2025 年第 1 期（总第 38 期）

# 国际战略与安全研究报告

# INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

## 人工智能风险相关术语研究（一）

## Glossary Research on Artificial Intelligence Risks (Part I)

清华大学战略与安全研究中心

CENTER FOR
INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

# 人工智能风险相关术语研究（一）

# Glossary Research on Artificial Intelligence Risks (Part I)

CISS人工智能与国际安全项目术语工作组

*By CISS Working Group on Artificial Intelligence Glossary*

**编者按**：2025年2月13-14日，清华大学战略与安全研究中心与美国布鲁金斯学会在慕尼黑举行第12轮“中美人工智能与国际安全二轨对话”。在术语讨论环节，双方共同提出20个人工智能风险相关的概念术语进行探讨交流，并就其中9个术语给出了各自的具体解释，以此增进彼此对人工智能潜在安全风险的认知和理解。

*On February 13-14, 2025, the Center for International Security and Strategy (CISS) of Tsinghua University and the Brookings Institution held the 12th round of the“China-U.S. Track 2 Dialogue on Artificial Intelligence and International Security” in Munich. During the Glossary Session, both sides proposed 20 AI risk-related terms for exchange, and then provided their respective definitions for nine of these terms to enhance mutual understanding of the potential security risks posed by artificial intelligence.*

## ●灾难性风险

灾难[1]性风险[2]特指破坏力极强、远超常规应对能力的极端事件，其核心特征可归纳为三点：破坏烈度极高、影响范围广泛、持续时间漫长。这类风险既可能源自自然力量（如超级火山爆发），也可能源于人为因素（如核战争），其破坏性不仅会摧毁社会基础架构，更将

[1] 灾祸造成的苦难;灾祸。《汉语大词典》(2007版)，上海：上海辞书出版社；天灾人祸造成的严重损害和苦难。《辞海》(第6版)，上海：上海辞书出版社，第5582页。

[2] 可能发生的危险。《汉语大词典》(2007版)，上海：上海辞书出版社。

引发跨地域、跨代际的深远影响。

在人工智能领域，该类风险可能体现为两种形态：其一，系统因复杂架构产生不可预测的突发行为；其二，技术遭恶意滥用导致大规模危害。其形成机制通常源于两大矛盾，即技术迭代速度超越人类监管能力，或现有治理体系无法有效约束技术发展使其符合社会整体利益。

**●Catastrophic risk**

Catastrophic risk refers to events with exceptional destructive potential that exceeds ordinary response capacities, characterized by significant magnitude, geographic scope, and temporal persistence. Such risks, whether natural or anthropogenic in origin, disrupt fundamental societal structures and produce multi-generational consequences that transcend jurisdictional boundaries.

In the artificial intelligence domain, catastrophic risks emerge when systems potentially inflict widespread harm through either complexity-induced emergent behaviors or deliberate misuse. These scenarios typically develop through two primary mechanisms: technical systems evolving beyond effective human oversight, or governance frameworks proving insufficient to ensure responsible development aligned with broader societal interests.

●生存风险

作为风险谱系中最严峻的类别，生存[③]风险直接威胁人类物种存续，可能彻底阻断文明发展轨迹。相较于灾难性风险，这类风险（如小行星撞击、超级病原体或失控AI）具有不可逆特性，能够彻底消弭文明复苏的可能，根本上颠覆人类赖以生存的基础条件。

在人工智能领域，该类风险可能出现的情景包括：系统具备超越

[③] 活着，活下去；存在；生活。《汉语大词典》（2007版），上海：上海辞书出版社；活着；保存生命。《辞海》（第6版），上海：上海辞书出版社，第3960页。

人类认知边界的智能水平、文明演进方向出现根本性偏移、人类在关键领域丧失控制权、自主系统脱离监管破坏战略稳定，或有效治理框架之外运行的自主能力所导致的战略不稳定。

### ●Existential risk

Existential risk refers to threats capable of causing human extinction or permanently curtailing humanity's developmental trajectory, representing the terminal category within risk taxonomies. Such threats, whether cosmic, ecological, biological, or technological, are distinguished from catastrophic risks by their irreversible elimination of recovery pathways, potential to extinguish humanity entirely, and capacity to fundamentally alter parameters necessary for continued human existence.

In artificial intelligence contexts, existential risks arise where systems potentially undermine humanity's continued existence or drastically constrain its future potential. These scenarios typically emerge through mechanisms including intellectual capabilities surpassing human comprehension, fundamental transformation of civilizational trajectory, permanent displacement of human agency in critical domains, or strategic instabilities resulting from autonomous capabilities operating beyond effective governance frameworks.

### ●人工智能可控性

人工智能可控性是指人工智能系统在设定的条件（场景）和规定时间内，通过一系列策略、机制和技术保障，持续、稳定、可靠且安全地完成预定义的任务/功能的能力。特别是在发生故障时，它应能够：（1）即时隔离，切断故障模块，防止故障级联扩散；（2）维持安全状态，可即时切换到备用模式并保留基本功能；（3）具有可恢复性，在快速恢复后可重新投入任务。在当前条件下，鉴于前沿人工智能系统的“黑箱”技术特征，以及人类价值观的多元，实现对前沿人

工智能系统实施完全控制面临较大挑战。因此，需要更全面的风险缓解策略和敏捷治理方法。[4]

**●AI Controllability**

AI controllability refers to the ability of an artificial intelligence system to continuously, stably, reliably, and safely complete predefined tasks/functions under specified conditions and within a specified time frame through a series of strategies, mechanisms, and technical safeguards. Especially in the event of a failure, it should be able to (1) isolate instantly, cutting off the faulty module to prevent cascading failures; (2) maintain a safe state, switching to a backup mode and preserving basic functions; (3) support rapid recovery and re-entry into tasks after quick repairs. Considering the technical mechanisms of cutting-edge AI systems (LLM), as well as the inherent contradictions in human values, fully controlling advanced AI systems (including autonomous systems) will be very difficult, if not impossible. Therefore, more comprehensive risk mitigation strategies and agile governance approach are needed.

AI controllability refers to the ability of an artificial intelligence system to continuously, stably, reliably, and safely complete predefined tasks/functions within a specified timeframe and under given conditions (scenarios), through a series of strategies, mechanisms, and technical safeguards. Particularly in the event of a failure, it should be capable of: (1) immediate isolation, cutting off the faulty module to prevent cascading failures; (2) maintaining a safe state, being able to instantly switch to a backup mode

[4] 参考文献：

GJB 900B−2024，国军标，《装备安全性通用工作要求》。

GJB 900A−2001，国军标，《装备可靠性工作要求》。

ISO/IEC TS 8200:2024(en)Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems 国际标准化组织（ISO）/ IEC TS 8200：2024 信息技术 — 人工智能 — 自动人工智能系统的可控。

and preserve basic functions; (3) being recoverable, so that it can re-engage in tasks after rapid recovery. Under current circumstances, given the “black-box” technical characteristics of frontier artificial intelligence systems and the diversity of human values, achieving full control over such systems presents significant challenges. Therefore, more comprehensive risk mitigation strategies and agile governance approaches are needed.

### ●人工智能可治理性

人工智能可治理性是指通过制度、法律、技术、市场等机制，提高AI系统的可解释性、透明性、公平性、可靠性、安全性，以及AI事故的可追溯性和可问责性，切实维护国家主权、安全和发展利益，保障公民、法人和其他组织的合法权益，确保人工智能技术造福于人类。[5]

### ●AI Governability

AI governability refers to the enhancement of AI systems’ interpretability, transparency, fairness, reliability, and security, as well as the traceability and accountability of AI-related incidents, through mechanisms such as institutions, laws, technology, and markets. It aims to effectively safeguard national sovereignty, security, and development interests, protect the legitimate rights and interests of citizens, legal persons, and other organizations, and ensure that artificial intelligence technologies are used for the benefit of humanity.

### ●故意升级

故意升级指某一行为体为实现战略目标或影响对手行为而有计

[5] 参考文献：

国家新一代人工智能治理专业委员会，《新一代人工智能伦理规范》，2021年。

全国网络安全标准化技术委员会，《人工智能安全治理框架》1.0版，*AI Safety Governance Framework*, National Technical Committee on Cybersecurity of Standardization Administration olChina，2024年9月发布。

上海社科院、武汉大学等，《全球人工智能治理研究报告》，2024年11月发布。

划、有目的地提高冲突的强度或范围。这种升级行为可能包括公开的军事行动、激进的言辞或有控制地使用武力等，以表明决心或争取让步。战略信号传递和风险管理是这类升级内涵的重要组成部分。

●**Intentional escalation:**

A calculated and purposeful increase in the intensity or scope of conflict undertaken by an actor to achieve strategic objectives or influence an adversary's behavior. This may involve overt military actions, aggressive rhetoric, or controlled use of force aimed at signaling resolve or extracting concessions. It encompasses both strategic signaling and risk management.

●非故意升级

非故意升级指在危机当中由行动之非预期后果所引发的冲突强度上升。这种升级并非出于刻意的意图，而是由于沟通失误、误解或程序性错误等因素无意中加剧了紧张局势。这类升级凸显了有效沟通和程序管理的重要性。[⑥]

[⑥] 中国官方、军方与学术界的公开出版物并没有对有意升级、无意升级这两个概念进行专门、明确的定义，但相关讨论涉及对它们的理解，本词条参考了这些讨论，可参见如：胡平（中国国际战略学会）：《国际冲突分析与危机管理研究》，军事谊文出版社，1993年；中国现代国际关系研究院：《国际危机管理概论》，时事出版社，2003年；丁邦泉（国防大学）主编：《国际危机管理》，国防大学出版社，2004年；徐辉（国防大学）：《国际危机管理理论与案例解析》，国防大学出版社，2011年。

Publicly available publications from the Chinese government, military, and academic community do not provide a specific or explicit definitions of intentional escalation and unintentional escalation.  However, relevant discussions offer insights into their interpretations. For more details, see: Hu Ping (China Institute for International Strategic Studies, CIISS), *Analysis of International Conflict and Research on Crisis Management*, Junshi Yiwen Press, 1993; China Institutes of Contemporary International Relations (CICIR), *Introduction to International Crisis Management*, Shishi Press, 2003; Ding Bangquan (PLA National Defense University) etc. eds., *Management of International Crisis*, National Defense University Press, 2004; Xu Hui (PLA National Defense University), *Theory of Case Analysis of International Crisis Management*, National Defense University Press, 2011.

**●Unintentional escalation:**

An increase in conflict intensity that occurs as an unintended consequence of actions taken during a crisis. This form of escalation arises not from deliberate intent but from miscommunications, misinterpretations, or procedural missteps that inadvertently heighten tensions. It underscores the importance of effective communication and procedural management.

**●失控危机**

失控危机指人工智能系统因为技术原因作出与预设目标不相符的举动，且人类操作员无法对其进行有效控制，或无法在其酿成灾难性后果前终止系统[7]，原因可能在于人工智能系统链式反应的技术性故障[8]，也可能在于人工智能系统出现“自主意识”[9]。

**●Out of control crisis**

Out of control crisis: An artificial intelligence system takes actions that deviate from its predetermined objectives due to technical reasons, and human operators are unable to effectively control it or terminate the system before it causes catastrophic consequences. The underlying cause may be a technical malfunction leading to a chain reaction within the AI system or the emergence of “autonomous consciousness” within the AI itself.

**●战略稳定**

狭义的战略稳定特指核武器领域，包括相互确保摧毁、危机稳定和军备竞赛稳定等内涵，广义的战略稳定是指在全球范围内，各行为

[7] 张煌，杜雁芸. 俄美军用人工智能竞争的战略稳定风险及其治理路径[J]. 俄罗斯研究,2022, (06):157−190.

[8] 龙坤，徐能武. 人工智能军事应用的国际安全风险与治理路径[J]. 国际展望,2022,14(05):123−141+165−166.DOI: 10.13851/j.cnki.gjzw.202205007.

[9] 全国网络安全标准化技术委员会，《人工智能安全治理框架》, *AI Safety Governance Framework*, National Technical Committee on Cybersecurity of Standardization Administration of China，2024年9月发布。

主体通过保持自我克制并进行相互制约，从而在国际体系层面形成的一种相对稳定、平衡的战略态势。[10]

●**Strategic stability**

Strategic stability, in its narrow sense, specifically pertains to the realm of nuclear weapons, including mutual assured destruction, crisis stability, and arms race stability. (Tsinghua) In its broader sense, strategic stability refers to a relatively stable and balanced strategic situation at the international level, where actors maintain self-restraint and engage in mutual constraints, thereby contributing to global stability. (NDU)

●两用技术

两用技术：既有民事用途，又有军事用途[11]或有助于提升军事潜力，特别是可以用于设计、开发、生产或者使用大规模杀伤性武器及其运载工具的技术，包括相关的技术资料等数据[12]。

●**Dual-use technology**

Dual-use technology: Technologies that have both civil and military uses or help enhance military potential, especially those that can be used for the design, development, production, or use of weapons of mass destruction and their delivery vehicles, including relevant technical information and other data.

[10] 参考文献：

Bin Li and Gaochen Hu, China’s Nuclear Deterrence from the American Perspective, Vol.35, 2018, pp.21 - 41.

Yi Yang, Global Strategic Stability Theory (Quanqiu Zhanlve Wending Lun), China People’s Liberation Army National Defence University Press, 2005, p.3.

[11] 胡冬梅,王建卿,王海涛,等.军民两用技术研究现状及发展思路[J].科技导报,2018,36(10):14−19.

[12]《中华人民共和国两用物项出口管制条例》第一章第二条，Regulations on Export Control of Dual−use Items，Chapter I Article 2，https://www.gov.cn/zhengce/zhengceku/202410/content_6981400.htm

## 附：CISS人工智能与国际安全项目术语工作组成员名单

肖　茜　清华大学战略与安全研究中心副主任、
　　　　人工智能国际治理研究院副院长

陈　琪　清华大学战略与安全研究中心副主任、国际关系学系教授

朱启超　国防科技大学国防科技战略研究智库主任、教授

谢海斌　国防科技大学智能科学学院副主任、教授

董　汀　清华大学战略与安全研究中心副研究员

孙成昊　清华大学战略与安全研究中心副研究员

李　强　中国政法大学军事法研究所所长、副教授

鲁传颖　清华大学战略与安全研究中心特约专家，
　　　　同济大学政治与国际关系学院教授

祁昊天　北京大学国际安全与和平研究中心副主任、副教授

徐纬地　国防大学原战略研究所研究员

张　伶　原国防大学国家安全学院副教授

郑乐锋　清华大学战略与安全研究中心人工智能项目专员

刘　源　清华大学战略与安全研究中心研究助理

发表日期：2025年3月18日

**CISS Working Group on Artificial Intelligence Glossary Research**

**XIAO Qian**, Deputy Director, Center for International Security and Strategy, Tsinghua University

**CHEN Qi**, Deputy Director, Center for International Security and Strategy, Tsinghua University

**ZHU Qichao**, Director and Professor, Institute for Defense Technology and Strategic Studies, National University of Defense Technology

**XIE Haibin**, Deputy Director and Professor of Department of Intelligent Science and Technology, College of Intelligence Science and Technology, National University of Defense Technology

**DONG Ting**, Fellow, Center for International Security and Strategy, Tsinghua University

**SUN Chenghao**, Fellow, Center for International Security and Strategy, Tsinghua University

**LI Qiang**, Director and Associate Professor of the Military Law Institute, Law School at China University of Political Science and Law

**LU Chuanying**, Nonresident Fellow, Center for International Security and Strategy, Tsinghua University; Professor, School of Political Science and International Relations, Tongji University□

**QI Haotian**, Associate Professor and Deputy Director, Center for International Security and Peace Studies, School of International Studies, Peking University

**XU Weidi**, Senior Colonel (Ret.), former Research Fellow, Institute for Strategic Studies, National Defense University

**ZHANG Ling**, Senior Colonel (Ret.), Former Associate Professor,

National Security College, National Defense University

**ZHENG Lefeng**, Project Manager of AI and International Security, Center for International Security and Strategy, Tsinghua University

**LIU Yuan**, Research Assistant, Center for International Security and Strategy, Tsinghua University

审编：肖茜　　签发：达巍

**清华大学战略与安全研究中心编印**

办公地点：北京市海淀区清华大学明理楼428房间

联系电话：010-62771388

http://ciss.tsinghua.edu.cn　邮箱：ciss@tsinghua.edu.cn

扫码关注我们