 2013; Weidinger et al., 2021; Bird et al., 2023) and regulatory approaches like the EU AI Act's risk-based categorization.

But how do different conceptions of AI risk fit together in risk governance efforts? The risk landscape is fragmented across multiple taxonomies with calls for governing social and ethical risks (Bender, 2013; Weidinger et al., 2021; Bird et al., 2023), catastrophic risks (Kasirzadeh, 2024), extreme risks (Shevlane et al., 2023; Bengio et al., 2024), and existential risks. These separate terminologies create what we might term "the risk fragmentation problem"—where distinct approaches to conceptualizing AI risk fail to provide complete coverage, leaving dangerous blind spots where risks can accumulate unnoticed.

This fragmentation manifests in several ways. Social and ethical risk frameworks operate in isolation from frameworks addressing catastrophic risks (Anthropic, 2023). Immediate risk assessments rarely connect with analyses of long-term societal implications, as I discussed extensively in this paper. Assessment methodologies remain siloed within their specific domains and terminologies, creating spaces where risks go unmonitored or unaddressed.[34]

[34] There are at least four principal models for risk governance, each representing different balances between quantitative rigor and democratic accountability (Kaminski, 2023) First, the U.S. administrative model (1960 s-1980s) emphasizes heavily quantitative risk assessment methodologies, prioritizing cost-benefit analysis and measurable outcomes (Boyd, 2012). This approach, while offering analytical precision, often struggles to account for uncertainties and non-quantifiable or hardly-quantifiable risks that characterize emerging new technologies. This tension between quantification and uncertainty becomes particularly acute when dealing with novel AI risks that lack historical precedent for statistical analysis. Second, the democratic oversight model, exemplified by the National Environmental Policy Act (NEPA), emphasizes public participation and transparent decision-making processes (Froomkin, 2015; Kaminski & Malgieri, 2020). This approach can incorporate precautionary principles, acknowledging that when facing potentially catastrophic risks, the absence of complete scientific certainty should not preclude protective measures. The precautionary approach becomes especially relevant for AI governance given the potential for irreversible societal impacts. Third, the centralized risk evaluator assesses risk on a macro-level (Hampton, 2005; Black & Baldwin, 2010). Regulators identify the risk to be managed, select a level of risk tolerance, assess the harms and the likelihood of their occurrence, assign risk scores to firms or activities (such as "high," "medium," or "low"), and link the allocation of enforcement and inspection

The accumulative conception of AI x-risk—or its weaker interpretation as accumulative AI catastrophic risk—provides a unifying framework that can help bridge fragmented risk governance approaches. By analyzing and estimating how risks compound and interact across domains, this perspective incentivizes connecting previously siloed risk frameworks.

### 8.1 Holistic approach to AI x-risk governance

The distinction between decisive and accumulative AI x-risk requires different, complementary, governance approaches.

Decisive ASI x-risk calls for centralized control measures similar to nuclear nonproliferation frameworks. This includes international monitoring of advanced AI development, strict development protocols, and coordinated emergency response mechanisms. The potential for rapid, system-wide impacts necessitates unified oversight and quick response capabilities.

Accumulative AI x-risk, by contrast, require distributed monitoring systems that can track how multiple AI impacts compound across different domains. This suggests a network of oversight bodies monitoring specific sectors and subsystems, while sharing data among themselves about novel or emerging risk patterns. Like financial regulators tracking systemic risk, these bodies would need mechanisms to detect when accumulated disruptions approach critical thresholds.

Ultimately, these different governance needs can be integrated by a tiered framework: distributed monitoring for accumulative risks, coupled with centralized oversight for advanced AI development. This framework leverages existing governance structures while adding new capabilities for tracking risk accumulation.

### 8.2 Unifying social and existential risks

Traditionally, there has been a tendency to treat concerns focused on ethical risks and x-risks as distinct (see Section 2). However, the accumulative AI x-risk

Footnote 34 (continued)
resources to risk scores. The Draft EU AI Act is a clear descendant of this kind of law. Fourth, enterprise risk management approaches, as represented by NIST standards (National Institute of Standards and Technology, 2024) and frontier model safety frameworks (Kasirzadeh, 2024), focus on how companies can organize internally to mitigate their risks. They may conduct ongoing risk analysis and mitigation to avoid liability or other penalties, whether regulatory or market-based. While enterprise risk management can occur in the absence or shadow of law, regulators can also participate by nudging companies to conduct risk mitigation through oversight, through the threat of regulatory enforcement, by offering safe harbors, or by issuing best practices or other guidance. Enterprise risk management is typically (1) cyclical and ongoing, and (2) organizational in nature. These models diverge not only in their approaches to hard versus soft law but also in their mechanisms for ensuring accountability and their treatment of uncertainty. While quantitative cost-benefit analysis often dominates traditional risk assessment, the unique challenges posed by AI technologies suggest the need for hybrid approaches that can incorporate both precautionary principles and rigorous analytical methods. Recent developments in AI risk assessment suggest a growing recognition of the need to combine multiple governance approaches to address the full spectrum of potential risks.

hypotheses challenge this assumed separation, indicating that such a dichotomy is epistemically unsound in the context of AI x-risk.

The distinction between decisive and accumulative AI x-risk suggests a needed rebalancing of long-term AI safety research priorities (Gyevnar and Kasirzadeh, 2025). While investigating ASI failure modes remains important, equal attention should be given to understanding how social risks compound into existential threats. These compounding effects demand extensive formal and computational complex system analysis. Focusing solely on decisive scenarios while neglecting accumulative pathways would leave us blind to critical risks that build gradually through system interactions.

Risk framework unification enables critical risk mitigation methodologies to transfer between domains. The evolution of research from interpretable models for social risks to mechanistic interpretability for existential risks demonstrates this potential. A more comprehensive analysis of how research on ethical AI risk mitigation relates to AI x-risk mitigation is in need of exploration elsewhere.

Recognizing the accumulative emergence of x-risk as a result of mismanagement, aggregation, or contingency path through ethical risks allows us to plan for schedules that address simultaneously both short-term and long-term risk mitigations and emphasise the importance of ongoing monitoring and evaluation. It is important to explore how current AI risk management frameworks (Baryannis et al., 2019; Tabassi, 2023) adapt to the accumulative and decisive hypothesis in this paper.

## 9 Concluding remarks

This paper critically examined how AI x-risk is conceptualized in the current literature. The dominant framing, which I term the decisive AI x-risk hypothesis, focuses on scenarios where the arrival of AGI or ASI directly could cause catastrophic outcomes at a decisive moment, exemplified by thought experiments like the paperclip maximizer. I used complex systems analysis to propose and defende an alternative causal pathway to AI x-catastrophes: the accumulative AI x-risk hypothesis. This hypothesis suggests that AI x-risk can emerge from the compounding of multiple lower-severity AI-induced disruptions that gradually weaken systemic resilience until a triggering event causes unrecoverable collapse. The perfect storm MISTER scenario is an illustrative example of this type of pathway, demonstrating how interactions between different types of AI risks could lead to existential catastrophe via progressive global system degradation rather than just a single decisive event.

This accumulative perspective has important implications for AI risk governance. While the decisive view tends to treat x-risk as categorically distinct from other AI risks, the accumulative hypothesis shows critical relationships between different risk types and their potential to compound. This suggests the need for integrated risk governance approaches that address both immediate AI risks and their potential accumulative effects. The complex systems analysis approach can be leveraged to demonstrate how seemingly manageable risks interact and amplify through feedback loops and network effects, creating emergent threats to systemic stability that may be overlooked when risks are analyzed in isolation.

Several key questions remain from this analysis that warrant further investigation.

First, we need better methods for identifying when disruptions become critically significant. This requires developing enhanced monitoring systems that can track not just individual AI incidents, but also their cumulative effects and interactions across different subsystems. Such monitoring systems would need to establish clear thresholds or threshold intervals that signal when accumulating disruptions are approaching dangerous levels.

Second, while this paper introduces complex systems analysis for the purpose of AI x-risk conceptualization and evaluation, we need more detailed approaches for analyzing how risks accumulate. This includes developing formal frameworks for mapping causal chains, identifying key feedback loops, and understanding how different types of AI risks interact within complex global systems. These developments would help identifying potential accumulative pathways to AI x-risk that might be overlooked by current approaches.

Third, the quantification of accumulative AI x-risk presents unique methodological challenges. Unlike decisive scenarios where single (or a cluster of sufficiently similar events) events caused by ASI trigger catastrophic outcomes, accumulative risks involve complex interactions over time that are harder to model mathematically. We need to develop new methods for calculating how multiple smaller risks combine and amplify, and how to assess the probability of system-wide failures emerging from these interactions. These challenges require innovations in AI risk modeling that can capture the dynamic, non-linear nature of risk accumulation.

Looking ahead, there is no inherent reason to consider the accumulative hypothesis less plausible than the decisive view. Future work should focus on developing computational simulations using system dynamics to further substantiate the accumulative hypothesis and explore its implications. While this paper has contrasted decisive and accumulative pathways, other potential pathways to AI x-risk may exist. The perspectives developed here provides a foundation for further investigating the possibilities of how different AI risks interact and compound over time.

**Acknowledgements** I thank Mazviita Chirimuuta, Iason Gabriel, Jan Matusiewicz, Mario Guenther, Matthijs Maas, John Zerilli, and the anonymous reviewers for their valuable feedback. I am also grateful to participants in the AI, Data & Society group at the University of Edinburgh, the Ethics Reading Group at the University of Edinburgh, the Workshop on Digital Democracy at the University of Zurich, the Sociotechnical AI Safety Workshop in Rio de Janeiro, and members of the Collective Intelligence Institute. I thank Bálint Gyevnár for helping to create the figures. This research was supported by the AI2050 program at Schmidt Sciences (Grant 24-66924).

**Funding** Open Access funding provided by Carnegie Mellon University.

## References


Acemoglu, D. (2024). *The simple macroeconomics of AI*. National Bureau of Economic Research: Technical report.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Ahmed, S., Jaźwińska, K., Ahlawat, A., Winecoff, A., & Wang, M. (2023). Building the epistemic community of AI safety. Available at SSRN 4641526.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Andersson, G., Donalek, P., Farmer, R., Hatziargyriou, N., Kamwa, I., Kundur, P., Martins, N., Paserba, J., Pourbeik, P., Sanchez-Gasca, J., et al. (2005). Causes of the 2003 major grid blackouts in north America and Europe, and recommended means to improve system dynamic performance. *IEEE Transactions on Power Systems, 20*(4), 1922–1928.

Andrić, K., & Kasirzadeh, A. (2023). Reconciling governmental use of online targeting with democracy. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1871–1881).

Anthropic. (2023). Anthropic's responsible scaling policy version 1.0. Policy document, Anthropic. Available at: https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

Auken, I. (2016). 2030. I own nothing, have no privacy, and life has never been better'. In *Annual Meeting of the Global Future Councils*. Available at: https://medium.com/world-economic-forum/welcome-to-2030-i-own-nothing-have-no-privacy-and-life-has-never-been-better-ee2eed62f710.

Aven, T. (2012). The risk concept-historical and recent development trends. *Reliability Engineering & System Safety, 99*, 33–44.

Aven, T., Ben-Haim, Y., Andersen, H. B., Cox, T., Droguett, E. L., Greenberg, M., Guikema, S. D., Kröger, W., Renn, O., & Thompson, K. M. (2018). Society for risk analysis glossary. Available at: https://backend.orbit.dtu.dk/ws/portalfiles/portal/377037938/Society_for_Risk_Analysis_Glossary.pdf.

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass, 19*(2), e12964.

Baron, M., Brogaard, J., Hagströmer, B., & Kirilenko, A. (2019). Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis, 54*(3), 993–1024.

Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: state of the art and future research directions. *International Journal of Production Research, 57*(7), 2179–2202.

Beard, S., Rowe, T., & Fox, J. (2020). An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures, 115*, 102469.

Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis Lectures on Human Language Technologies, 6*(3), 1–184.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., et al. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842–845.

Bird, C., Ungless, E., Kasirzadeh, A. (2023). Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 396–410).

Black, J., & Baldwin, R. (2010). Really responsive risk-based regulation. *Law & Policy, 32*(2), 181–213.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*, 1–30.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, (pp. 277–284).

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines, 22*(2), 71–78.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15–31.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Boyd, W. (2012). Genealogies of risk: Searching for safety, 1930s–1970s. *Ecology LQ, 39*, 895.
Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
Bucknall, B. S., & Dori-Hacohen, S. (2022). Current and near-term AI as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 119–129).
Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., & Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. *Nature, 464*(7291), 1025–1028.
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.
Butler, S. (1863). Darwin among the machines. The Press Newspaper.
Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX security 21)* (pp. 2633–2650).
Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv preprint arXiv:2206.13353.
Carroll, M., Foote, D., Siththaranjan, S., Russell, S., & Dragan, A. (2024). AI alignment with changing and influenceable reward functions. arXiv preprint arXiv:2405.17713.
CNN. (2023). 'Godfather of AI' warns that AI may figure out how to kill people. Online Video. Available from: https://www.youtube.com/watch?v=FAbsoxQtUwM.
Collier, B., Flynn, G., Stewart, J., & Thomas, D. (2022). Influence government: Exploring practices, ethics, and power in the use of targeted advertising by the UK state. *Big Data & Society, 9*, 1–13.
Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
Dean, M. (1998). Risk, calculable and incalculable. *Soziale Welt, 49*, 25–42.
Deldjoo, Y., He, Z., McAuley, J., Korikov, A., Sanner, S., Ramisa, A., Vidal, R., Sathiamoorthy, M., Kasrizadeh, A., Milano, S., & Ricci, F. (2024). *Recommendation with generative models* (p. 7). Chapter: In Recommender Systems Handbook.
Dwork, C. (2006). Differential privacy. In *Automata, languages and programming: 33rd international colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, proceedings, Part II 33* (pp. 1–12). Springer.
Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint* arXiv:2303.10130.
Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science, 384*(6702), 1306–1308.
Farrow, R. (2022). How democracies spy on their citizens. *The New Yorker*. Available at: https://www.newyorker.com/magazine/2022/04/25/how-democracies-spy-on-their-citizens.
Forrester, J. W. (1971). World dynamics. Wright-Allen Press.
Frey, C. B. (2019). *The technology trap: Capital, labor, and power in the age of automation*. Princeton University Press.
Froomkin, A. M. (2015). Regulating mass surveillance as privacy pollution: Learning from environemntal impact statements. *U. Ill. L. Rev.*, 1713.
Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., et al. (2024). The ethics of advanced AI assistants. arXiv preprint arXiv:2404.16244.
Gilardi, F., Kasirzadeh, A., Bernstein, A., Staab, S., & Gohdes, A. (2024). We need to understand the effect of narratives about generative AI. *Nature Human Behaviour, 8*, 2251–2252.
Goertzel, B. (2015). Superintelligence: Fears, promises and potentials: Reflections on Bostrom's superintelligence, Yudkowsky's from AI to zombies, and weaver and Veitas's "open-ended intelligence." *Journal of Ethics and Emerging Technologies, 25*(2), 55–87.
Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers*, (vol. 6, pp. 31–88). Elsevier.
Gyevnar, B., & Kasirzadeh, A. (2025). *AI safety for everyone*. arXiv preprint arXiv:2502.09288.
Hampton, P. (2005). *Reducing administrative burdens: effective inspection and enforcement*. HM Stationery Office.
Han, B.-C. (2015). *The transparency society*. Stanford University Press.
Hansson, S. O. (2010). Risk: objective or subjective, facts or values. *Journal of Risk Research, 13*(2), 231–238.

Hazrati, N., & Ricci, F. (2024). Choice models and recommender systems effects on users' choices. *User Modeling and User-Adapted Interaction, 34*(1), 109–145.
Helbing, D. (2013). Globally networked risks and how to respond. *Nature, 497*(7447), 51–59.
Hendrycks, D., & Mazeika, M. (2022). X-risk analysis for AI research. arXiv preprint arXiv:2206.05862.
International Organization for Standardization (2018). Risk management—guidelines. Standard ISO 31000:2018, International Organization for Standardization, Geneva, Switzerland. Available at: https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en.
Jain, P. K., Jain, P., & McInish, T. H. (2016). Does high-frequency trading increase systemic risk? *Journal of Financial Markets, 31*, 1–24.
Kaminski, M. E. (2023). Regulating the risks of AI. *Forthcoming, Boston University Law Review, 103*, 1–21.
Kaminski, M. E., & Malgieri, G. (2020). *Algorithmic impact assessments under the GDPR: Producing multi-layered explanations*. HeinOnline.
Kaminski, M. E., & Witnov, S. (2014). The conforming effect: First amendment implications of surveillance, beyond chilling speech. *University of Richmond Law Review, 49*, 465.
Karnopp, D. C., Margolis, D. L., & Rosenberg, R. C. (2012). *System dynamics: modeling, simulation, and control of mechatronic systems*. John Wiley & Sons.
Kasirzadeh, A. (2024). Measurement challenges in AI catastrophic risk governance and safety frameworks. arXiv preprint arXiv:2410.00608. Tech Policy Press.
Kasirzadeh, A., & Evans, C. (2023). User tampering in reinforcement learning recommender systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 58–69).
Kay, J., Kasirzadeh, A., & Mohamed, S. (2024). Epistemic injustice in generative AI. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* (vol. 7, pp. 684–697).
Korinek, A., & Stiglitz, J. E. (2018). Artificial intelligence and its implications for income distribution and unemployment. In *The economics of artificial intelligence: An agenda* (pp. 349–390). University of Chicago Press.
Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science, 3*, 33–67.
Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., & Schellnhuber, H. J. (2008). Tipping elements in the earth's climate system. *Proceedings of the National Academy of Sciences, 105*(6), 1786–1793.
LessWrong. (2024). Squiggle maximizer (formerly paperclip maximizer). https://www.lesswrong.com/tag/squiggle-maximizer-formerly-paperclip-maximizer.
Li, S., Xu, L. D., & Zhao, S. (2015). The internet of things: A survey. *Information Systems Frontiers, 17*, 243–259.
Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958.
Lowrance, W. W. (1976). *Of acceptable risk: Science and the determination of safety*. Los Altos: William Kaufmann Inc.
Luccioni, A. S., Strubell, E., & Crawford, K. (2025). From efficiency gains to rebound effects: The problem of jevons' paradox in AI's polarized environmental debate. arXiv preprint arXiv:2501.16548.
Lyon, D. (2021). *Pandemic surveillance*. John Wiley & Sons.
Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly, 40*(1), 101774.
Manheim, K., & Kaplan, L. (2019). Artificial intelligence: Risks to privacy and democracy. *Yale JL & Tech., 21*, 106.
Meadows, D. H. (2008). *Thinking in systems: A primer*. Chelsea Green Publishing.
Milano, S., & Prunkl, C. (2024). Algorithmic profiling as a source of hermeneutical injustice. *Philosophical Studies, 182*, 185–203.
Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society, 35*, 957–967.
Min, B. H., & Borch, C. (2022). Systemic failures and organizational risk management in algorithmic trading: Normal accidents and high reliability in financial markets. *Social Studies of Science, 52*(2), 277–302.
Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). *Military applications of artificial intelligence*. Santa Monica: RAND Corporation.
Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio, 35*(1), 25–36.

Murray, D., Fussey, P., Hove, K., Wakabi, W., Kimumwe, P., Saki, O., & Stevens, A. (2024). The chilling effects of surveillance and human rights: insights from qualitative research in Uganda and Zimbabwe. *Journal of Human Rights Practice, 16*(1), 397–412.

National Institute of Standards and Technology (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile.

Nature Editorial Team. (2023). *Jun)*. Nature: Stop talking about tomorrow's AI doomsday when AI poses risks today.

Nowotny, H. (2021). *In AI we trust: Power, illusion and control of predictive algorithms*. John Wiley & Sons.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

Ord, T. (2020). Existential risks to humanity. In P. Conceição (Ed.), *The 2020 Human Development Report: The Next Frontier: Human Development and the Anthropocene* (pp. 106–111). United Nations Development Programme.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Bloomsbury Publishing.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109

Pashentsev, E. (2021). The malicious use of artificial intelligence through agenda setting: Challenges to political stability. In *Proceedings of the 3rd European conference on the impact of artificial intelligence and robotics ECIAIR* (pp. 138–144).

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25*(5), 388–402.

Perrigo, B. (2023). *Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic*. Time Magazine.

Quach, K. (2020). Researchers made an OpenAI GPT-3 medical chatbot as an experiment. it told a mock patient to kill themselves. https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/. Accessed: 2023-04-10.

Richards, B., Agüera y Arcas, B., Lajoie, G., & Sridhar, D. (2023). The illusion of AI's existential risk. *Noema Magazine*. https://noemamag.com/the-illusion-of-ais-existential-risk/.

Rillig, M. C., & Kasirzadeh, A. (2024). Ai personal assistants and sustainability: Risks and opportunities. *Environmental Science & Technology, 58*(17), 7237–7239.

Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. *The Internet Society (ISOC), 80*(15), 1–53.

Rujevic, N. (2024). Serbia monitors government critics with spyware. *DW News*. Available at: https://www.dw.com/en/serbia-monitors-journalists-and-dissidents-with-spyware/a-71132881.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3 ed.). Upper Saddle River, New Jersey 07458: Pearson Education, Inc.

Schaake, M. (2024). *The tech coup: How to save democracy from silicon valley*. Princeton University Press.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. WW Norton & Company.

Schechner, S., & Seetharaman, D. (2023). How worried should we be about AI's threat to humanity? Even tech leaders can't agree.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., Van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature, 461*(7260), 53–59.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(3), 1–42.

Sharma, M., & Kaur, M. (2022). A review of Deepfake technology: an emerging AI threat. *Soft Computing for Security Applications: Proceedings of ICSCS, 2021*, 605–619.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. (2023). Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search*. MIT press.

Steffen, W., Rockström, J., Richardson, K., Lenton, T. M., Folke, C., Liverman, D., Summerhayes, C. P., Barnosky, A. D., Cornell, S. E., Crucifix, M., et al. (2018). Trajectories of the earth system in the anthropocene. *Proceedings of the National Academy of Sciences, 115*(33), 8252–8259.

Tabassi, E. (2023). Artificial intelligence risk management framework (ai rmf 1.0).
Technology and Science Insights and Foresight (2023). Future Risks of Frontier AI: Which capabilities and risks could emerge at the cutting edge of AI in the future? https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf. Accessed: 2024-01-01.
Tegmark, M. (2018). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.
Thom, R. (1974). Stabilité structurelle et morphogenèse. *Poetics, 3*(2), 7–19.
Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. *Risk Analysis, 33*(10), 1772–1787.
Torres, É. P. (2023a). *Human extinction: A history of the science and ethics of annihilation*. Routledge.
Torres, P. (2023). Existential risks: a philosophical analysis. *Inquiry, 66*(4), 614–639.
Tucker, C. (2018). Privacy, algorithms, and artificial intelligence. In *The economics of artificial intelligence: An agenda* (pp. 423–437). University of Chicago Press.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.
United Nations International Strategy for Disaster Reduction. (2009). *UNISDR terminology on disaster risk reduction*. United Nations, Geneva, Switzerland: Technical Report.
Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence, 4*(3), 189–191.
U.S. Environmental Protection Agency. (2024). About risk assessment. Accessed: 29 December 2024.
Von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*. New York: George Braziller.
Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., & Gabriel, I. (2022). Taxonomy of risks posed by language models. In *FAccT '22: 2022 ACM conference on fairness, accountability, and transparency* (pp. 214–229). ACM.
Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": The mad world of AI-generated media. *IT Professional, 22*(5), 90–99.
Wiener, N. (1960). Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science, 131*(3410), 1355–1358.
World Economic Forum. (2024). *The global risks report 2024* (19th ed.). World Economic Forum: Insight Report.
Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2023). The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.
Zeeman, E. C. (1977). *Catastrophe theory*. Reading: Addison-Wesley.
Zuboff, S. (2019). *The age of surveillance capitalism*. London: Profile Books.