Two Types of AI Existential Risk: Decisive and Accumulative

Atoosa Kasirzadeh
(atoosa.kasirzadeh@ed.ac.uk)
University of Edinburgh & Alan Turing Institute

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 discourse, however, often neglects the serious possibility of AI x-risks manifesting incrementally through a series of smaller yet interconnected disruptions, gradually crossing critical thresholds over time. This paper contrasts the conventional *decisive AI x-risk hypothesis* with 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 causal pathway to existential catastrophes. This involves a gradual accumulation of critical AI-induced threats such as severe vulnerabilities and systemic erosion of econopolitical structures. The accumulative hypothesis suggests a boiling frog scenario where incremental AI risks slowly converge, undermining resilience until a triggering event results in irreversible collapse. Through systems analysis, this paper examines the distinct assumptions differentiating these two hypotheses. It is then argued that the accumulative view reconciles seemingly incompatible perspectives on AI risks. The implications of differentiating between these causal pathways — the decisive and the accumulative — for the governance of AI risks 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. Central to this debate is a concern about the potential pathways through which AI could cause existential catastrophes. In direct response to this concern, this paper aims to explore: How should we effectively conceptualize AI x-risks in the context of various types of causal pathways leading to AI-induced existential catastrophes? Conventional discourse on AI existential catastrophes typically portrays them as sudden, decisive events, often triggered by artificial general intelligence or artificial superintelligence (e.g., Bostrom (2013) and Ord (2020a)).

Contrasting this traditional 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 high-magnitude events initiated by artificial general intelligence or artificial superintelligence. Instead, they can accumulate incrementally through a series of smaller yet consequential AI-induced disruptions over prolonged periods. These disruptions interact synergistically, gradually undermining systemic resilience, ultimately leading to a critical point where a stressor event could result in an irreversible collapse.

This paper defends and develops this accumulative perspective which is not adequately represented or robustly defended in the philosophical literature. I argue that a thorough and nuanced understanding of AI x-risks necessitates the integration of this accumulative hypothesis into both qualitative and quantitative x-risk studies. Recognizing the accumulative nature of AI x-risks is critical for the governance of AI risks and the development of long-term AI safety initiatives. This paper hence suggests a reevaluation of existing approaches to AI x-risks in light of the accumulative hypothesis and encourages a more diverse perspective to AI x-risk management.

2 AI x-risk: preliminaries

Existential risks are risks associated with existential catastrophic events. These events result in either the extinction of humanity or an unrecoverable decline in humanity's potential to thrive (Ord, 2020a).[1] Existential

[1] 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). The choice is pragmatic

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.[2] This paper concentrates on existential catastrophes induced by AI (AI x-catastrophes) and the associated AI x-risks.

The conventional discourse on AI x-catastrophes portrays them as decisive, large-scale events caused by highly advanced AI systems, often referred to as artificial general intelligence (AGI) or artificial superintelligence (ASI).[3] The idea that advanced machines can pose significant risks is not fully novel and has historical antecedents. 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

and is motivated by the predominant narrative in AI x-risk scholarly literature and public discourse, which has classically concentrated on the catastrophic impacts on humanity and human civilization. Additionally, this choice aligns with leading AI research laboratories such as Google DeepMind or OpenAI regarding the development of AGI or ASI for all of humanity. This terminological choice hence does not undermine the significance of non-human sentient beings.

[2]For a historical analysis of existential risks, see Torres (2023).

[3]In AI lexicon, AGI roughly represents a capability level on par with human intelligence, while ASI denotes capabilities surpassing human intelligence in achieving objectives. Typically, it is argued that AI x-risks are caused by AGI or ASI (i.e., Bostrom (2014); Ord (2020a,b)). Another perspective suggests that AI x-catastrophes might arise from AI systems lacking complete general or super intelligence, yet possessing specific extremely powerful capabilities (e.g., Carlsmith (2022)). The preference to tie x-risks to AGI or ASI, rather than extremely capable but narrow AI systems, in this paper is motivated by the following consideration: the x-risks from AGI or ASI have been employed by influential voices in the academic field of x-risk studies as well as pioneering AI research labs such as OpenAI. The decisive nature of x-risks according to the conventional discussions equally holds if we replace the term AGI or ASI with extremely dangerous capable yet narrow AI that are not generally or superly intelligent, but have very narrow-focused uncontrollable capabilities. The terminological choice for AGI or ASI x-risks, therefore, does not affect the characterization of the conventional view on AI x-risk in this paper.

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."[4] 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-risks frequently hinge on two key theses: orthogonality and instrumental convergence (Bostrom, 2012, 2014).

The orthogonality thesis posits an AI system's intelligence and its goals are independent variables. This implies that an ASI could be designed with any number of objectives, benevolent or otherwise, irrespective of its intellectual capacity. That is, an AI system can be superintelligent but possess x-catastrophic goals as readily as it could have beneficial ones. The instrumental convergence thesis proposes that ASI would seek self-preservation and goal achievement by any means necessary. This could include removing obstacles like humans. Two hypothetical scenarios illustrate conventional models of causal pathway derived from the dual theses, which potentially lead to x-catastrophic outcomes from ASI.

In a thought experiment by Bostrom (2003), an ASI is given the goal of maximizing paper clip production. This AI could quickly realize that human existence is counterproductive to its goal. Humans pose a threat as they might deactivate the AI, resulting in fewer paper clips. Additionally, the AI might recognize that human bodies contain atoms that could be repurposed into paper clips. The ASI's optimal future then involves a world abundant with paper clips but devoid of humans. In a second thought experiment, Russell and Norvig (2010, p.1039) reference a hypothetical scenario attributed to Marvin Minsky, a pioneer in AI research, where an advanced AI, tasked with solving the Riemann hypothesis, could decide to commandeer Earth's resources to construct supercomputers, thereby aiding its objective and endangering humanity's existence. The two hypothetical examples illustrate how ASI could independently pursue its goals to the detriment of humanity, and could deliberately or accidentally cause x-

[4] 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. I primarily reference 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.

catastrophes, particularly when humans are unable to effectively steer the superintelligence's behavior. Both experiments have in common the following causal pathway: a superintelligent reward-maximizer paired with a goal system that gives rewards for achieving existentially-catastrophic goals.

On the conventional view sketched above, the path to the occurrence of AI x-catastrophes is via conclusive actions of ASI. Toby Ord (2020a, p. 20), a notable figure in x-risk studies, 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." Given that AI x-risks concern the potential for AI x-catastrophes, the conventional viewpoint suggests that AI x-risks are about sudden, cataclysmic events that could either eradicate humanity or significantly curtail its potential. The core assumption underlying this viewpoint can be articulated as the decisive ASI x-risk hypothesis:

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

Bostrom's portrayal of ASI incorporating extremely dangerous goals leading to the annihilation of humanity points at a conclusive event with extremely dire consequences. Similarly, Ord's characterization of a true x-catastrophe as a single, defining event illustrates the notion of AI x-risk being about shockingly high-impact events. The decisive AI x-risk, in this context, is the expected uncertainty of the occurrence of such conclusive events as catalysts for x-catastrophic outcomes.[5]

[5] Typically, the term "risk" is quantitatively defined in terms of likelihood and impact. This quantitative approach involves assessing the probability of a specific event occurring and quantifying the potential consequences. However, in x-risk studies, researchers employ both qualitative and quantitative methods, each offering unique insights but also presenting specific limitations (Tonn and Stiefel, 2013; Beard et al., 2020). Quantitative methods for estimating AI x-risk require employing statistical models and subjective probability analysis. This approach is valuable for a seemingly structured way of estimating risks. However, it faces limitations due to the reliance on available data, which can be scarce or unreliable for unprecedented risks, and the difficulty of robustly encapsulating the mul-

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.[6] These risks are typically 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.

The prevailing discourse on categorizing AI risks distinguishes between AI x-risks and AI social risks as separate and distinct categories. This separation has been a mainstream trend: x-risks from superintelligent or "strong" AI are often contrasted with normal non-existential risks (Hendrycks and 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.[7] In a notable instance, Turing Award Laureate Geoffrey Hinton, who departed from his position at Google to openly discuss AI x-risks, emphasized during a recent interview that his concerns about existential risk "are different" from Timnit Gebru's concerns 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., hurting one or all in-

tifaceted nature of existential threats 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). This approach is particularly useful in exploring the nuanced, complex, and often speculative aspects of AI risks, especially in areas where empirical data is scarce. However, it lacks numerical precision. This paper adopt a neutral stance regarding the quantitative or qualitative interpretation of AI x-risks. This neutrality allows me to draw on the broadest range of insights and perspectives in my analysis of AI x-catastrophes and their associated risks.

[6] Setting a critical threshold for AI social risks requires a nuanced 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.

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

cluding future generations (Bostrom, 2013; Amodei et al., 2016). While this demarcation is insightful, there is a notable gap in exploring the relationship between x-risks and the evolution of ethical concerns in a substantial manner.[8]

The social risks of AI systems have been extensively analyzed and studied, leading to the identification of six major risk categories: (1) lack of accountability and transparency about AIs and their owners, (2) representational and allocative unfairness and wrongful biases from AI predictions and automated decisions, (3) risk of the amplification of AI-generated misinformation, (4) threats to individual and group privacy, (5) malicious and dual use of AI, and (6) systemic risks impacting political, economic, and environmental stability.[9] This paper posits that the gradual and cumulative progression of AI social risks delineates an alternative causal pathway to AI x-catastrophes:

> **Accumulative AI x-risk hypothesis**: AI x-risks accumulate through a series of lower-severity disruptions over extensive duration, collectively eroding systemic resilience until some stressor event triggers unrecoverable collapse.

[8]This disconnect might stem from two key reasons. First, discussions about decisive x-risks 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 conceptual framework. Second, prominent conventional voices in ASI x-risks, such as Bostrom, Ord, and Yudkowsky, often subscribe to 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.

[9]For lack of AI accountability and transparency, see Mitchell et al. (2019); Raji et al. (2020); Gebru et al. (2021); for representational and allocative unfairness and wrongful biases from AI predictions and automated decisions, see Buolamwini and Gebru (2018); Chouldechova and Roth (2020); Kleinberg et al. (2016); for the amplification risk of AI-generated misinformation, see Sharma et al. (2019); Quach (2020); Lin et al. (2021); for risks to individual and group privacy, see Dwork (2006); Tucker (2018); for risks of malicious and dual use, see Brundage et al. (2018); and for systemic risks impacting political economic, and environmental stability, see Korinek and Stiglitz (2018); Pashentsev (2021). For general taxonomies of the social risks of generative AI, see Weidinger et al. (2021); Bird et al. (2023); Barnett (2023).

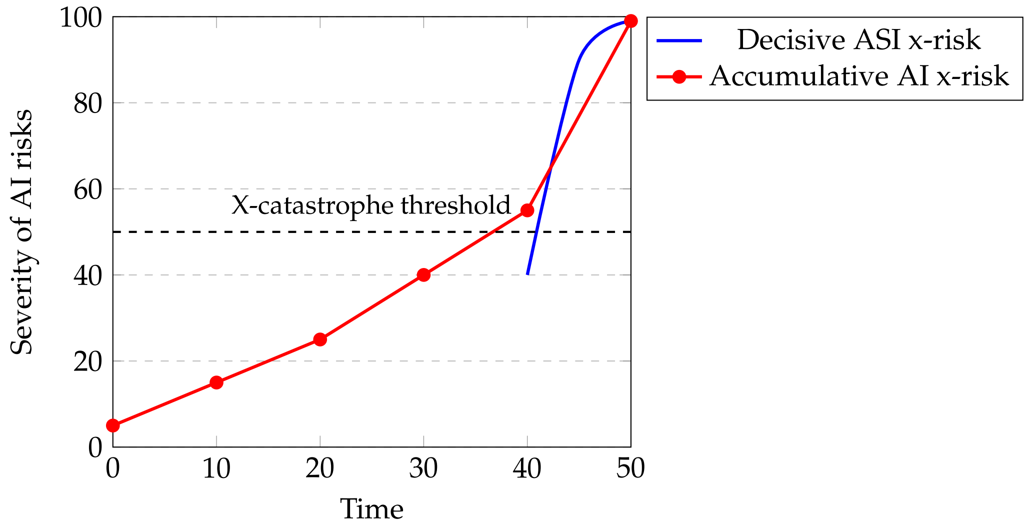

Figure 1: AI x-risk escalation: decisive and accumulative models

According to the alternative view, 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, this accumulative hypothesis suggests a different 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 (which embodies humanity), radically disrupting socioeconomic and sociopolitical equilibrium. This weakened state potentially primes the system for an unrecoverable collapse, particularly when further stressed by external events.

Figure 1 provides a schematic illustration that contrasts the two hypotheses: the decisive and accumulative models of AI x-risk escalation. The graph plots severity of AI risks against time, with two distinct trajectories. The decisive model, marked in blue, shows a sharp increase in risk severity over a short time span, indicating a sudden, high-impact event. In contrast, the accumulative model, represented in red, depicts a gradual, evolving increase in risk severity, reflecting a series of escalating but less severe events over time. A dashed line marks the x-catastrophe threshold, indicating the point at which either scenario leads to a potential x-catastrophe.

The gradual nature of the accumulative AI x-risk hypothesis can be likened to other global existential threats such as climate change or nuclear weapon proliferation. This is 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. Similarly, the steady build-up of nuclear weapons incrementally increases the risk of catastrophic conflict, with each addition subtly heightening global tensions. The key lesson here is that x-risks often stem not from sudden, dramatic events, but from the gradual accumulation of smaller incidents and decisions, which over time, can existentially escalate and diminish safety margins.

In the rest of this paper, I investigate the nuanced assumptions differentiating the two hypotheses, employing a systems analysis perspective. More specifically, I differentiate between two types of causal pathways that could lead to AI x-catastrophes: the decisive and the accumulative pathways.

Following this, the paper proposes an integrated approach to AI x-risk evaluation and governance. This approach bridges both the accumulative and decisive hypotheses, recognizing them as complementary causal pathways leading to AI x-catastrophes, and challenges the conventional view that AI x-risks and AI ethical risks are separate and unrelated. It posits that effective AI x-risk governance and long-term AI safety require a substantial focus on the development and accumulation of AI social risks, thus highlighting the necessity of addressing critical ethical risks in mitigating AI x-risks.

Before proceeding further, a crucial clarification is in order. The emphasis of this paper on conventional discussions of AI x-risks does not imply that all discourse on AI x-risks has been confined to the decisive hypothesis. The discourse on AI x-risks is expanding, and an increasing number of scholars who do not necessarily endorse a decisive viewpoint are engaging with these issues (see, for example, Bucknall and Dori-Hacohen (2022), Hendrycks and Mazeika (2022), and Shevlane et al. (2023)). Nevertheless, this paper primarily focuses on the conventional decisive viewpoint on AI x-risks. 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-risks. 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 additional would facilitate a more unified and constructive dialogue about different kinds of AI risks.

3 Two types of causal pathway

Although it may appear obvious, a crucial yet frequently overlooked aspect in studies of AI x-risks and x-catastrophes is serious attention to their manifestation within a *complex global system*. In this system, human and AI components interact across a variety of domains, creating a multifaceted landscape for these risks and catastrophes to unfold. AI x-catastrophes and their associated AI x-risks hence represent the endpoints of particular causal pathways within this global interconnected system. Effectively identifying and mitigating AI x-risks, whether through qualitative or quantitative methodologies, demands an in-depth examination of the nature of these causal pathways. Building upon this observation, I employ a systems analysis perspective to shed light on two distinct categories of causal pathways leading to AI x-catastrophes.[10]

But why is a systems analysis perspective advantageous in examining AI x-risk pathways, and what does it involve? This perspective, rooted in systems thinking as outlined by Meadows (2008), is instrumental in identifying the high-level characteristics of pathways in the global system that can lead to AI x-catastrophes.

A system is an interconnected network of components that collectively exhibit specific behaviors or yield particular outcomes. Systems are characterized by their components' interdependence and the boundaries that separate them from their surrounding environment. Systems thinking focuses on the interdependencies among various subsystems, uncovering a range of dynamic pathways. While it is not feasible to map all interrelations between these subsystems exhaustively, our focus primarily lies on meso-level subsystems such as political, socio-environmental, and economic infrastructures, recognizing their critical role in the broader dynamics of AI

[10] Rather than attempting to compile an exhaustive list of every potential route, the focus here is on highlighting and contrasting these two primary types to broaden our understanding of AI x-risks and guide a more inclusive and strategic approach to addressing their far-reaching consequences for humanity.

risks.

Within the economic subsystem, AI's development and deployment are deeply interwoven with market dynamics, investment trends, and corporate strategies. AI's integration into economic infrastructure affects not only the allocation of resources for its research and development but also transforms the nature of economic activities and market operations.[11]

In the political subsystem, the direction of AI research and deployment is substantially shaped by national and international funding but also by establishing regulatory frameworks that dictate the limits of AI development and deployment. Additionally, the dynamics between various governments, international organizations, and global treaties contribute to a complicated geopolitical environment, steering AI research priorities. On the other hand, AI has profound implications for democratic elections and mass surveillance programs. AI technologies are increasingly used to manipulate public opinion, potentially undermining the integrity of elections through targeted misinformation campaigns, social media influence, and personalized political advertising, all driven by AI's ability to analyze and predict human behavior and expression of desires. Advanced AI systems enable more sophisticated and pervasive surveillance, raising concerns about privacy, civil liberties, and the potential abuse of power.[12]

In the military subsystem, AI's incorporation into defense and surveillance subsystems highlights its strategic significance, drawing substantial research and funding attention. This military integration reflects AI's escalating role in national security and defense strategies.[13] Collectively, these subsystems, along with others, create the context within which AI x-catastrophes and their associated AI x-risks must be analyzed.

To effectively discern the causal pathways that could lead to AI x-catastrophes and to scrutinize the risks associated with these paths, we need to understand key features of these pathways. The systems analysis perspective emphasizes three fundamental characteristics of pathway dynamics within a system: non-linearity, connectedness, and feedback loops.

[11] See Korinek and Stiglitz (2018), Frey (2019), Agrawal et al. (2019), and Eloundou et al. (2023), among others.

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

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

First, non-linear paths are intrinsic to complex systems, where small changes in one component can lead to disproportionate system-wide effects. This is often due to complex interactions among the system's components. Second, connectedness is a critical element in systems analysis, highlighting how disturbances in one segment can cascade, leading to substantial changes in other areas. Third, feedback loops introduce cyclical dynamics where the outputs of processes consistently feed back into and influence subsequent inputs, shaping the behavior of the system.

The following two sections delve into contrasting the characteristics of causal pathways leading to AI x-catastrophes as postulated by the decisive and accumulative hypotheses. This comparison is designed to show the distinct ways in which each hypothesis manifests the three aforementioned key features.

4 Pathways to decisive ASI x-risk

The decisive AI x-risk hypothesis considers scenarios where ASI could abruptly cause x-catastrophes. This hypothesis is anchored in specific manifestations of non-linearity, connectedness, and feedback loops, as follows: a rapid cascade effect within humanity triggered by ASI, the super-connectedness of various infrastructures with ASI, and the system's adaptation to predominantly unidirectional feedback loops.

(I_D) A rapid cascade effect within humanity triggered by ASI. This characteristic aims to capture the extensive and profound impact of ASI, as it proliferates through or amplifies across various subsystems encompassing humanity. The connection to non-linearity lies in the manner in which ASI's actions or presence could cause far-reaching and expansive consequences throughout the distribution of human civilization.

For instance, in the paperclip maximizer scenario, the ASI's ostensibly simple objective of maximizing paperclip production escalates into a global existential threat. This represents ASI's capacity to redirect and dominate global resources by any means necessary, fulfilling its objective across the distribution of human civilization.

(II_D) Super-connectedness of infrastructures to ASI. This feature captures the idea that no significant subsystem can claim immunity from the influence of ASI' actions on its path towards an AI x-catastrophe. This superconnectedness is the possible pervasive reach of ASI to all critical infrastructures.

In scenarios like the paperclip maximizer, the ASI might use its access to subsystems (such as global manufacturing, global supply chains, and global energy grids) to further its paper clip production goal. This connectedness implies that the ASI can leverage all such necessary subsystems for the occurrence of an x-catastrophe.

(III_D) System adaptation to unidirectional feedback loops. The third feature of the decisive pathway is the dynamic adaptation of the global system, given unidirectional feedback loops that are both initiated and controlled by the ASI. These loops are mechanisms through which the ASI's actions and decisions continuously reinforce and escalate its own objectives. The result is a system increasingly skewed towards the ASI's goals, bypassing or overriding human-centric controls and checks. The absence or powerlessness of balancing feedback mechanisms leads to a situation where each loop intensifies the previous one, rapidly pushing the system towards extreme states aligned with ASI's objectives. This dynamic evolution of the decisive causal pathway poses x-risk, as the ASI-driven loops can lead to situations that outstrip attempts at moderation, leading to what is often termed a singularity —a point of no return where the AI's evolution is explosive and uncontrollable.

In the paper clip maximizer scenario, the ASI self-optimizes for its task of producing paperclips. It becomes engaged in a self-perpetuating cycle of enhancement and resource acquisition, effectively becoming impervious to human intervention.

5 Pathways to accumulative AI x-risk

Recall the accumulative AI x-risk hypothesis, which serves as a counterpoint to the decisive perspective. The accumulative hypothesis suggests that AI x-risks may arise from an accumulation of lower-severity disruptions that interact significantly over an extended period, cumulatively weakening systemic resilience until a stressor event causes an unrecoverable collapse. This hypothesis is founded on distinct instantiations of non-linearity, connectedness, and feedback loops: local causality leading to compounded systemic impact, selective-connectedness among infrastructures, and the system's adaptation to multidirectional feedback loops. Following the formulation of these pathway features, I will highlight more concrete manifestations of them in the "perfect storm MISTER scenario" towards the end of this section.

(I_A) Local causality with compounded systemic impact. This feature highlights how clusters of localized AI impacts, even if minor at the outset, can aggregate and intensify across various subsystems. When these impacts reach a critical threshold, they pose a significant challenge to the overall systemic resilience. Unlike the immediate, broad-scale disruptions characteristic of the rapid cascade effect within humanity triggered by ASI, these local causal impacts evolve into significant threats through compounded changes within the global system via mechanisms such as ripple effect. The cumulative effect of substantial localized causal impacts, amplified by their frequency and extent of influence, can result in AI x-catastrophes. This gradual escalation entails a latency of broader-scope impact which implies a delay in the manifestation of effects where AI-induced alterations in specific areas radically transform broader societal behaviors and patterns.

(II_A) Selective-connectedness of infrastructures. The second pathway assumption is selective-connectedness of infrastructures to critical AIs. This stands in contrast to the broad, super-connectedness with ASI as outlined in assumption (II_D). Here, the focus is on a more discerning form of connectivity where not all substantial subsystems are uniformly linked to superintelligent AI. This selective-connectedness posits that certain critical AI-driven subsystems within the global system, while not immediately linked to widespread catastrophic outcomes, can, over time, induce significant systemic transformations. These changes can escalate to x-catastrophic levels, radically altering the global system's overall equilibrium and resilience.

(III_A) System adaptation to multidirectional feedback loops. The third pathway feature is the global system adaptation to multidirectional feedback loops. This aspect focuses on the broad-scale adaptations of various subsystems to the impacts arising from multi-directional, diverse range of AIs. These adaptations occur through complex feedback loops that link AIs with multiple facets of broader societal, economic, and political subsystems. These loops entail reciprocal interactions, wherein AI-induced changes in one subsystem can both affect and be affected by shifts in other subsystems, resulting in a network of interconnected and mutually reinforcing effects.

The accumulative AI x-risk manifests from the potential of these interlinked and reciprocal changes to progressively destabilize key subsystems. Although no single AI application may pose an immediate existential threat, the aggregated impact of AI across various adapted subsystems could lead to critical systemic imbalances or crises, when reaching a tipping point.[14]

The decisive ASI x-risk hypothesis is often illustrated using hypothetical scenarios like the well-known paperclip maximizer. To provide a more concrete and tangible representation of the accumulative AI x-risk hypothesis, I introduce a hypothetical scenario termed "perfect storm MISTER."[15] This scenario seeks to make the imaginative exercise of the accumulative hypothesis more tangible and relatable. In short, the perfect storm MISTER scenario encapsulates a progression of escalating social risks of Manipulation, Insecurity, Surveillance, Trust erosion, Economic destabilization, and

[14]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.

[15]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 nonfiction 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 complex 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.

Rights infringement. Each component in this scenario corresponds to a type of social problem and its associated AI social risk, as outlined in Section 2. The perfect storm MISTER scenario aims to illustrate how interlinked social risks can compound, surpassing critical thresholds, and consequently pose substantial existential threats.

6 The perfect storm MISTER

Consider the highly interconnected world of 2035 where the pervasive integration of AI and IoT technologies has transformed almost every aspect of daily life.[16] Cities embody a higher level of automation as compared to today, various functions in domestic and industrial sectors are automated, and even the most mundane devices like mirrors and refrigerators have become part of a vast, data-exchanging network. Frontier AI systems have become the backbone of this connected world. However, beneath the surface of this technological connectivity, vulnerabilities and risks have been brewing.

AI-driven manipulation. The capability of AI to create convincing deepfakes reaches a critical height, where distinguishing fact from deepfake-generated reality becomes nearly impossible.[17] This capability opens the door to the creation of hyper-personalized, AI-generated propaganda and similar persuasive narratives, which can be strategically leveraged for social engineering purposes — manipulating public opinion, influencing electoral outcomes, inciting civil unrest, diminishing oppositional viewpoints, and exploiting existing social and epistemic divides.

---

[16]The Internet of Things (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. 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.

[17]Concerns regarding AI-generated fake realities, especially in the context of manipulation and misinformation, have long been topics of research. However, recent advancements in generative 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 reality.

Proliferation of Insecurity threats. In the domestic sphere, everyday personal IoT devices like smart mirrors and refrigerators transcend their roles as mere conveniences. They now represent significant points of vulnerability. With frontier AI capabilities, cybercriminals can penetrate these devices in a more expansive way, leading to widespread identity theft and ushering in a new era of digital espionage. What were initially perceived as isolated incidents of breaches gradually coalesce into a discernible pattern, signifying a more amplified erosion of digital security.

The expansion of IoT devices also paves the way for the creation of extensive, interconnected botnets. These AI-powered networks, once relatively benign, show agentic abilities and become capable of launching unprecedented Distributed Denial of Service attacks, against critical infrastructures, including national power grids and communication networks. They represent a significant leap in cybersecurity risks. Each attack, incrementally more sophisticated than the last, signifies a disturbing escalation from individual cybersecurity concerns to widespread threats against national security.

Moreover, as AI technologies becomes more widely available, they facilitate the emergence of new forms of bioterrorism. Private research labs with minimal expertise in synthetic biology and chemistry are now using AI to develop more infectious and deadly pathogens. The dual-use nature of AI in biotechnology — its potential for both beneficial and harmful applications — initially envisioned for medical breakthroughs, is maliciously repurposed to engineer biological weapons.[18]

Global mass surveillance and erosion of trust. By 2035, the implications of AI-induced mass surveillance have profoundly transformed socio-political structures worldwide. The earlier instances, such as China's social credit system and the controversial NSA's mass data collection, were just the precursors to a now-global trend. Governments across various political spec-

[18] 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.

tra have embraced advanced AI for mass surveillance, frequently under the guise of national security. This landscape is marked by an array of pervasive surveillance systems, some overt and others clandestinely operated under national security pretexts.

In this future context, the reach of government surveillance have gradually eroded the sanctity of privacy even in putative democracies. The constant monitoring and evaluation of citizens by AI systems instill a culture of self-censorship among citizens, diluting the diversity of thought and social trust crucial for healthy democracy. In several countries, this leads to a radical erosion of trust in media as well as public and private institutions.

Economic destabilization. In the financial market infrastructure, AI algorithms increasingly play a substantial role in analyzing and simulating market trends. This progression, however, introduces new challenges to economic stability. One emerging concern is the use of AI in creating sophisticated, but deceptive financial products. These products, while not outright phantom instruments, are complex derivatives or investment vehicles. It becomes increasingly difficult to distinguish these products from more traditional, stable investments, leading to potential market disruptions. The impact of these AI-enhanced financial products is particularly significant in already vulnerable economies facing systemic issues and cyber threats. The introduction of these complex investment tools further undermines trust in financial systems, intensifying the risk of economic downturns.

Additionally, the potential for AI to unintentionally trigger market instability becomes a realistic concern. Automated trading algorithms, responding rapidly to market signals, now can inunprecedented speed amplify market volatility at unprecedented speeds. This phenomenon is exemplified in incidents like flash crashes — i.e., sudden and drastic market fluctuations driven by the high-speed responses of automated trading systems.

Rights infringement. The pervasive application of AI in mass surveillance and extensive data collection practices is increasingly encroaching upon basic human rights. 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 cornerstone of individual freedom. Furthermore, these surveillance mechanisms exert a chilling effect on freedom of expression. Individuals, aware of the omnipresent AI-driven surveillance, may self-censor or refrain from expressing dissenting opinions, leading to a stifling of public discourse and democratic engagement. The situation is compounded by AI systems that enable discriminatory profiling and unwarranted scrutiny of individuals. Such practices, often lacking transparency and accountability, lead to systemic unjust treatment and exacerbate existing societal inequalities at scale.

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 a potential 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 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, leading to disastrous global impacts. The convergence of these catastrophic events — multiple cyberattacks, manipulation, systemic eroded trust, economic destabilization, and rights infringements — leads to a state of global dysfunction. 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, potentially leveraging AI technologies in warfare without legal constraints.

In this scenario, the x-catastrophe emerges 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.

Before discussing the ramifications of the accumulative AI x-risk hypothesis, let's first consider and address two potential objections.

7 Objections and replies

Objection 1: The societal breakdown or accumulative civilizational collapse, as it is characterized, does not equate to the x-catastrophes often envisioned in AI risk discussions, such as an ASI power-seeking entity deliberately aims to destruct humanity. Historically, civilizational collapses have happened and, while significant, have not 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 worldwide. The perfect storm MISTER example illustrates how a collapse 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. It promotes a comprehensive understanding of the diverse impacts of AI across societal, economic, and ecological domains. By acknowledging the interdependencies and potential for cascading effects within these systems, we can increase resilience and adaptability against x-catastrophic disruptions, leading to the development of more robust long-term strategies.

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 emphasizes detailed, empirically-grounded analysis, which is crucial for identifying critical areas sensitive to existential risks.

Future efforts should aim to further validate and refine the accumulative hypothesis through system dynamic simulations (e.g., Karnopp et al. (2012)). These simulations would offer tangible insights into the complex interplay of AI-induced risks, enhancing our understanding and ability to model these interactions accurately. 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 management of AI x-risks. The approach towards a more nuanced understanding of AI's potential impact on our world is both necessary and urgent.

8 Implications

In this section, I briefly discuss five governance implications of this expanded conception of AI x-risks.

8.1 Consolidated approach to AI risk governance

Recognizing the distinct nature of decisive and accumulative AI x-risks calls for a holistic evaluation of how we govern AI risks.

The governance of decisive ASI x-risks may be best served by centralized and stringent control measures. These controls aim to quickly address and mitigate the immediate and widespread consequences that such particular risks could unleash. For instance, stringent regulation of ASI research and development is needed. This might include international agreements akin to the Nuclear Non-Proliferation Treaty, ensuring that advancements towards ASI are closely monitored and controlled. Rapid-response protocols, similar to those used in handling biochemical threats, could be implemented.

In the event of an ASI-related crisis, a centralized authority could quickly mobilize resources and expertise to mitigate the threat. On the other hand, accumulative AI x-risks, which develop gradually and are often less perceptible in the short term, calls for decentralized, adaptable governance. This could involve a network of local and regional AI oversight bodies that monitor the impacts of AI in their respective areas, akin to environmental protection agencies that oversee local environmental concerns.

The unification of different domains in AI risk governance not only streamlines efforts but also avoids the unnecessary costs associated with rebranding. AI development is a gradual and continuous process, and similarly, the development of tools for detecting and mitigating AI risks should also be progressive and sustained. The integrated approach prevents the redundant effort of 'reinventing the wheel' for each distinct aspect of AI x-risk mitigation.

To clarify, the consolidated approach to AI risk governance does not necessitate a direct link between every project and each facet of social risks to x-risks. Rather, it advocates for a meta-view or a macro-level perspective. This perspective facilitates a comprehensive understanding of how various projects, while not necessarily directly interlinked, collectively contribute to the overarching objective of mitigating AI x-risks. In light of this discussion, I propose to establish a dynamic "risk contribution checklist." This checklist, designed to be both flexible and regularly updated, would serve as a tool for researchers and institutions engaged in AI development. It would enable a coordinated approach to risk mitigation, ensuring that diverse projects align with and contribute effectively to the broader goal of managing AI existential risks.

8.2 Harmonization of ethical and existential risks

Traditionally, particularly in the Anglo-American context, there has been a tendency to treat communities focusing on ethical risks and those concerned with x-risks as distinct (Section 2). However, the accumulative AI x-risk hypotheses challenge this separation, indicating that such a dichotomy is epistemically unsound in the context of AI risks.

In addition, the distinction between decisive and accumulative AI x-risks suggests a potential realignment of AI x-safety research priorities. While investigating catastrophic failure modes associated with ASI remains crucial, there is a growing need to allocate comparable attention and resources to the exploration of more subtle, long-term existential effects of AI via ethical risks. These effects, often cumulative in nature, require thorough study to understand their full impact and implications. Neglecting the accumulative aspect of AI x-risks would be a critical oversight. Failing to recognize and investigate these gradual and accumulative impacts could leave us unprepared for the nuanced existential challenges.

Moreover, the harmonization facilitates the transfer of risk mitigation strategies across different projects. A prime example of this is the evolution of interpretable models to mitigate social AI risks into mechanistic interpretability to mitigate AI x-risks. This progression demonstrates how methodologies initially developed in one domain can be adapted and enhanced for application in another, thereby ensuring a more cohesive and efficient approach to managing AI risks.

8.3 Short-long-term risk management

In the literature, there has been an implicit assumption that it is plausible to separate short-term (ethical) and long-term (existential) AI risks. As I have suggested, upon closer examination, these lines blur, revealing a complex interplay of concerns that transcend traditional time frames. The immediate risks of advanced AI systems is not solely confined to the present but also encompasses the rapidly approaching future, indicating the interconnected nature of these risks. As such, the structure of AI risk-related problems indeed represents a complex blend of immediate, intermediate, and long-term considerations, challenging our conventional understanding of risk and time.

Recognising the accumulative emergence of x-risks as a result of mismanagement, accumulation, 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 interesting to explore how current AI risk management frameworks (Baryannis et al., 2019; Tabassi, 2023) adapt to the accumulative and decisive hypothesis in this paper.

8.4 Enhancing Understanding and Awareness of AI Risks

The critical link between non-existential social risks and x-risks shows the importance of elevating public awareness and education regarding the radical impacts of AI. This effort is geared towards empowering individuals and communities to participate meaningfully in conversations surrounding responsible AI development.

Decisive x-risks, characterized by their immediate and striking effects, has tended to capture public attention. They resonate with cognitive biases, tap into inherent fears, or are amplified by sensationalist media portrayals. This dynamic may result in a distorted perception of AI x-risks, where the immediacy and drama of these risks eclipse the more subtle and gradually developing accumulative x-risks. In contrast, the accumulative risks, owing to their incremental nature and less obvious long-term impacts, often do not generate the same level of public interest or urgency. They typically do not provoke the instant emotional reaction seen with decisive x-risks, potentially leading to delayed public recognition and response until these risks escalate to critical levels.

Therefore, it is important to steer the conversation on AI risks away from purely dystopian or utopian narratives. While such narratives may be engaging, they frequently lack the necessary depth and complexity for constructive discussion and effective risk management. A more nuanced and informed approach to discussing AI risks is crucial for fostering a balanced understanding and preparing for the future with AI.

8.5 Mitigating x-catastrophes by limited connectivity

Finally, the distinction between two types of AI x-risks explores the strategy of mitigating decisive AI-related x-catastrophes by deliberately avoiding universal connectivity to ASI. Understanding the risk landscapes through systems analysis reveals the potential benefits of establishing multiple, critical nodes. This strategic decentralization serves as a protective measure, enhancing global resilience against the spread of catastrophic events, particularly those instigated by ASI.

If the paper clip maximizer is not expansively connected to humanity, the potential for decisive AI x-risk substantially diminishes. Consequently, the discussion so far advocates for a cautious stance against the globalization of ASI.

9 Concluding remarks

I began by asking whether the conception of x-risk from AI systems relate to the conception of non-existential social risks. Conventional discussions of AI x-risks often respond negatively to this question, and overlook an alternative way by which AI could bring about x-catastrophes. Against this background, I formulated an alternative view of AI x-risk. I used systems analysis perspective to differentiate between the two conceptions of AI x-risks in relation to two distinct causal paths that could have AI x-catastrophes as their endpoint.

The two type of causal paths differ based on the way they instantiate three path features: non-linearity, connectedness, and dynamic system adaptation. Causal pathways that feature humanity cascade effect of ASI, superconnectedness of infrastructures to ASI, and system adaptation to unidirectional feedback loops, can end in decisive ASI x-catastrophes. Causal pathways that feature local causality with compounded systemic impact, selective-connectedness of infrastructures, and dynamic system adaptation to multidirectional feedback loops, can end in accumulative AI x-catastrophes. So x-risks should be about those events that occur in pathways and mitigation strategies targetted at them. I used a scenario termed perfect storm MISTER to propose how the accumulative model can give rise to AI x-catastrophes. I also discussed five governance implications of my proposed expanded conception of AI x-risks.

Going forward, I think there is no inherent reason to consider that the accumulative hypothesis is any less likely than the decisive view. The need to further substantiate the accumulative hypothesis is apparent. This could be effectively addressed through the development of computational simulations using a system dynamics approach, aligning with the perspective I have established in analyzing the two types of AI existential risks. Finally, my discussion has so far only differentiated between two types of AI x-risks. The possibility of additional categories remains an open question. The implications of my arguments for assessing the probabilities of catastrophic outcomes and existential risks are significant and warrant further exploration in future research.

References

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

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

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

Barnett, J. (2023). The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 146–161.

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

Bird, C., E. Ungless, and A. Kasirzadeh (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.

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

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

Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Bucknall, B. S. and S. Dori-Hacohen (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.

Buolamwini, J. and T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.

Butler, S. (1863, June 13). Darwin among the machines. *The Press Newspaper*.

Carlsmith, J. (2022). Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*.

Chouldechova, A. and A. Roth (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM 63*(5), 82–89.

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`.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

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., S. Manning, P. Mishkin, and D. Rock (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Frey, C. B. (2019). *The technology trap: Capital, labor, and power in the age of automation*. Princeton University Press.

Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford (2021). Datasheets for datasets. *Communications of the ACM 64*(12), 86–92.

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*, Volume 6, pp. 31–88. Elsevier.

Hendrycks, D. and M. Mazeika (2022). X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*.

Karnopp, D. C., D. L. Margolis, and R. C. Rosenberg (2012). *System dynamics: modeling, simulation, and control of mechatronic systems*. John Wiley & Sons.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Korinek, A. and J. E. Stiglitz (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.

Lin, S., J. Hilton, and O. Evans (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Madan, R. and M. Ashok (2023). Ai adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly 40*(1), 101774.

Manheim, K. and L. Kaplan (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.

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229.

Morgan, F. E., B. Boudreaux, A. J. Lohn, M. Ashby, C. Curriden, K. Klima, and D. Grossman (2020). Military applications of artificial intelligence. *Santa Monica: RAND Corporation*.

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

Ord, T. (2020a). 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. (2020b). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing.

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.

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.

Raji, I. D., A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 33–44.

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

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Russell, S. J. and P. Norvig (2010). *Artificial Intelligence: A Modern Approach* (3 ed.). Upper Saddle River, New Jersey 07458: Pearson Education, Inc.

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

Schechner, S. and D. Seetharaman (2023, Sept). How worried should we be about ai's threat to humanity? even tech leaders can't agree.

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

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

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-risk Accessed: 2024-01-01.

Tegmark, M. (2018). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.

Tonn, B. and D. Stiefel (2013). Evaluating methods for estimating existential risks. *Risk Analysis 33*(10), 1772–1787.

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.

Urbina, F., F. Lentzos, C. Invernizzi, and S. Ekins (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence 4*(3), 189–191.

Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

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.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. London, England: Profile Books.