

Research Article

Using Probabilistic Methods to Study Technological Extinction Is Unscientific, Meaningless, and Harmful

Hu Jiaqi^{1,*}

*Correspondence: Hu Jiaqi, hujiaqi1122@163.com

© The Author(s) 2026.

This article is published by Wisdom Academic Press Ltd. under a Creative Commons Attribution 4.0 International License. The license permits use, sharing, adaptation, distribution, and reproduction in any medium or format, provided that appropriate credit is given to the original author(s) and the source, a link to the license is supplied, and any modifications are indicated. Unless otherwise stated, all images and third-party materials included in this article are also covered by the same license. For any material not covered by this license, if the intended use is neither permitted by applicable law nor allowed under the license terms, direct permission must be obtained from the copyright holder. To view a copy of this license, please visit <http://creativecommons.org/licenses/by/4.0/>.

Abstract

In this paper, the term “technological extinction” denotes the total biological extinction of the human species—leaving zero survivors—resulting from scientific and technological means. The term “unchecked development of science and technology” signifies the advancement of science and technology in the absence of effective oversight, permitting it to move forward indefinitely without directional control. Humanity is currently caught in a systemic extinction risk driven by the unchecked development of science and technology. The structural contradiction between the infinite destructive capacity of technological iteration and humanity’s limited governance capability has become the core bottleneck restricting the long-term survival of human civilization. Researchers specializing in technological existential risks, represented by Nick Bostrom and Toby Ord, have adopted mathematical modeling and probabilistic quantification as their core paradigm for assessing humanity’s ultimate extinction crisis and guiding global technology governance policies. They have successively proposed quantitative conclusions such as a 25% probability and a 1/6 probability of human extinction by technology within this century (Bostrom, 2013; Ord, 2021), profoundly influencing global technology crisis governance and top-level public policy-making. This paper clearly distinguishes the core value boundaries of two types of probabilistic assessments: the localized risk probability assessments made by front-line technology practitioners and domain-specific technical experts on individual technology tracks are grounded in industry practice and aligned with technological patterns, possessing sufficient industry reference value and technical rationality. However, the academic community engaged in holistic research on humanity’s technological crises and ultimate existential risks has elevated probabilistic quantification methods to the core basis for guiding global technology crisis governance and top-level public policy-making—a move that suffers from fundamental and insurmountable defects. The core model parameters of this paradigm rely on

expert subjective judgment and empirical assessment, resulting in widely divergent scholarly estimates, no unified scientific standards, and no objective empirical support—completely lacking the rigor and reliability required to guide national or global governance decisions. Based on five core axioms—the constancy of human nature, the one-way increase of technological destructive power, the irreversibility of group competition, the perpetual lag of defense behind destruction, and the inherent uncertainty of technology—and relying on the research framework of “extinction path analysis + defense-limit testing,” (Hu, 2018) this paper demonstrates through closed-loop reasoning that the extinction of human civilization caused by unchecked technological development is a 100% inevitable outcome, with no room for probability fluctuation. The risk perception and governance orientation formulated by dedicated crisis researchers based on the probabilistic paradigm not only violate core scientific principles but also completely deviate from the practical realities of global governance. Moreover, they numb global crisis consensus, misguide technology control policies, and delay humanity’s critical window for self-rescue—posing a significant threat to the continuation of human civilization. Therefore, this paradigm is absolutely unsuitable for top-level governance and policy guidance regarding ultimate technological crises.

Keywords

Technological extinction, Human survival, Human extinction, Probabilistic research paradigm, Nick Bostrom, Toby Ord, Extinction Path Analysis, Defense Limit Testing, Human existential crisis, Global technology governance policy

1. Introduction

The author has been engaged in research on technological crises and human survival security for nearly half a century. After twenty-eight years of systematic refinement, in 2007 he published the 800,000-character mono-graphic study *Saving Humanity*, establishing a dedicated research framework of “extinction path analysis + defense-limit testing.” Combined with five fundamental axioms, this framework has undergone rigorous multi-level deduction to reach a definitive core conclusion: if humanity continues to allow the unchecked, disorderly iteration of science and technology, then—at the earliest within this century, and at the latest within two to three hundred years—it will inevitably face overall species extinction (Hu, 2018). This conclusion is not subjective speculation or alarmism, but an objective prediction grounded in the evolutionary patterns of human civilization, the logic of social strategic interaction, and the essential nature of technological iteration. It contains no probabilistic ambiguity or uncertainty.

Currently, the field of global technological existential risk research suffers from a pronounced problem of paradigm misapplication. The core controversy does not lie in the technical form of

probability assessment itself, but in the severe overreach of its applicable scenarios and application levels. In specific technology domains such as artificial intelligence, synthetic biology, nanotechnology, and nuclear technology, the probabilistic predictions made by front-line researchers and technical experts—based on long-term industry accumulation—are professionally grounded assessments aligned with technological iteration patterns and practical experience. They can effectively guide industry-level technology risk control, R&D norms, and safety optimization, possessing strong contextual rationality and application value. The risk assessments of industry-leading practitioners such as Geoffrey Hinton and Elon Musk, as well as the 2023 survey data from AI Impacts involving over 2,700 global AI researchers and the industry predictions from the Metaculus global forecasting platform (BBC Radio 4, 2024; Joy, 2000; Metaculus, 2018; PowerfulJRE, 2025), all fall within the category of domain-specific technical risk references that do not bear the responsibility of global governance. Their assessment logic and application value are worthy of recognition.

However, dedicated researchers of human technological crises, represented by Bostrom and Ord (Armstrong et al., 2013; Bostrom, 2013; Cotton-Barratt & Ord, 2015; Ord et al., 2010; Ord, 2021), have deliberately blurred the contextual boundaries of probabilistic assessment. They have elevated probabilistic methods—which are only suitable for domain-specific, internal industry reference—into the core research framework for assessing humanity’s ultimate existential crisis, guiding global technology governance policies, and building international risk prevention systems. This essentially deviates from the scientific pursuit of truth and practicality, and violates the rigor required for global public governance. Such research appears to present an objective and rational academic posture through quantified data, but in reality, it uses complex mathematical models to package subjective speculation, deliberately avoids the core structural contradictions of technological extinction crises, and weakens an absolutely inevitable civilizational survival crisis into a random probabilistic event. This not only misguides the policy decisions of nations and international organizations but also continuously undermines the formation of a unified global crisis consensus and delays humanity’s critical window for self-rescue. At the fundamental level, the extinction of human civilization is a unique, non-replicable ultimate public event concerning the survival fate of all humanity, which cannot be measured by conventional probabilistic statistical rules. The probabilistic models built on subjective value assignments, when used by dedicated crisis researchers to guide top-level governance, completely forfeit the objectivity, rigor, and reliability that public policy research should possess.

The author once regarded Bostrom as an academic peer, after he proposed the core view that “science and technology are the primary trigger for human extinction, and humanity may face the end of civilization within the next century.” (Bostrom, 2002, 2013, 2019, 2025) However, after long-term engagement with his research framework, the author found that his core deviation lies not in localized

risk prediction, but in the erroneous application of probabilistic paradigms to global crisis assessment and policy guidance, progressively weakening the inevitability of technological extinction crises, replacing deterministic crisis judgment with fluctuating probabilities, and severely misleading global technology governance orientation. Based on this, this paper precisely defines the boundaries of its inquiry: it fully acknowledges the industry-level rationality of front-line experts' localized probabilistic assessments, and only specifically critiques the serious fallacy of dedicated technological crisis researchers who abuse the probabilistic paradigm to guide global crisis governance and public policy. The paper deconstructs the core loopholes of this paradigm from three dimensions—unscientific nature, meaninglessness, and harmfulness—elucidates the core advantages of the “extinction path analysis + defense-limit testing” scientific paradigm, demonstrates the absolute inevitability of technological extinction, and calls upon the global academic community and governance levels to abandon probabilistic, ambiguous governance approaches and establish a crisis prevention and control system that aligns with objective laws and is suited to top-level policy design.

2. The Unscientific Nature of Applying Probabilistic Methods to Top-Level Crisis Governance Research: Deviating from the Scientific Principles of Public Policymaking

Objective rigor, verifiability, reproducibility, and falsifiability are the core characteristics of modern scientific research, and are mandatory criteria for guiding public policy and global governance decisions. Probabilistic assessments in specialized technical fields serve industry-level technology optimization and localized risk prevention, with relatively large margins for error, and reliance on heuristic judgment is reasonable. However, top-level governance research oriented toward humanity's overall survival and global technology control—whose conclusions directly determine the fate of all humanity and influence the policy directions of nations—must be derived from invariant natural laws and objective facts, and must never rely on subjective speculation or ambiguous conclusions. The fundamental defect of applying probabilistic research paradigms to the overall assessment and policy guidance of ultimate crises lies in the fact that the entire system is built on subjective conjecture, devoid of objective data, constant laws, and unified verification standards, rendering it entirely inadequate for the ultimate public governance issue of human civilization extinction. (Ord et al., 2010; Russell, 2015; Tegmark & Bostrom, 2005; Yudkowsky, 2008)

The “extinction path analysis + defense-limit testing” research method adopted in this paper strictly relies on five core axioms for its argumentation. These five axioms are invariant natural laws repeatedly verified throughout the history of human civilization development and technological iteration, constituting the rigid theoretical foundation for top-level research and governance design on technological existential risks (Hu, 2018). Specifically: First, the Axiom of the One-Way Increase of

Technological Destructive Power. Technological iteration is inherently an upgrade of tools and destructive capacity, exhibiting only a one-way strengthening trend, with no possibility of technological regression, weakening of destructive power, or spontaneous stagnation. This provides the eternal basis for the continuous emergence of new extinction means and the continuous escalation of risk levels. Second, the Axiom of the Constancy of Human Nature. Humanity's core instincts of self-interested competition, vengeance, and hatred cannot be eradicated through morality, institutions, or law. Only surface-level behavior can be regulated; the underlying logic remains immutable. Third, the Axiom of the Irreversibility of Group Competition. Competition among nations, ethnic groups, and interest blocs is the normal state of human civilization. Technological competition, rivalry, and covert arms races are permanent features. Fourth, the Axiom of the Perpetual Lag of Defense Behind Destruction. Destructive technologies are low-cost, easy to proliferate, and quick to deploy, while corresponding defense and countermeasure technologies have long development cycles, high thresholds, and are passive and lagging, creating intrinsic and ineliminable risk windows. Fifth, the Axiom of the Inherent Uncertainty of Technology. The potential risks of unknown technological domains are difficult to predict in advance, and many seemingly benign technological breakthroughs may generate fatal secondary risks. These five axioms are mutually reinforcing and form a closed-loop, self-consistent system, serving as the fundamental prerequisite for assessing humanity's ultimate existential crisis and designing global governance systems. In contrast, the probabilistic research paradigm adopted by dedicated crisis researchers completely departs from scientific theoretical foundations and governance practice, relying throughout on researchers' personal experience and subjective value assignments, producing widely divergent scholarly estimates with no unified academic consensus, and possessing no scientific qualification to guide global policy.

2.1. Model Parameters Are Entirely Subjectively Assigned, Unable to Support Authoritative Public Decision-Making

Probabilistic methods constitute a widely employed approach in scientific inquiry, and the tools themselves are inherently unproblematic. These methods are generally suited to contexts where there are abundant historical observations and repeated validation—such as the estimation of seasonal climatic changes in a region (e.g., spring, summer, autumn, winter) or the assessment of probability distributions in coin-tossing experiments. The core logic of probabilistic risk research is to complete risk quantification through mathematical model parameter assignment. However, technological extinction is the ultimate public event of human civilization—there are no historical extinction samples, no controllable experimental data, and no unified global measurement standards. In the global risk models constructed by dedicated researchers, all core parameters are subjective personal judgments and expert empirical assignments. This is an inherent defect that the paradigm cannot

overcome, and the core reason why it can never be used to guide global governance policy. Bostrom's existential risk analysis framework lists artificial intelligence, nuclear war, synthetic biology, and nanotechnology as core extinction risks (Bostrom, 2002, 2011, 2013, 2019, 2025), yet it cannot specify the basis for the probability values assigned to each type of risk—there is no historical data, experimental results, or objective laws to support them. It is essentially personalized empirical conjecture, rather than an actionable governance basis. A comprehensive review of the mainstream estimates in the global dedicated crisis research field reveals that the conclusions of different top scholars and authoritative institutions vary enormously and conflict with each other, fully proving that such probability parameters have no objective standards and no reference value for public decision-making. The mainstream estimates in the field are as follows:

First, Oxford philosopher Toby Ord, in his book *The Precipice: Existential Risk and the Future of Humanity*, systematically qualifies the risk of global catastrophic events within this century, proposing that the probability of technology-induced ultimate extinction is approximately 1/6 (16.7%) (Ord, 2021), and further disaggregated the proportions of various risk categories, attempting to use this data as the basis for governance recommendations such as global technology regulation and international risk cooperation. His estimate appears meticulous but is in fact entirely without empirical support or legal foundation—it is merely a personal experience-based numerical disaggregation. Yet it has been cited extensively in international governance reports and has become a reference basis for top-level policy design, posing serious governance risks.

Second, Nobel laureate Geoffrey Hinton, hailed as the “godfather of AI,” and other front-line technology practitioners have predicted a 10–20% (BBC Radio 4, 2024) probability of AI loss of control within the next 30 years. Such judgments focus on the iterative risks of the technology itself and are used only for industry-level technology safety R&D reference, without involving global governance or public policy. These are reasonable domain-specific empirical assessments and do not constitute paradigm misuse.

Third, technology entrepreneur Elon Musk has proposed a 20% (PowerfulJRE, 2025) probability of AI-induced destruction. This judgment is grounded in frontier technology deployment scenarios, serves industry-level technology risk control and industrial norms, and is a typical localized technical assessment with contextual rationality, bearing no responsibility for global governance.

Fourth, dedicated crisis research teams led by Nick Bostrom, founder of the University of Oxford's Future of Humanity Institute, have published authoritative reports stating that the probability of human extinction by technology within this century is no less than 25% (Bostrom, 2002, 2013, 2025; Tegmark & Bostrom, 2005), and have used this as the core basis for guiding global AI safety governance and high-risk technology control policies, while subjectively asserting that risk probabilities can be reduced through international cooperation and technological optimization,

arbitrarily imposes governance decisions with considerable randomness, seriously departing from objective laws.

Fifth, survey data from platforms such as Metaculus and AI Impacts (Joy, 2000; Metaculus, 2018) aggregate the localized technical risk judgments of front-line practitioners, which are only suitable for Intra-industry risk warning. However, dedicated researchers have magnified these into global civilizational extinction probabilities, directly using them to derive global governance policies—a classic case of contextual misuse and hierarchical mismatch.

Sixth, a 2025 special study by MIT dedicated risk researcher Max Tegmark concluded that the probability of humanity losing the AI competition and ceding civilizational dominance exceeds 90% (Milmo, 2025). The model parameters and logical assumptions of this study are purely personal academic speculative innovations, with no industry-wide verification standards. Yet they have been incorporated into the risk assessment systems of some international organizations, misleading top-level risk perception.

A comprehensive review of industry data reveals that the global extinction probabilities given by dedicated crisis researchers range from 9% to over 90% (BBC Radio 4, 2024; Bostrom, 2013; Joy, 2000; Metaculus, 2018; Milmo, 2025; Ord, 2021; PowerfulJRE, 2025)—an extremely wide span, mutually contradictory, with no unified academic consensus or objective value-assignment standards. Once such subjective estimates are elevated to the basis for global public policy, they directly lead to fragmented governance standards and disjointed risk perceptions among nations. In stark contrast, the localized probability assessments of front-line technical experts serve only domain-specific technological optimization, with clear boundaries and contextual appropriateness, and thus possess full rationality and reference value.

Thus, the core boundary can be clearly defined: the probabilistic judgments of domain-specific practitioners are experience-based technical references with no public governance orientation—reasonable and usable. The global probabilistic estimates of dedicated technological crisis researchers are subjective academic conjectures, yet they are used to guide global governance and top-level policy—completely violating the objectivity and rigor of public decision-making, and possessing no scientific validity whatsoever.

2.2. Misapplication of the Applicable Boundaries of Probability Theory, Unsuitable for Global Governance Decision-Making Contexts

Probabilistic statistics has strict preconditions and boundary conditions. They are only applicable to repeatable, statistically quantifiable, and sufficiently sample-accumulated routine events, deriving effective values through the frequency patterns of many repeated experiments, and can then be used for localized risk prediction. Equipment failures, technological loss of control, and experimental risks

in specialized technical fields are repeatable, observable routine risks that are well-suited to probabilistic assessment logic. Therefore, the localized probabilistic judgments of front-line experts are scientifically sound and reasonable.

However, the extinction of human civilization is a unique, non-replicable, irreversible ultimate global public event—with no historical extinction samples, no repeatable experimental contexts, and no extraterrestrial civilizational reference systems—completely beyond the applicable scope of probability theory. Dedicated crisis researchers deliberately conflate the boundary between “localized technological risk” and “global civilizational extinction risk,” forcibly applying conventional probabilistic models to ultimate crisis governance research. This constitutes a serious theoretical misapplication and conceptual substitution. For example, Toby Ord draws on the physical probability of asteroid impacts to derive the probability of human civilization extinction (Ord et al., 2010; Ord, 2021), conflating natural random events with the systemic risks of human technological competition and civilizational survival, transforming a narrow physical probability into a basis for global governance, severely misleading global risk perception and policy-making.

In short, probabilistic methods are suited to localized, repeatable, technical-level risk prediction—the application by front-line experts is reasonable and compliant. However, they are entirely unsuited to global, unique, irreversible, public-governance-level civilizational extinction assessment. The cross-disciplinary misuse by dedicated researchers is the core source of this paradigm’s harm to global governance.

2.3. Conflating Probabilistic Risk with Inevitable Outcome, Misleading Top-Level Governance Logic

The most fatal governance fallacy of the probabilistic research paradigm is its distortion of the inherent trajectory of human extinction—caused by unregulated technological development—from an objective inevitability into a controllable, avoidable random probability event. While this cognitive bias may be inconsequential for localized industry risk management, when elevated to the core guiding logic of global technological crisis governance, it directly subverts the foundational orientation of top-level policy frameworks. This entraps nations and the international community in the erroneous belief that “risks can be fortuitously avoided and governance pursued through a gradualist approach.”

This paper, grounded in the “Extinction Path Analysis + Defense Limit Testing” framework and supported by five constant axioms, empirically demonstrates that the destructive power of technology is infinitely iterative and without upper bound; that human defense systems are perpetually lagging and inherently limited; and that, combined with the invariant patterns of human nature (game-playing, group competition, hatred and revenge) and technological uncertainty, the unchecked development of

technology will inevitably breach the boundaries of human control, leading to the total extinction of civilization. This conclusion carries 100% certainty, with no room for probabilistic variation.

The theories of Bostrom's AI value alignment and Ord's international cooperation to reduce risk (Armstrong, 2013; Armstrong et al., 2013; Bostrom, 2012, 2013; Cotton-Barratt & Ord, 2015; Orseau & Armstrong, 2016; Ord, 2021; Russell, 2014; Russell et al., 2005; Yudkowsky, 2001, 2016, 2022) are, at their core, idealized assumptions based on probabilistic thinking—believing that risks can be reduced through technological optimization, ethical constraints, and localized cooperation. Such ideas have value as references for industry-level technological optimization, but as the basis for global top-level governance policy, they are completely divorced from reality: the irreversibility of national interest competition, the impossibility of unifying human values, and the perpetual lag of defense behind destruction all determine that probabilistic flexible control and localized governance cannot avert the ultimate extinction crisis. Dedicated researchers' use of probability to weaken the inevitability of the crisis directly leads to a misalignment of global governance priorities and the evasion of core contradictions.

3. The Meaninglessness of Applying Probabilistic Methods to Guide Global Crisis Governance: Unable to Support Top-Level Decision-Making or Solve Fundamental Problems

Technological crisis research operates on two levels: first, domain-specific technical risk control research, whose core goal is to optimize technology and mitigate localized risks—probabilistic methods are suited to this scenario, and expert assessments have practical value. Second, holistic research on humanity's overall survival and global governance, whose core goal is to avert civilizational extinction, reconstruct global governance systems, and defend humanity's survival red-line. Its core mission is to reveal inevitable crises, provide fundamental solutions, and guide top-level policy—not to engage in meaningless probabilistic numerical games. Dedicated crisis researchers who apply probabilistic methods to the field of global governance completely evade core contradictions, depart from governance practice, fail to address humanity's most fundamental survival questions, and forfeit the core significance of ultimate risk research in serving public governance.

3.1. Floating Probabilistic Values Cannot Guide Unified Global Governance Decisions

From the perspective of public governance logic, the survival of human civilization holds absolute, infinite, and unique ultimate value. Top-level decisions concerning the fate of all humanity must be based on deterministic laws and rigid conclusions, rather than floating, subjective, and highly divergent probabilistic values. Regardless of whether the extinction probability given by dedicated

researchers is 9%, 16.7%, 25%, or 90% (Armstrong, 2013; Armstrong et al., 2013; Bostrom, 2013; Cotton-Barratt & Ord, 2015; Ord, 2021) as long as civilizational extinction is a real possibility, global governance must take “preventing it at all costs” as its absolute bottom line. The numerical fluctuation of probability values cannot change the fatal nature of the crisis, nor can it adjust the absolute priority of global ultimate risk prevention.

The localized probabilistic values of front-line practitioners are used for corporate R&D, industry norms, and technological iteration optimization—precisely matched to their contexts and with clear practical value. However, the global extinction probabilities that dedicated researchers expend extensive academic resources to calculate are widely divergent, mutually contradictory, and devoid of unified standards. They not only fail to guide global governance but actually create fragmented policy perceptions among nations.

3.2. Evading Core Structural Contradictions, Offering No Fundamental Governance Solutions

The core structural contradictions of humanity’s technological extinction crisis are the triple conflicts between the infinite destructive capacity of technological iteration and humanity’s limited control capability, between national interest competition and global unified risk control, and between individual self-interest and civilizational survival. The core mission of global crisis governance research is to confront these contradictions directly, address their root causes, and construct a fundamental solution system. The probabilistic research paradigm, however, evades these essential issues throughout.

The governance proposals put forward by dedicated researchers such as Bostrom and Ord—AI safety technology optimization, industry ethics improvements, fragmented international cooperation, localized risk regulation (Armstrong, 2013; Armstrong et al., 2013; Bostrom, 2013; Cotton-Barratt & Ord, 2015; Ord, 2021)—are all shallow end-of-pipe governance measures. They are only suitable for industry-level localized risk control optimization and are completely incapable of solving the global structural contradictions. In the context of irreversible national technological competition, probabilistic flexible control and sporadic risk prevention cannot halt the iteration and proliferation of high-risk technologies; they can only temporarily delay risks, with no fundamental solution value.

In contrast, the scientific research paradigm of this paper directly confronts the core contradictions. Through extinction path analysis and defense-limit testing, it empirically demonstrates the limitations of end-of-pipe governance, and ultimately proposes the only fundamental solution suited to global top-level governance: advancing the Great Unification of humanity and constructing a unified global technology governance system to curb unchecked technological development at its source. This deterministic conclusion is something that probabilistic ambiguous governance thinking

can never replace, and it represents the true core value of global crisis research.

3.3. Falling into Academic Formalism, Departing from the Original Purpose of Public Governance Research

Currently, the field of dedicated technological crisis research has developed a formalistic academic culture centered on probabilistic modeling and numerical calculation. Researchers no longer take “protecting the survival of human civilization and providing scientific governance solutions” as their core goal, but are instead oriented toward producing quantified data, publishing academic results, and monopolizing academic discourse. They unwarrantedly elevate probabilistic methods—applicable solely to niche technical domains—to the core paradigm of global governance, consuming vast research resources to conduct meaningless numerical games.

This paradigmatic misuse directly distorts the trajectory of global crisis research: the academic community becomes excessively preoccupied with subjective probability calculations, neglecting core research directions such as extinction path tracing, defense system innovation, and global governance restructuring. It reduces public governance research—which should serve all humanity—to a tool for advancing individual academic reputations and clique interests, completely deviating from the social responsibility and original public mission of ultimate risk research.

4. The Harmfulness of Applying Probabilistic Methods to Guide Global Crisis Governance: Misleading Global Perception and Top-Level Policy, Delaying Humanity’s Self-Rescue

The localized probabilistic assessments of front-line technical experts are confined to Intra-industry use and technological optimization, with no public harm. However, when dedicated crisis researchers apply the probabilistic paradigm to global technological crisis governance and public policy-making, it produces global-scale, fatal negative effects—directly numbing global crisis consensus, misleading national governance decisions, and delaying humanity’s window for self-rescue, posing an irreversible threat to the continuation of human civilization.

4.1. Weakening Global Crisis Consensus, Fostering Collective Complacency

Human nature inherently harbors complacency, and the floating probabilistic conclusions produced by dedicated researchers precisely cater to the risk-avoidance inertia and developmental demands of nations and societies. Medium-to-low probability conclusions such as 1/6 and 25% (Bostrom, 2013; Ord, 2021) lead the global public to form the erroneous consensus that “the extinction crisis is highly unlikely to occur,” weakening the absoluteness and urgency of the crisis. Meanwhile, high-probability conclusions such as 90% (Milmo, 2025), due to the enormous scholarly disagreement,

are dismissed as extreme views and fail to form global consensus. This probabilistic tiered misleading completely dissolves the internal motivation for unified global crisis prevention, allowing nations to prioritize technological development, national competition, and economic growth while relaxing top-level control over ultimate risks.

4.2. Providing False Scientific Justification for Global Unchecked Technological Competition

Against the backdrop of irreversible national interest competition, global technological rivalry already possesses internal momentum toward unchecked expansion. (Armstrong et al., 2013; Bostrom, 2003) The various medium-to-low extinction probabilities published by dedicated researchers become the “theoretical excuse” for nations to relax control over high-risk technologies, accelerate technological iteration, and compete for technological hegemony. Relying on the erroneous perception that probabilistic risks are controllable, nations continue to push forward competition in high-risk fields such as AI, genetic engineering, nanotechnology, and nuclear technology, lowering regulatory thresholds and weakening ethical constraints—directly accelerating the accumulation of global risks and worsening the crisis. The localized risk assessments of front-line experts, confined to Intra-industry use, do not produce such global governance harm.

4.3. Misleading Top-Level Public Policy, Missing Humanity’s Window for Self-Rescue

Humanity currently has only a limited window for self-rescue, lasting roughly through this century. The orientation of global technology governance policy directly determines the fate of civilizational survival. The governance thinking dominated by the probabilistic research paradigm traps global governance in the shallow misconception of “reducing probabilities, controlling localized risks, and patching end-of-pipe solutions.” Nations invest massive resources in end-of-pipe measures such as AI value alignment, nuclear safety protections, and biosurveillance, (Armstrong, 2013; Bostrom, 2012; Joy, 2000; Orseau & Armstrong, 2016; Russell, 2014, 2020; Russell et al., 2005; Yudkowsky, 2001, 2016, 2022) while deliberately avoiding the fundamental solutions of “strictly controlling high-risk technology R&D, breaking down national competition barriers, and constructing a unified global governance system.”

The probabilistic governance logic of dedicated researchers misleads the global academic community and national political circles into believing that fragmented international cooperation and localized industry regulation can reduce risk probabilities and avert extinction crises. This completely ignores the irreconcilability of structural contradictions, causing the continuous misalignment of global governance priorities and causing humanity to repeatedly miss the optimal opportunity for fundamental self-rescue.

4.4. Distorting Academic and Governance Ecosystems, Suppressing the Implementation of Scientific Fundamental Solutions

The probabilistic research paradigm, through its monopoly over academic discourse, has become the mainstream guiding framework for global crisis governance, completely distorting the academic and governance ecosystems in related fields. Vast amounts of research resources, talent, and funding are invested in meaningless probabilistic modeling, while scientific research focusing on crisis inevitability, path tracing, and global governance reconstruction is marginalized, questioned, and suppressed. The deterministic governance solution proposed in this paper, because it does not conform to the mainstream probabilistic academic paradigm, struggles to be effectively disseminated and implemented, hindering the construction of humanity's fundamental self-rescue system.

5. The Scientific Research Paradigm: Extinction path Analysis + Defense-Limit Testing, Locking in the Inevitability of Crisis

What this paper critiques is not the probabilistic tool itself, but the severe misapplication of its paradigm context—forcibly elevating probabilistic methods that are only suitable for localized, repeatable, technical risk optimization to the level of assessing humanity's civilizational survival, guiding global public policy, and constructing top-level governance systems. To thoroughly replace this subjective, ambiguous, and governance-harmful probabilistic research system, this paper relies on five constant core axioms to construct an ultimate risk assessment framework that involves no subjective value assignment, no probabilistic gaming, and yields rigid, empirically verifiable conclusions—namely, the “extinction path analysis + defense-limit testing” scientific paradigm (Hu, 2018). This paradigm has its own exclusive and rigorous core definitions: Extinction path Analysis uses the five constant axioms as its sole underlying basis, employing a standardized five-step deductive reasoning process to rigidly derive, from the underlying evolutionary patterns of civilization and technology, the inevitable path through which unchecked technological development leads to human extinction. Defense-Limit Testing conducts comprehensive empirical testing of all technological extinction means, with the core argument that all risk offsetting and defense solutions have ineliminable development windows. As multiple extinction mechanisms undergo continuous iteration, these high-risk exposure intervals compound indefinitely, ultimately leaving humanity in permanent exposure to extinction threats and rendering catastrophe inevitable. The entire paradigm is supported by objective laws, closed-loop reasoning, and extreme-case empirical testing, definitively ending the erroneous perception that “technological extinction is a random probabilistic event,” and providing scientific theoretical support for fundamental global technological governance reform.

5.1. Paradigm Underpinning: Five Constant Core Axioms

The five constant axioms are the theoretical foundation of the two research modules. They have been verified throughout the entire cycles of human civilization evolution, technological iteration, and social competition—with no exceptions, no reversals, and no failure scenarios. They are the core elements that distinguish this paradigm from subjective probabilistic models (Hu, 2018).

First, the Axiom of the One-Way Increase of Technological Destructive Power. Technological iteration is inherently an upgrade of tool-making and destructive capacity, exhibiting only a one-way strengthening trend, with no possibility of technological regression, weakening of destructive power, or spontaneous stagnation. This provides the eternal basis for the continuous emergence of new extinction means and the continuous escalation of risk levels.

Second, the Axiom of the Constancy of Human Nature. Humanity's core instincts of self-interested competition, vengeance, and hatred cannot be eradicated through morality, institutions, or law. This determines that global technological competition, high-risk R&D, and interest competition will persist permanently, with no possibility of voluntarily restricting high-risk technological development.

Third, the Axiom of the Irreversibility of Group Competition. Competition among nations, ethnic groups, interest blocs, and research entities is the normal state of civilization. With no global unified mandatory control authority, high-risk technologies will inevitably continue to be developed, proliferated, and iterated.

Fourth, the Axiom of the Perpetual Lag of Defense Behind Destruction. Destructive technologies are low-cost, easy to deploy, and quick to implement, while corresponding defense and offsetting technologies have long development cycles and are difficult to deploy. There are inherent and ineliminable risk windows.

Fifth, the Axiom of the Inherent Uncertainty of Technology. Technological exploration has no boundaries and no end. Positive breakthroughs inevitably bring unknown secondary risks. New extinction means will continue to emerge unexpectedly, impossible to predict or preempt. (Tegmark, 2017) These five axioms are mutually reinforcing and form a closed-loop, self-consistent system, providing absolutely stable objective foundations for law-based reasoning and extreme-case empirical testing.

5.2. Extinction path Analysis: Standardized Five-Step Deductive Reasoning Based on the Five Axioms

Extinction path Analysis is not a fragmented risk deconstruction of various high-risk technologies, but a universal standardized reasoning framework built upon five constant axioms, applicable to all technological extinction risks. It locks in the sole inevitable path from uncontrolled technological development to human extinction at the fundamental law level, completely overturning the core logic

of the probabilistic paradigm that “risks are controllable and outcomes are random.” The five-step reasoning is interlocked and layered, with no logical loopholes or room for optimism bias. The specific derivation is as follows:

Step One: Technological destructive power increases unidirectionally and without limit, with the risk base continuously expanding. Relying on the Axiom of the One-Way Increase of Technological Destructive Power, human technology always iterates along the unidirectional trajectory of “capability amplification and destructive power enhancement.” From the Stone Age, where one arrow or knife could kill only one person, to the invention of gunpowder, where one explosive shell could kill a crowd, to the discovery of the mass–energy equation and the invention of nuclear weapons, where a single nuclear bomb can destroy an entire city—all of this confirms this axiom. In the absence of global mandatory root-level control, the boundaries of technological exploration continue to widen, and the types and paths of technology that could trigger human extinction increase without cease, providing an inexhaustible source of risk for civilizational crisis and completely eliminating any possibility of natural risk dissipation.

Step Two: Group competition drives the deployment of high-risk technologies, with extinction means continuously taking shape. From early scientists’ fear that nuclear weapons could exterminate humanity (nuclear weapons are only means of mass destruction, not extinction means), to today’s possession not only of nuclear weapons but also of genetic technologies, nanotechnology, and AI technologies that are far closer to the possibility of human extinction—all of this confirms the Axiom of the Constancy of Human Nature and the Irreversibility of Group Competition. Under this axiom, global technological competition is permanent; seizing technological hegemony and obtaining development dividends are the core objectives. Under the competition mechanism, all frontier technologies with extinction potential will continue to be developed and deployed, and will not be voluntarily abandoned due to potential risks. Even if some nations tighten controls, competitive entities will continue to break through technological boundaries, continuously transforming theoretical risks into real threats (Armstrong et al., 2013).

Step Three: Following the Industrial Revolution, technological iteration has accelerated exponentially, rapidly advancing to the brink of humanity’s extinction capability, with technological destructive power completely surpassing the safety threshold of human civilization. Traditional technologies can only inflict localized, limited damage, whereas contemporary cutting-edge technologies, through a single breakthrough, can deliver universal, catastrophic, and irreversible civilizational strikes, marking humanity’s definitive exit from the low-risk survival phase. If this trajectory continues, even if none of the currently concerning technology types can extinguish humanity, as long as technological development persists, extinction means will inevitably emerge. Moreover, once the first extinction means appear, new ones will sequentially arise, and higher-tier

extinction means—initially accessible only to nations—are increasingly becoming attainable by corporations, and even by individual scientists. While national behavior is somewhat controllable, individual human behavior is extremely difficult to regulate.

Step Four: Defense systems are inherently lagging, with fatal high-risk vulnerability windows permanently existing. Based on the axiom that defense permanently lags behind destruction, the deployment and proliferation speed of new extinction technologies far exceeds that of corresponding defense technologies. Destructive technologies can be rapidly developed and deployed, whereas defense systems require a lengthy process of scientific research, experimental validation, and universal deployment, forming an unresolvable defense gap window: when extinction means become mature and usable, humanity lacks effective prevention and control capabilities, remaining in a high-risk state without any safety net throughout.

Step Five: Multiple-source vulnerability windows compound indefinitely, making extinction an inevitable outcome. Due to humanity's inherent nature of hatred and retaliation, and the axiom of inherent technological uncertainty, after extinction means emerge, not only will extremists deliberately trigger them, but laboratory accidents and inadvertent use of technological products may also ignite extinction events. Moreover, as long as technology continues to develop without restraint and humanity has not yet gone extinct, new extinction means will persistently arise. Before old risk defense systems are perfected, new high-risk technologies have already taken shape, causing the defense vulnerability windows corresponding to various technologies to continuously accumulate, overlap, and compound indefinitely. Under the triple effect of infinite time, infinite risks, and infinite high-risk vulnerability windows, regardless of the probability of any single risk event, a triggering event will inevitably occur, rendering human extinction a 100% certain outcome.

The five-step reasoning forms a complete closed-loop of fundamental laws, proving from the essence of civilizational and technological evolution that the ultimate endpoint of uncontrolled technological development must be human extinction, with no probability fluctuation, no exception avoidance, and no room for optimism bias, thereby fundamentally negating the probabilistic paradigm's notion of contingency.

5.3. Defense-Limit Testing: Empirical Verification of Inevitability Through the Window-Overlap Mechanism

If extinction path analysis is macro-level law-based deduction, defense-limit testing is micro-level operational empirical verification—a core module of this paradigm. Although some destructive means may be offset, some means are impossible to offset. For example, a super-plague developed through genetic technology might potentially be offset through pharmaceutical treatment, but the energy released by a nuclear explosion cannot be offset. General AI might potentially be offset

through alignment techniques (though the author does not believe alignment theory can succeed), but the unlimited replication of nanobots cannot be offset (Armstrong, 2013; Bostrom, 2012; Orseau & Armstrong, 2016; Russell, 2014, 2020; Russell et al., 2005; Yudkowsky, 2001, 2016, 2022; Yudkowsky & Soares, 2025). Moreover, any offsetting measure has a development time window. To maximize the rigor of the argument, this paper adopts the most favorable assumption for the probabilistic paradigm: assuming that humanity possesses sufficient resources, time, and technical capability to develop 100% of the corresponding defense and offset technologies for all extinction means in the future, and that all high-risk threats can theoretically be completely neutralized (in reality, this assumption is impossible—there will always be extinction technologies that cannot be defended against). Even under this optimal premise, extinction can still be empirically proven to be inevitable, thoroughly puncturing the probabilistic paradigm’s idealized governance cognition.

The core logic of defense-limit testing focuses on the ineliminability and infinite overlap of development windows. All extinction-level technologies—nuclear weapons, synthetic biology, nanotechnology, superintelligence (Bostrom, 2002, 2019; Rees, 2000, 2010; Tegmark, 2018; Yudkowsky & Soares, 2025)—require fixed development cycles for their defense, offset, and control solutions. The development process from “no defense” to “defensible” is an absolutely fatal high-risk window. During this window, extinction technologies are mature and available, disasters can be triggered at any time, and humanity has no fallback control capability. This gap phase cannot be skipped or offset.

The example of nuclear weapons provides intuitive confirmation: nuclear weapons can reach technological maturity and military proliferation within decades, while the comprehensive defense system of global nuclear early warning, interception, conflict management, and post-disaster recovery requires decades or even centuries of continuous construction and improvement. During the long window between nuclear technology maturity and the absence of defense systems, humanity faces the existential threat of nuclear miscalculation, nuclear conflict, and nuclear proliferation for an extended period. Even if humanity eventually perfects the nuclear defense system and offsets nuclear risks, the previously existing high-risk gap periods have already exposed human survival to sheer luck rather than absolute safety throughout that entire time.

The risk of a single technology’s vulnerability window is already fatal, and the continuous iteration of technology will cause high-risk vulnerability windows to compound indefinitely and persist permanently. Before nuclear risk defenses are fully secured, synthetic biology for custom pathogen technology has already proliferated, creating a biological disaster vulnerability window; before biological defense systems are perfected, nanotechnology and super AI extinction risks continue to emerge, adding new high-risk vulnerability windows (Bostrom, 2002, 2011, 2019; Rees, 2000, 2010; Tegmark, 2018; Yudkowsky & Soares, 2025). This does not even account for the

emergence of even more dangerous technologies in the future. Consequently, these vulnerability windows layer over one another and continuously compound, leaving human civilization permanently in a high-risk state of comprehensive defense deficit.

This leads to the core conclusion: humanity's research capabilities and defense technologies can continuously iterate, but can never fill the infinitely overlapping high-risk windows. Every new extinction technology inevitably comes with its own exclusive defense gap period. The infinite addition of new windows ultimately makes the outbreak of extinction forces inevitable. This is the ultimate defense limit of technological risk control: risks can be resolved after the fact, but cannot be preemptively prevented. The infinitely accumulating antecedent high-risk windows doom human extinction to be inescapable.

5.4. Core Comparison of the Two Paradigms: Deterministic Scientific System Overwhelms Probabilistic Subjective System

The empirically grounded scientific paradigm of this paper exhibits an essential hierarchical gap compared to the probabilistic research paradigm, fully demonstrating that the probabilistic paradigm is unqualified to guide global top-level technology governance.

First, the logical foundations are fundamentally different. The probabilistic paradigm relies on researchers' subjective value assignments, empirical conjecture, and artificial assumptions, with no fixed axioms or objective laws support, producing contradictory scholarly conclusions—essentially subjective academic speculation. This paradigm relies on five constant axioms, using standardized five-step reasoning and window-based extreme-case empirical testing. The entire process is objectively verifiable, logically closed-loop, and produces rigid conclusions with no room for subjective interpretation.

Second, the risk perception logic is completely opposed. The probabilistic paradigm defines civilizational extinction as a random probabilistic event, believing that technological optimization and international cooperation can reduce risk probabilities and avert disaster, ignoring the inevitable laws of technological evolution. This paradigm does not rely on the probability of any single risk event; it relies only on the objective mechanism of infinite window overlap to demonstrate that, in the long run, human extinction is the only inevitable outcome.

Third, the governance orientation is vastly different. The probabilistic paradigm leads to shallow end-of-pipe governance, indulging in probability reduction and localized patching, unable to resolve the structural contradiction between technology and control—treating the symptoms but not the disease. This paradigm precisely reveals the core crisis of window overlap, pointing directly to the only fundamental solution: through unified global governance and root-level control of high-risk technologies, blocking the birth of new extinction means at the source, terminating the continuous

addition of windows, and achieving the fundamental guarantee of civilizational survival.

Fourth, the application boundaries are clearly defined and insurmountable. The probabilistic paradigm is only suited to domain-specific, single-event, repeatable, small-scale technical failure assessment. When cross-applied to global civilizational assessment and top-level policy design, it inevitably leads to cognitive and governance failures. This paradigm is specifically designed for humanity's ultimate existential crisis, global and comprehensive technology governance, and fundamental institutional reform. It is the only scientific system currently suited to global governance and capable of solving the technological extinction problem.

6. Conclusion and Call to Action

This paper precisely defines the application boundaries of probabilistic assessment of technological risk, aiming to correct the previous one-size-fits-all cognitive bias: the probabilistic assessments of localized technological risk made by front-line technology practitioners and domain-specific technical experts are grounded in industry practice, suited to contextual needs, and possess full scientific validity and reference value—they are important foundations for technological risk control. The true fundamental fallacy that severely threatens the survival of human civilization lies in the academic community engaged in holistic research on humanity's technological crises and ultimate existential risks, which has erroneously elevated probabilistic methods—only suitable for specialized technical fields—to the core guiding paradigm for global technological crisis governance and top-level public policy-making.

Dedicated crisis researchers represented by Nick Bostrom and Toby Ord (Armstrong et al., 2013; Bostrom, 2013, 2025; Cotton-Barratt & Ord, 2015; Ord et al., 2010; Ord, 2021), relying on subjectively assigned probabilistic models, have produced widely divergent, objectively unfounded global extinction probability conclusions, using these to dominate global risk perception and governance policy. This suffers from three fatal defects: unscientific nature, meaninglessness, and high harmfulness. This paradigm disregards the uniqueness and inevitability of human civilization extinction, weakening a rigid survival crisis into a floating random risk, misleading nations into relaxing root-level control and indulging in end-of-pipe governance and localized optimization. It seriously obstructs the construction of a unified global crisis prevention system and continuously delays humanity's critical window for self-rescue.

Drawing upon nearly half a century of in-depth research, and utilizing the “extinction path analysis + defense-limit testing” empirical paradigm, combined with the rigorous deduction of the five constant axioms, the author establishes the following core conclusion: if humanity allows the unchecked iteration and boundless development of science and technology, it will inevitably face overall species extinction and civilizational collapse within this century to two to three hundred years.

This outcome has 100% absolute certainty, with no probabilistic buffer or room for wishful thinking (Hu, 2018). Humanity's current window for self-rescue is already extremely urgent. Probabilistic ambiguous governance and complacency-based risk prevention are currently the greatest cognitive obstacles and policy barriers to global civilizational self-rescue.

Based on the core imperative of the holistic survival of human civilization, this paper hereby issues the following precise calls:

First, the global academic community must strictly distinguish the boundaries of research paradigms. It should affirm the industry value of front-line technical probability assessments, and thoroughly abandon the erroneous paradigm of "probability guiding global governance" in dedicated crisis research, focusing instead on deterministic crisis research and fundamental solution exploration.

Second, nations and international organizations must correct the misconception of probabilistic governance. They should abandon the superficial approach of "probability reduction and localized control," and, grounded in the inevitable laws of crisis, promote an awakening movement to re-understand science and technology. This should further advance the process of the Great Unification of humanity, construct a unified global high-risk technology governance system, and curb unchecked technological development at its source.

Third, all of humanity must establish an absolutely deterministic crisis awareness, abandon probabilistic complacency, and support comprehensive root-level technology control and global governance system reconstruction, jointly safeguarding the foundation of human civilization's survival.

The ultimate survival of human civilization cannot rely on subjective probabilistic gaming or localized technological optimization. It can only be achieved through a precise cognition of objective inevitable laws and fundamental, root-level governance transformation. Strictly distinguishing the paradigm boundaries between technical risk control and global governance, and abandoning probabilistic top-level governance misconceptions, are necessary prerequisites for humanity to emerge from crisis and achieve civilizational perpetuity.

Acknowledgments

Not applicable.

Author contributions

The sole author was responsible for conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, and project administration, and has read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data availability

The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of interest

The author declares that there is no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable. (This manuscript contains no identifiable personal data in any form.)

Additional information

Received: 25 August 2026

Accepted: 8 September 2026

Published online: 22 September 2026

Publisher's Note

Wisdom Academic Press Ltd. remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- [1] Hu Jiaqi. Saving Humanity (First Edition, Volume I: The Extinction of Humanity) [M]. Beijing: Tongxin Publishing House, 2007. ISBN: 978-7-80716-508-8[Chinese]
- [2] Hu Jiaqi. Saving Humanity (First Edition, Volume II: The Path for Humanity) [M]. Beijing: Tongxin Publishing House, 2007. ISBN: 978-7-80716-508-8[Chinese]
- [3] Hu Jiaqi. The Greatest Problem [M]. Shenyang: Northeastern University Press, 2007. ISBN: 978-7-81102-478-4[Chinese]
- [4] Hu Jiaqi. On Human Extinction [M]. Hong Kong: Ha Fai Yi Publication limited, 2008. ISBN: 978-988-99113-4-8
- [5] Hu Jiaqi. Saving Humanity (Third Edition) [M]. Singapore: South Ocean Publishing House, 2019. ISBN: 978-981-17-0336-2[Chinese]
- [6] Martin Rees. Our Final Century [M]. Translated by Chen Yanyan, Zhao Min. Changsha: Hunan Science & Technology Press, 2010. ISBN: 978-7-5357-6058-6
- [7] Martin Rees. The Future of Humanity [M]. Translated by Yao Song, Ding Dingchong. Shanghai: Shanghai Jiao Tong University Press, 2020. ISBN: 978-7-313-22731-7
- [8] Max Tegmark. Life 3.0: Being Human in the Age of Artificial Intelligence [M]. Translated by Wang Jieshu. Hangzhou: Zhejiang Education Publishing House, 2018. ISBN: 978-7-5536-7278-6
- [9] Max Tegmark. Our Mathematical Universe: My Quest for the Ultimate Nature of Reality [M]. Translated by Wang Jieshu. Hangzhou: Zhejiang People's Publishing House, 2017. ISBN: 978-7-213-07980-1
- [10] Nick Bostrom. Deep Utopia: Life and Meaning in a Solved World [M]. Translated by Huang Feifei. Beijing: CITIC Press, 2025. ISBN: 978-7-5217-7164-0
- [11] Stuart Russell. Human Compatible: Artificial Intelligence and the Problem of Control [M]. Beijing: CITIC Publishing Group Co., Ltd., 2020. ISBN: 978-7-5217-1951-2
- [12] Stuart Russell, Peter Norvig. Artificial Intelligence: A Modern Approach [M]. Beijing: People's Posts and Telecommunications Press, 2023. ISBN: 978-7-115-59811-0.



- [13] Toby Ord. *The Precipice: Existential Risk and the Future of Humanity* [M]. Translated by Wei Silin. Beijing: CITIC Press, 2021. ISBN: 978-7-5217-3226-9.
- [14] Hu Jiaqi. *Saving Humanity* [M]. Canada: FriesenPress, 2018. ISBN: 978-1-5255-3173-6
- [15] Joy, B. Research Priorities for Robust and Beneficial AI [J]. *Wired*, 2000(4).
- [16] Joy, B. Research Priorities for Robust and Beneficial AI [J]. *Wired*, 2000(4). (Note: Duplicate entry as in original)
- [17] BBC Radio 4. (2024.12.27). *Today* [Video]. YouTube. <https://www.youtube.com/watch?v=gnr0otD4HD0>
- [18] Milmo, D. AI Firms Warned to Calculate Threat of Super Intelligence or Risk It Escaping Human Control [N]. *The Guardian*. 2025.5.10.
- [19] Yudkowsky, E & Soares, S. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* [M]. US: Little, Brown and Company, 2025.9. ISBN: 9780316595643.
- [20] Yudkowsky, E. AGI Ruin: A List of Lethalities. [EB/OL]. <https://intelligence.org/2022/06/10/agi-ruin/>, 2022(6).
- [21] Yudkowsky, E. *The AI Alignment Problem: Why It's Hard, and Where to Start* [R]. Machine Intelligence Research Institute, 2016(5).
- [22] Yudkowsky, E. Cognitive Biases Potentially Affecting Judgement of Global Risks [A]. *Global Catastrophic Risks* [M], 2008(7): 91-119.
- [23] Yudkowsky, E. *AGI Ruin: Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. [R]. Machine Intelligence Research Institute, 2001(6).
- [24] Orseau, L & Armstrong, S. Safely Interruptible Agents [J]. *UAI'16: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016(6).
- [25] Metaculus. If a Global Catastrophe Occurs, Will It Reduce the Human Population By 95% Or More? (2018.9.19). <https://www.metaculus.com/questions/9060/#comments>
- [26] Tegmark, M & Bostrom, N. How Unlikely is a Doomsday Catastrophe? [J]. *Nature*, 2005(12).
- [27] Rees, M. *Our Final Century* [M]. *Wired*, 2000(4).
- [28] Bostrom, N. Existential Risk Prevention as Global Priority [J]. *Global Policy*, 2013(4): 15-31.

- [29] Bostrom, N. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards [J]. Journal of Evolution and Technology, Vol. 9, No. 1, 2002.
- [30] Bostrom, N. Astronomical Waste: The Opportunity Cost of Delayed Technological Development [J]. Utilitas, Vol. 15, No. 3, 2003: 308-314.
- [31] Bostrom, N. Information Hazards: A Typology of Potential Harms from Knowledge [J]. Review of Contemporary Philosophy, Vol. 10, 2011: 44-79.
- [32] Bostrom, N. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents [J]. Minds and Machines, Vol. 22, Iss. 2, 2012(5).
- [33] Bostrom, N. Existential Risk Prevention as Global Priority [J]. Global Policy, 2013(4): 15-31.
(Note: Duplicate of [28])
- [34] Bostrom, N. The Vulnerable World Hypothesis [J]. Utilitas, Vol. 15, No. 3, 2019.11: 455-476.
- [35] PowerfulJRE. (2025.2.28). Joe Rogan Experience #2281 - Elon Musk [Video]. YouTube.
<https://www.youtube.com/live/sSOxPJD-VNo>
- [36] Cotton-Barratt, O & Ord, T. Existential Risk and Existential Hope: Definitions [J]. Future of Humanity Institute – Technical Report, 2015(1).
- [37] Ord, T & Hillerbrand, R & Sandberg, A. Probing the Improbable: Methodological Challenges For Risks With Low Probabilities And High Stakes [J]. Journal of Risk Research, Volume 13, 2010(4): 191-205.
- [38] Armstrong, S. Risks and Mitigation Strategies for Oracle AI [J]. Studies in Applied Philosophy, Epistemology and Rational Ethics, 2013(5): 335-347.
- [39] Armstrong, S. & Bostrom, N. & Shulman, C. Racing to the Precipice: a Model of Artificial Intelligence Development [J]. Technical Report, 2013(1): 2-7.
- [40] Russell, S. Unifying Logic and Probability: Recent Developments [J]. Communications of the ACM, Vol 58, 2015(7).
- [41] Russell, S. Of Myths and Moonshine [J]. Edge.org, 2014(11).
- [42] Russell, S, Dewey, D, Tegmark, M. Research Priorities for Robust and Beneficial AI [J]. Nature, 2005(12).