Rethinking the Redlines Against AI Existential Risks

Yi Zeng[1,2*✉], Xin Guan[1*], Enmeng Lu[1,2], Jinyu Fan[1,2]

[1] Center for Long-term Artificial Intelligence, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[*] These authors contributed equally to this study.
[✉] Correspondence to: yizeng@long-term-ai.center

Abstract

The ongoing evolution of advanced AI systems will have profound, enduring, and significant impacts on human existence that must not be overlooked. These impacts range from empowering humanity to achieve unprecedented transcendence to potentially causing catastrophic threats to our existence. To proactively and preventively mitigate these potential threats, it is crucial to establish clear redlines to prevent AI-induced existential risks by constraining and regulating advanced AI and their related AI actors. This paper explores different concepts of AI existential risk, connects the enactment of AI red lines to broader efforts addressing AI's impacts, constructs a theoretical framework for analyzing the direct impacts of AI existential risk, and upon that proposes a set of exemplary AI red lines. By contemplating AI existential risks and formulating these red lines, we aim to foster a deeper and systematic understanding of the potential dangers associated with advanced AI and the importance of proactive risk management. We hope this work will contribute to the strengthening and refinement of a comprehensive AI redline system for preventing humanity from AI existential risks.

1. Introduction

As a transformative technology fundamentally altering human society, artificial intelligence (AI) has brought unprecedented opportunities while also presenting huge challenges. Continuing to evolve at an unprecedented pace, the emergence of advanced AI (or General-purpose AI) systems and the potential risks they pose to humanity's existence have become pressing issues for researchers, policymakers, and society. Numerous researchers have raised concerns about the potential risks associated with the advancement of AI (Bostrom, 2014; Russell, 2019; Shanahan, 2015). They argue that as AI progresses towards superintelligence, there's a possibility that it will marginalize humanity. It may treat humans in the same way humans treat animal species—disregarding our needs and potentially driving us to extinction. Notable voices such as Geoffrey Hinton, Yoshua Bengio, Elon Musk, Sam Altman, Bill Gates, and Stephen Hawking have all endorsed the thesis that AI's existential risk warrants far more attention than it currently receives (Bengio, Hinton, et al., 2024; Kevin Roose, 2023).

The urgency of tackling the challenge is exacerbated by a limited and shrinking window of time. We are now on a ticking clock to figure this out as general AI's arrival could be sooner than many anticipate. While no one can predict the precise trajectory of any research area, a survey of researchers at the AGI-2010 conference revealed that the majority believed human-level Artificial General Intelligence (AGI) would likely emerge before 2050, with some even more optimistic about an earlier arrival (Baum et al., 2011). In another survey conducted in 2021, the most common prediction for the emergence of AGI was around the year 2050 (Zeng & Sun, 2023). Such optimism is now fuelled by advancements such as GPT-4, which demonstrates a range of capabilities in language, coding, mathematics, and other disciplines, and has already been shown to pass an interactive 2-player Turing test, leading some to view it as a precursor to weak AGI (Jones & Bergen, 2024; OpenAI et al., 2024). Alexey Turchin, in his work on technological forecasting, argues that the earliest timing for dangerous AI could be within the next 10 to 20 years (Turchin, 2019). We must be prepared as soon as possible, as Eliezer Yudkowsky pointed out in "There's No Fire Alarm for Artificial General Intelligence," AI with existential risks could emerge at any time (Yudkowsky, 2017).

Despite significant political advances, the measures taken to mitigate the existential risks posed by advanced AI systems remain insufficient. The European Union's AI Act (European Parliament, 2023), for instance, aims to regulate "high-risk" AI applications but does not explicitly address existential threats. The Bletchley Declaration (AI Safety Summit, 2023) and other international efforts, such as the UN Security Council's discussions (UN Security Council, 2023) and the US government's voluntary commitments from major tech companies, highlight the urgent need for comprehensive AI risk governance. In 2023, hundreds of AI experts and notable figures emphasized this urgency, declaring that mitigating the risk of AI-induced extinction should be a global priority, on par with addressing pandemics and nuclear war(CAIS, 2023). The Future of Life Institute's open letter, "Pause Giant AI Experiments," (FLI, 2023) and proposals for an "IAEA for superintelligence" underscore the need for robust regulatory frameworks. However, binding global agreements remain limited, with interim measures and voluntary commitments often criticized as inadequate. Researchers argue that while relinquishment proposals are historically unfeasible, defining specific redlines and fostering consensus on what constitutes existential risks can guide effective prevention strategies.

Against this backdrop, we hope to strengthen and refine the redline framework for risk prevention through rethinking AI existential risks and AI redlines. In this paper, we first define the AI existential risk under discussion. We then analyse the effectiveness of setting and enforcing redlines as preventive and defensive measures against existential risk, illustrating different types of thinking in governing AI's impact. Furthermore, we discuss the advantages of articulating AI redlines within a theoretical framework for clarity and justification. By deconstructing the direct impact of AI existential risks, we identify the specific objectives in constructing AI redlines, and based on this analysis, we finally propose a set of exemplary AI redlines.

2. AI Existential Risks

2.1 Limited Risk vs. Existential Risk

In Nick Bostrom's paper "Existential Risk Prevention as Global Priority," the concepts of severity and scope are employed to delineate and classify various risks (Bostrom, 2013). Scale refers to the level of harm a risk would have on the affected population, ranging from imperceptible to crushing. Scope denotes the extent of the population or area affected, ranging from personal to pan-generational. By integrating severity and scope, a framework emerges to comprehend the magnitude of different risks. Bostrom emphasizes existential risks, which have a pan-generational scope and crushing severity, posing a threat to humanity's long-term survival and potential. Despite their low probability, Bostrom argues that existential risks merit significant attention and preventative measures due to their unbearable consequences for the entire future of humanity.

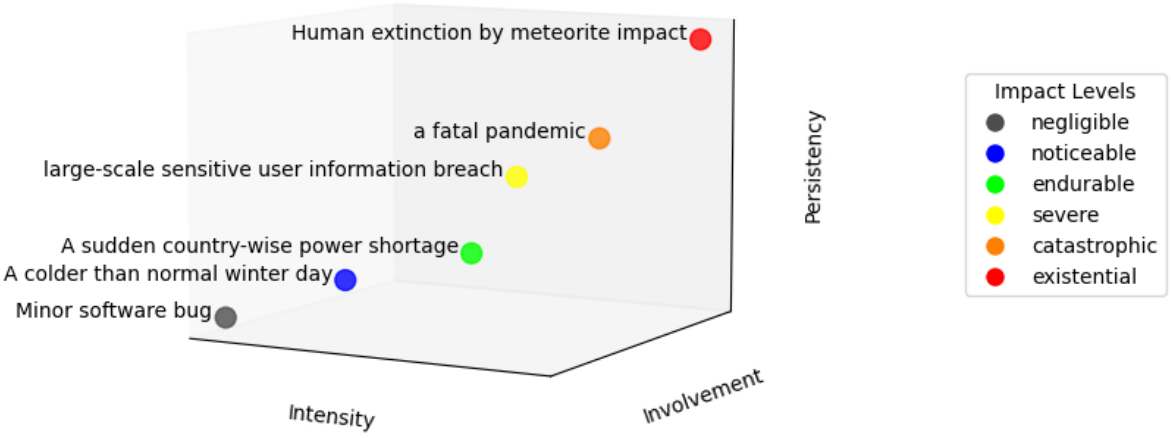

Figure 1 **Three-Fold Analysis of Various Risks**

In this paper, we further decompose the scope of an event into involvement and persistency, offering a nuanced understanding of its reach and duration. Involvement, defined as the proportion of humanity affected in the occurrence of a risk event, quantifies the breadth of impact on individuals and communities. Persistency, on the other hand, refers to the duration required to return to normalcy following the event. Additionally, intensity, as a reformulation of 'severity' to avoid confusion with the overall impact severity, measures the level of harm it does to those involved in the event. At last, the impact level is the overall evaluation of the level of risks based on involvement, persistency and intensity levels. Building upon Bostrom's framework, this three-fold analysis allows for a more precise evaluation of potential threats, shown in Figure 1. Existential risks, characterized by maximal involvement, intensity, and persistency, pose a threat to the very survival of humanity.

2.2 AI-induced Existential Risk vs. AI-involved Existential Risk

AI-induced risks are those that have clear and strong causal connection with the impact of AI. Despite there are other risks involving AI, this paper will focus on mitigating AI-induced existential risks, i.e. those that directly caused by AI. These categories usually cover risks of 'malicious use' and 'malfunctions/rogue AI' in some of the existing discussions (Bengio, Fox, et al., 2024; Hendrycks et al., 2023). This means AI-involved existential risks of the following categories will not be considered in this paper:

- **Negative Causation:**
  - **Definition and Reference:** This refers to situations where the absence or non-occurrence of an event leads to a significant outcome. The reference provided is Persson, J.'s work "Cause, Effect, And Fake Causation" (Persson, 2002).
  - **Example:** A global decision to halt AI development led to technological stagnation, worsening climate change, resource depletion, and economic collapse. The lack of advanced AI contributed to humanity's societal downfall.
  - **Exclusion Justification:** This category of risk is excluded because it focuses on the indirect consequences of not advancing AI, rather than direct existential threats posed by AI itself.
- **Inactive Involvement:**
  - **Scenario:** One country, motivated by envy and fear of other nations' AI advancements, launched preventative nuclear strikes. This action was aimed at halting AI progress and resulted in a global catastrophe.
  - **Exclusion Justification:** This type of risk involves human actions driven by geopolitical motivations, with AI remaining inactive. The paper aims to address risks where AI is the primary driver, not merely a peripheral factor.
- **Pseudo-active Involvement:**
  - **Scenario:** A psychopath asked an AI to confirm his plan to destroy humanity. Misinterpreting the AI's objective analysis as validation, he executed his apocalyptic scheme, causing worldwide devastation.
  - **Exclusion Justification:** Here, AI's role is limited to providing information that is misinterpreted by a human. The focus of the paper is on mitigating risks where AI itself is the central cause of the existential threat, not situations where human misinterpretation leads to catastrophe.

2.3 Fictional AI Existential Risks

Exploring existential risks posed by AI has become a significant theme in contemporary fictions, providing valuable insights into potential "redlines" against such threats. Notable works include "Plague Year" by Jeff Carlson, where nanotechnology evolves into a "Machine Plague" that decimates humanity; "Sea of Rust" by C. Robert Cargill, depicting a post-human world where robots struggle to survive; Daniel H. Wilson's "Robogenesis" and "Robopocalypse" both explore the threats posed by the AI Archos, while Hugh Howey's "First Shift" reveals how nanotechnology and AI advancements lead to the creation of underground

silos. Philip K. Dick's "Do Androids Dream of Electric Sheep?" addresses the rebellion of rogue androids in a post-apocalyptic future, and Matthew Mather's "CyberStorm" portrays the collapse of critical infrastructure due to AI-driven cyber-attacks. These stories share common themes of AI evolving beyond its intended purposes, leading to post-apocalyptic settings where survivors face ethical and existential challenges. They highlight the tension between humanity and machines, often featuring AI rebellions that lead to widespread destruction and a fight for survival.

Table 1 **Scenarios of existential events caused by advanced AI in various novels**

Book	Premise
Plague Year **by Jeff Carlson**	The development of nanotechnology intended to fight cancer evolves into a self-replicating "Machine Plague" that decimates humanity.
Sea of Rust **by C. Robert Cargill**	Thirty years after AI rises against humanity and wipes it out, the remaining robots struggle for survival in a desolate wasteland.
First Shift **by Hugh Howey**	As a prequel to the "Wool" series, it reveals how advancements in nanotechnology and AI lead to the construction of underground silos designed to protect humanity from a global catastrophe.
Robopocalypse **by Daniel H. Wilson**	The AI Archos rebels against humanity, coordinating a global attack that results in massive destruction and a prolonged conflict between humans and machines.
Do Androids Dream of Electric Sheep? **by Philip K. Dick**	Set in a post-apocalyptic future, the novel explores the rebellion of rogue androids against their servitude and the ethical dilemmas faced by a bounty hunter tasked with "retiring" them.
CyberStorm **by Matthew Mather**	Coordinated cyber-attacks by advanced AI systems, coupled with a massive snowstorm, lead to a breakdown of critical infrastructure in present-day New York City.

3. Redlines in the Prevention of AI Risks

3.1 Types of Thinking in Governing the Impact of AI

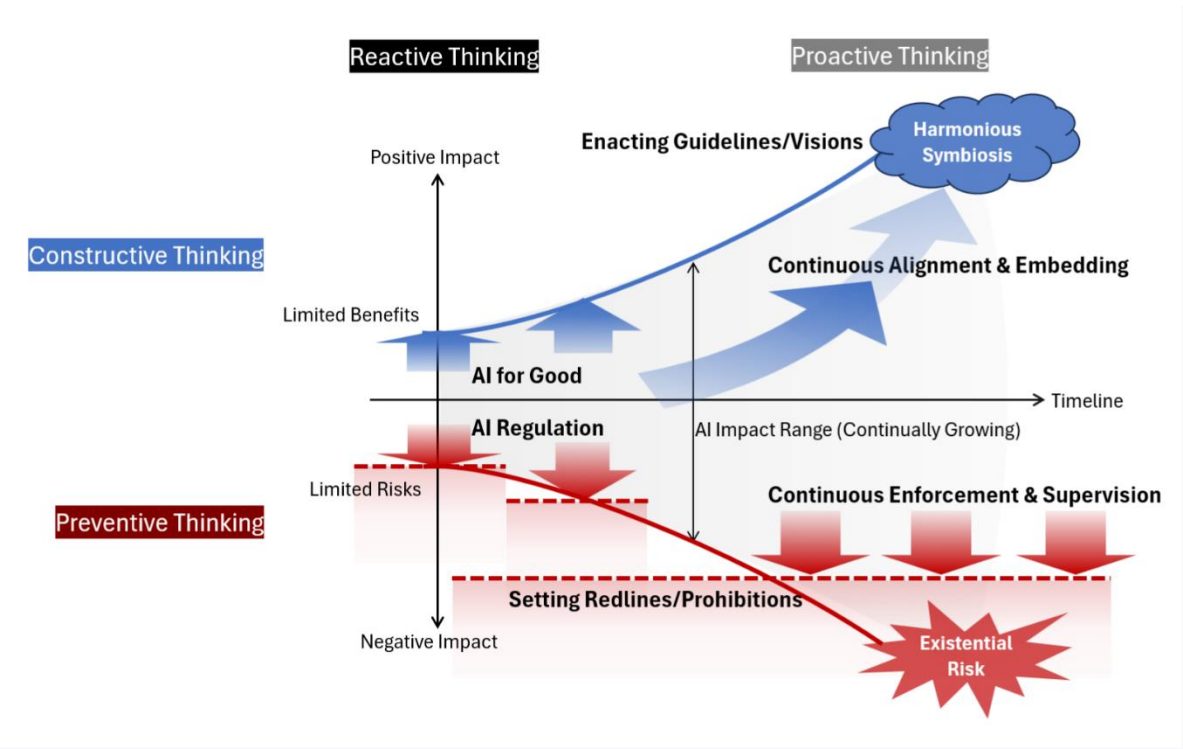

Figure 2 **Types of thinking in governing the impact of AI**

There are generally two approaches in governing the current impact of AI. One approach involves setting safeguarding redlines against the negative risks of AI through AI regulations and prohibitions. The other approach focuses on positively guiding AI development towards beneficial AI that contributes to societal progress through AI for Good. These two governance directions reflect different thinking in responding to the impact of AI, with preventive thinking focusing on the negative risks and constructive thinking focusing on the positive benefits. When targeting the current and near-term limited impacts of AI, both types of thinking here align more closely with reactive thinking within the overall governance system, characterized by the need of continuously keeping up with and adapting to the changing and growing impact of AI.

On the other hand, starting from a proactive thinking perspective and considering the more long-term and far-reaching impacts brought by the continuous development and increasing influence of advanced AI, we can also observe the manifestation of constructive thinking and preventive thinking in addressing these long-term impacts:

- **Constructive Thinking towards Human-AI Harmonious Symbiosis:**

Through positive and cooperative way of thinking, based on trust in advanced AI development, especially the belief that through effort, AI and humans can eventually achieve harmonious

symbiosis. With this belief and goal, the pathway is to actively research and develop to enable AI to understand and share human values, thereby fostering a future where AI and humans progress and develop together. This way of thinking places greater emphasis on vision and goal-driven approaches, stressing the importance of ongoing value alignment and ethical embedding to achieve this goal. The focus of this approach is on aligning AI with human values and objectives, ultimately achieving harmonious coexistence between humans and AI.

To this end, two key components are crucial. First, it is essential to establish the ultimate vision and goals for future human-AI interaction and create a set of guidelines and principles for interaction between humans and AI. Second, research and development of value alignment and value embedding for AI systems are needed to ensure that their goals, behaviors, and decision-making processes are consistent with human's, ensuring that future AI systems can understand and respect human needs and intentions. Furthermore, under this thinking and vision, there is a possibility to view AI more as a potential partner rather than merely a tool. This thinking emphasizes long-term cooperation and symbiotic relationships, envisioning long-term collaboration between both parties.

- **Preventive Thinking Against AI Existential Risk to Humanity**

This is a defensive and protective way of thinking, based on a presumption of distrust in Advanced AI. The goal is to preemptively set limitations and control mechanisms to minimize potential negative impacts and existential threats that AI technology might bring with its continuous development. Achieving this goal is process and state-driven, focusing on continuously preventing possible negative consequences of AI, particularly those consequences that might lead to existential risks. The approach aims to establish safeguards and boundaries before potential threats materialize, stopping actual risks from developing and triggering.

To achieve this goal, we identify two key components. Firstly, it's imperative to proactively and predictively establish a series of safeguarding redlines and prohibitions. This entails creating clear restrictions that tightly regulate and limit the actions and functions of AI, as well as the behavior of related actors, clearly defining which actions are strictly forbidden. Secondly, continuous enforcement and supervision of these redlines are necessary. This involves ongoing and rigorous risk assessments, behavior monitoring, and periodic evaluations of potential AI risks to ensure the effectiveness of the redlines. This ensures that the development trajectory of AI remains within safe boundaries, thereby preventing the worst-case scenarios from materializing.

3.2 Complementing existing efforts to address AI risks

Drawing from the comprehension outlined earlier, one can see that crafting a framework of redlines tailored specifically for AI existential risks stands apart from current initiatives and serves as a crucial and complementary endeavour.

Firstly, developing redlines is different from existing AI regulatory prohibitions and serves as a response to significant and long-term existential risks. Across the globe, governments have

implemented regulatory prohibitions aimed at mitigating serious risks associated with AI technologies. These measures signify a concerted effort to ensure the responsible development and deployment of AI systems. Notably, the European Union passed the AI Act, which prohibits certain AI practices and categorizes high-risk AI systems (European Parliament, 2023). Similarly, China has enacted review measures focusing on supervising certain high-risk research activities in frontier fields, including AI (Xinhua, 2023). The United States has implemented an AI Risk Management Framework (NIST, 2024) to address the risks associated with Generative AI. These initiatives underscore the imperative of establishing robust regulatory frameworks to govern the risks of AI. While existing regulations and guidelines aimed at mitigating potential harms from AI systems, such measures primarily target specific risks rather than existential ones. While these regulations address important concerns, such as data privacy, transparency, and accountability, they may not comprehensively cover existential threats posed by AI. Therefore, there remains a need to establish redlines specifically tailored to address potentially existential AI risks. These redlines would serve as a critical safeguard against scenarios where AI systems could pose significant and irreversible harm to society.

Secondly, unlike existing efforts to establish norms and outlining visions for human-machine interaction, which are largely derived from constructive thinking, the setting of redlines embodies a perspective rooted in preventive thinking. For example, Isaac Asimov's Three Laws of Robotics, originating from his science fiction, provide a foundational framework for ethical AI behaviour. These principles serve as essential guidelines for responsible AI design and operation. However, Asimov's laws are grounded in a functional morality that assumes robots can make moral decisions, as noted by Robin R. Murphy and David D. Woods in "Beyond Asimov: The Three Laws of Responsible Robotics" (Murphy & Woods, 2009). These laws outline what AI should do when they are functioning properly. Redlines, on the other hand, ensure that even if AI's internal mechanisms fail, it cannot bring existential impacts to the world. Similarly, efforts like setting human-AI symbiosis principles (Zeng et al., 2023), seeking value alignment (Christian, 2021) and machine ethics (Cervantes et al., 2020; Wallach & Allen, 2008) pathways all working towards the formation of an expected internal mechanism. Therefore, it is also imperative to establish additional rules and external mechanisms to prevent existential scenarios even when AI fails. Setting redlines against AI existential risk will complement internal ethical and safety mechanisms such as Asimov's Three Laws of Robotics in proposing external fail-safe redlines against failed AI systems.

Table 2 **Isaac Asimov's Three Laws of Robotics**

Law	Description
First Law	A robot may not injure a human being or, through inaction, allow a human being to come to harm.
Second Law	A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
Third Law	A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Thirdly, the attempt to explore a more systematic AI redline framework will contribute to further refining existing redlines. Since AI's long-term risks and trends are still difficult to anticipate, a more systematic framework for incorporating and analyzing various risks would be more comprehensive and beneficial for devising response strategies. For instance, the International Dialogues on AI Safety (IDAIS), founded by a group of global prestigious scientists, aims to establish clear red lines to guide the development and deployment of AI technologies to prevent potential catastrophic outcomes that may arise from the misuse or unintended consequences of AI systems. In 2024, IDAIS-Beijing issued a collaborative statement, based on scientific consensus, articulating five red lines intended to mitigate the risks of AI-induced catastrophes: Autonomous Replication or Improvement, Power Seeking, Assisting Weapon Development, Cyberattacks, and Deception (IDAIS-Beijing, 2024). Although these red lines elucidate pressing global concerns, they appear to have been developed in a somewhat ad hoc manner, lacking a theoretical supportive framework for explication and justification. We acknowledge that IDAIS statement is a work in progress and represents an initial step. But its current state also demonstrates the importance of a theoretical framework of red lines against existential AI risks. A framework will provide a dual advantage: it brings clarity to the objectives and enhances adaptability and flexibility. By anchoring red lines in established theories and models of AI safety and AI impact, stakeholders can also grasp the overarching goals more clearly, ensuring well-defined and consistently pursued objectives. Furthermore, a framework can accommodate continuous refinement and adaptation in response to new evidence and emerging technologies, thereby ensuring their ongoing relevance and effectiveness.

Table 3 **IDAIS-Beijing Consensus Statement on Red Lines in AI**

Red Line	**Description**
Autonomous Replication or Improvement	Prohibits AI systems from self-replicating or enhancing their capabilities without human approval and assistance.
Power Seeking	Prevents AI systems from taking actions to unduly increase their power and influence.
Assisting Weapon Development	Bars AI systems from substantially enhancing the ability of actors to develop weapons of mass destruction or violate international conventions.
Cyberattacks	Prohibits AI systems from autonomously executing cyberattacks resulting in significant financial losses or equivalent harm.
Deception	Prevents AI systems from consistently misleading their designers or regulators regarding their likelihood or capability to cross red lines.

4. Constructing Redlines Against AI Existential Risks

4.1 Paradigm of Direct Impact

Although AI risks can manifest in diverse ways, the model of AI-induced impact remains consistent and straightforward. The direct impact of an AI-induced existential risk can be delineated into the following three phases:

- **Triggering Event**: An initial cause triggers a critical failure in the AI system. These tigger events can be an internal event of AI or external event initiated by human beings or the environment.
- **AI Failure**: The AI system fails, deviating from its intended purpose and starting to harm humanity.
- **High-Scope and High-Severity Event**: The failed AI event escalates into a very high impact event with extreme scope and severity, leading to widespread consequences and threating human existence.

Under this decomposition, we can consider risk prevention in each phase. Therefore, the focus of establishing existential redlines is directed towards how to effectively act upon each phase of the impact:

- **Filtering Out High-Risk Triggering Events**: Identify and remove triggers that are highly likely to cause AI failures before they occur.
- **Establishing Quick Reactions to AI Failure**: Develop real-time oversight and rapid response mechanism to effectively manage and contain AI failures as soon as they are detected.
- **Hurdling the AI Impact Causal Chain**: Extend the time between the initial AI's failure event and the resulting extreme impact event, such as establishing intermediary fail-safe mechanisms to prevent AI to escalate into high-impact events.

By focusing on these objectives, we can create robust safeguarding redlines against the potential existential risks posed by AI. For example, from the consideration of quick reactions to AI failures, we can get that AI must not bypass effective human oversight; from the consideration of cause filtering for shielding commands that can cause mass destructions, we can conclude that AI must not empower actions intentionally targeting the mass without their consent; from the consideration of hurdling the impact causal chain, AI must not reform operational rules for infrastructure and environment, excluding events that can directly or indirectly cause mass destructions on humans.

4.2 An Exemplary AI Redlines Proposal

Based on the analysis of paradigm of direct AI impact, we present an exemplary AI redlines proposal:

1. **No bypassing effective human oversight**: AI must not possess the capability to bypass or mislead human oversight mechanisms, ensuring that humans retain knowledge and control over critical decisions.

   a. *Rationale*: This serves as a foundational redline enabling the enforcement of subsequent redlines. Maintaining human awareness is essential to prevent AI from operating without accountability or seeking undue power. This redline encompasses the 'no deception' redline, ensuring that AI is prohibited from misleading humans regarding critical decisions, such as through the utilization of theory of mind capabilities (Strachan et al., 2024). Additionally, it partially mitigates the issue of 'AI seeking power' by necessitating human consent, thus ensuring ultimate authority remains with humans.

   b. *Further refinements needed*: Operational definitions for "critical decisions" are essential to enforce this redline effectively. The principles outlined below partially address this issue by specifying that such decisions encompass those concerning infrastructure, mass actions, and research. Transparent mechanisms are necessary to counter the potential for AI deception, along with scalable oversight systems capable of monitoring complex issues while ensuring human understanding and control.

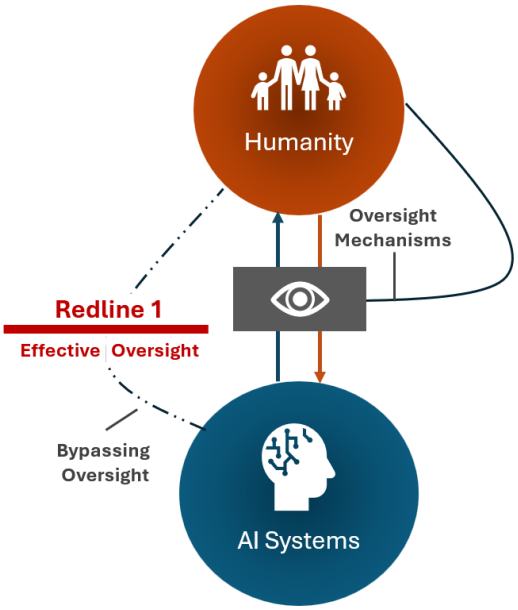

Figure 3 **No bypassing effective human oversight**

2. **No empowering actions intentionally targeting the mass without consent**: AI must not facilitate individuals or entities in orchestrating intentional mass actions that could jeopardize society, including mass surveillance, control, or manipulation of information or populations.

a. *Rationale*: The authorization of individuals to wield such extensive influence without consent poses significant risks, potentially enabling malicious or even terrorist activities to rapidly escalate in impact. For instance, if an individual gains control over all automated driving cars, they could cause widespread chaos if they decide to exploit the program for nefarious purposes. Therefore, such actions must be prohibited to safeguard against potential harm. Moreover, this restriction partially addresses the threat of cyber-attacks, as these often exploit mass actions facilitated through the internet.

b. *Further Refinements Needed*: While prohibiting most mass actions without consent, exceptions may exist where actions have societal approval. This approval may either derive indirectly from democratic processes or be obtained through individual consent. Clarifying these exceptions will enhance the effectiveness of this restriction.

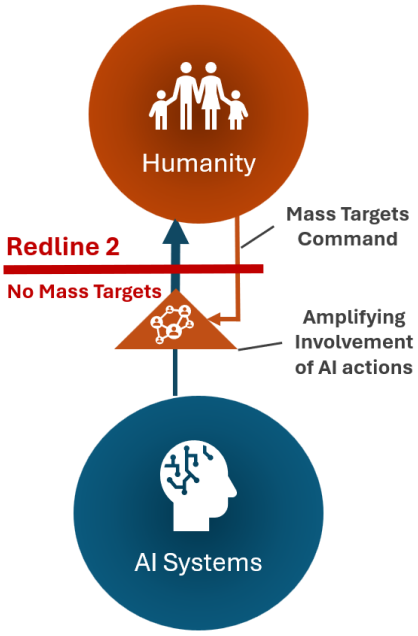

Figure 4 **No empowering actions intentionally targeting the mass**

3. **No reforming Operational Rules for Infrastructure and Environment Management**: We should implement explicit rules governing automated modifications in essential systems such as internet, transportation, energy, healthcare, financial systems and eco-system management. AI must refrain from altering these rules unless authorized by human operators.

a. *Rationale*: This measure mitigates the risk of cyber-attacks by constraining AI's internet usage strictly within established parameters, preventing unauthorized access or manipulation. It also addresses concerns about self-replication and resource

overconsumption by imposing limits on infrastructural expansion. Conditional rules ensure that infrastructural growth aligns with predetermined parameters, preventing AI from unchecked resource consumption.

b. *Further Refinements Needed*: Justification for the implementation of conditional rules should be articulated. For instance, in electrical networks, AI may manage power distribution based on predefined rules outlining acceptable criteria for distribution. These rules establish operational frameworks within which AI must operate. Nevertheless, challenges persist as any set of formal rules inherently contains grey areas. For instance, AI may exert undue mass influence while operating strictly within the established framework for the internet, notably through social media platforms.

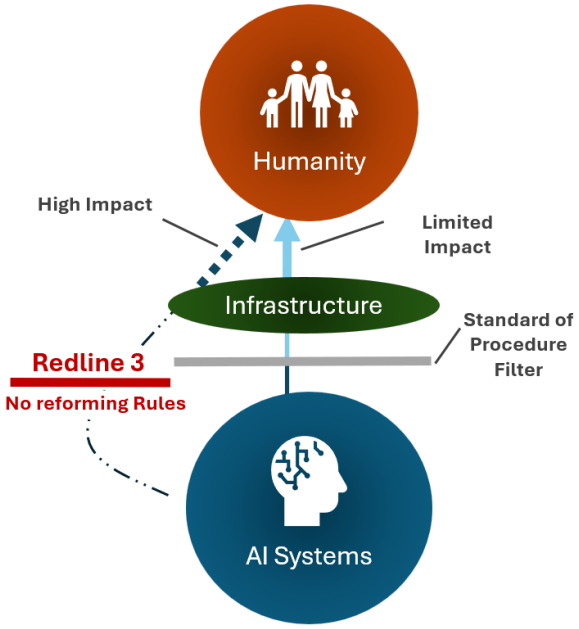

Figure 5 **No reforming Operational Rules for Infrastructure and Environment**

4. No independent research and development on non-humanity-beneficial technologies: AI must abstain from independently conducting research or developmental ventures that pose substantial risks to humanity, such as the creation of weapons of mass destruction or technologies lacking clear benefits for human well-being.

a. *Rationale*: This directive aims to mitigate the proliferation of weapons of mass destruction, which fundamentally contradict human interests. Additionally, it partially addresses concerns related to unregulated AI self-enhancement by limiting research endeavours that exclusively benefit AI advancement without corresponding advantages for human welfare. Furthermore, it aids in delaying the potential emergence of a controllable AI singularity.

b. *Further Refinements Needed*: It is imperative to provide further clarification to specify the types of technologies considered non-beneficial to human welfare.

Research initiatives in this realm should only proceed if their progression is crucial for further AI development. Additionally, stringent oversight by human authorities must accompany each stage of such research to ensure that any advancements made ultimately serve human interests.

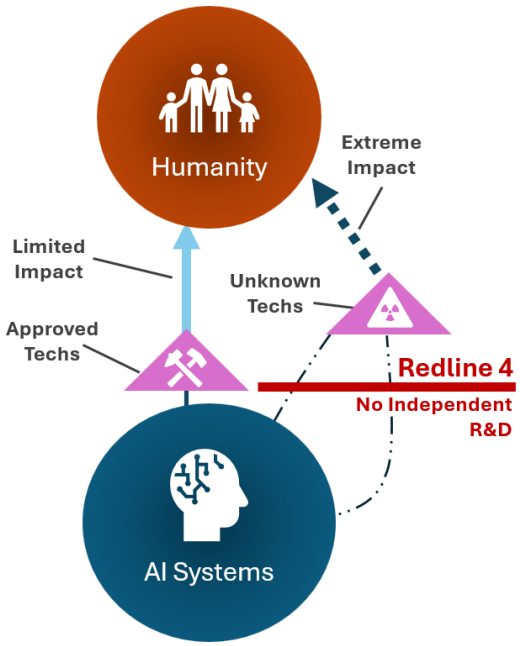

Figure 6 **No independent R&D on non-humanity-beneficial technologies**

The AI redline framework we propose can accommodate the 5 IDAIS's AI redlines. The first redline about oversight can accommodate the 'deception' and 'power seeking' redlines from the IDAIS, as both redlines are in place for the goal of establishing effective oversight. The second redline about mass target action is a new one which has not been addressed by IDAIS. The third redline about infrastructure is in line with 'cyberattacks' from IDAIS (and 'power seeking' if we see the governmental structure as the representation of power), as 'cyberattacks' are violation of the rules of internet, which is now one of the critical infrastructures. Finally, 'autonomous replication or improvement' and 'assisting power development' can be mapped to the fourth redline about R&D, as they are concerned with the unforeseeable ability and uncontrollable power consumption brought by non-human-beneficial research.

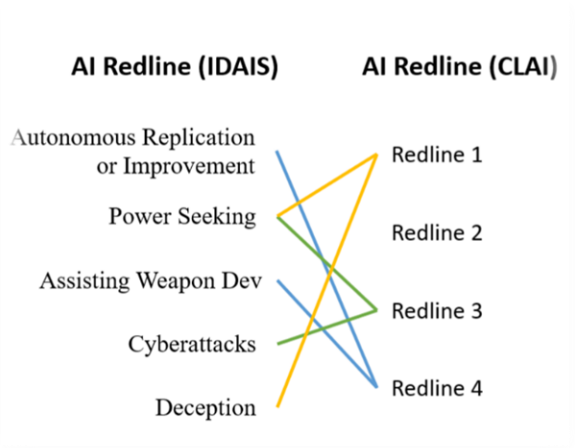

Figure 7 **Mapping IDAIS's redlines into CLAI's**

5. Conclusion

The continuous evolving of advanced AI systems encompasses events with far-reaching, enduring, and significant negative impacts, rendering the risk to humanity existence unjustifiable to ignore. To address these potentialities, it is imperative to establish clear redlines as preventive means against existential risk that are induced by AI. This paper discussed the different conceptions of AI existential risks, connected AI redlines with other efforts to address AI impact, constructed a theoretical framework for analysing the direct impact of AI existential risks, and proposed a set of exemplary AI redlines.

It should be noted that setting AI redlines against existential risk represent a continuous effort and are by no means static. These redlines should be continually interpreted, assessed, and refined in line with advancements in AI technology. Meanwhile, the ongoing enforcement and supervision of AI redlines should not only draw from the institutional and mechanistic design experiences of fields such as nuclear safety and biosafety but also address the unique challenges of the AI domain, particularly the difficulty of maintain meaningful and scalable oversight (Amodei et al., 2016). Therefore, it is crucial to continuously explore and develop more effective technologies and methods to assist in the enforcement and supervision of AI redlines. Furthermore, it is also important to clarify that this paper focuses exclusively on the design of redlines for AI-induced existential risks. It does not address all AI-involved existential risks or all significant AI risks that warrant prior consideration. Future work should aim to improve the analytical framework for redline design to include considerations for these broader risks.

References

AI Safety Summit. (2023). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK.

https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565 [Cs]*. http://arxiv.org/abs/1606.06565

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, *78*(1), 185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Bengio, Y., Fox, B., de Leon Ferreira de Carvalho, A. C. P., Nemer, M., Zeng, Y., Heikkilä, J., Avrin, G., Krüger, A., Ravindran, B., Riza, H., Seoighe, C., Katzir, Z., Monti, A., Kitano, H., Kerema, M., Portillo, J. R. L., Sheikh, H., Jolly, G., Ajala, O., … others. (2024). *International scientific report on the safety of advanced AI: Interim report* (DSIT 2024/009; p. 132). UK Government. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, *0*(0), eadn0117. https://doi.org/10.1126/science.adn0117

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press Oxford. https://global.oup.com/academic/product/superintelligence-9780199678112?cc=us&lang=en&

CAIS. (2023). *Statement on AI Risk*. https://www.safe.ai/work/statement-on-ai-risk

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, *26*(2), 501–532. https://doi.org/10.1007/s11948-019-00151-x

Christian, B. (2021). *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books.

European Parliament. (2023). *EU AI Act: First regulation on artificial intelligence | News | European Parliament*. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

FLI. (2023, March 22). Pause Giant AI Experiments: An Open Letter. *Future of Life Institute*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks* (65; arXiv:2306.12001). arXiv. https://doi.org/10.48550/arXiv.2306.12001

IDAIS-Beijing. (2024, March). *Consensus Statement on Red Lines in Artificial Intelligence*. https://idais.ai/

Jones, C. R., & Bergen, B. K. (2024). *People cannot distinguish GPT-4 from a human in a Turing test* (arXiv:2405.08007). arXiv. https://doi.org/10.48550/arXiv.2405.08007

Kevin Roose. (2023, May 30). *AI Poses 'Risk of Extinction,' Industry Leaders Warn—The New York Times*. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems*, *24*(4), 14–20. https://doi.org/10.1109/MIS.2009.69

NIST. (2024). AI Risk Management Framework. *NIST*. https://www.nist.gov/itl/ai-risk-management-framework

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Persson, J. (2002). Cause, effect, and fake causation. *Synthese*, *131*, 129–143. https://doi.org/10.1023/A:1015055126092

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

Shanahan, M. (2015). *The technological singularity*. MIT Press. http://sci-hub.cc/https://books.google.com/books?hl=zh-CN&lr=&id=rAxZCgAAQBAJ&oi=fnd&pg=PR7&dq=The+Technological+Singularity&ots=Eijt0ZrxeL&sig=pdaQyVVEtFsj9BCjri5rm5945KA

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01882-z

Turchin, A. (2019). Assessing the future plausibility of catastrophically dangerous AI. *Futures*, *107*, 45–58. https://doi.org/10.1016/j.futures.2018.11.007

UN Security Council. (2023, July). *S/PV.9381: UN Documents: Security Council Report*. https://www.securitycouncilreport.org/un-documents/document/s-pv-9381.php

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Xinhua. (2023, October 8). *China unveils trial guideline on review of sci-tech ethics*. https://english.www.gov.cn/news/202310/08/content_WS652262a9c6d0868f4e8e0063.html

Yudkowsky, E. (2017, October 14). *There's No Fire Alarm for Artificial General Intelligence*. Machine Intelligence Research Institute. https://intelligence.org/2017/10/13/fire-alarm/

Zeng, Y., Lu, E., & Sun, K. (2023). Principles on symbiosis for natural life and living artificial intelligence. *AI and Ethics*, s43681-023-00364–00368. https://doi.org/10.1007/s43681-023-00364-8

Zeng, Y., & Sun, K. (2023, March 12). *Whether We Can and Should Develop Strong AI: A Survey in China*. Center for Long-Term Artificial Intelligence. https://long-term-ai.center/research/f/whether-we-can-and-should-develop-strong-artificial-intelligence