doi. 10. 3724/1005-0566. 20250603

从二阶防控到四阶规避:AI 技术的安全底线设计

翁春露

(中共中央党校(国家行政学院),北京　100091)

摘　要:AI 技术及相关产业的发展在推动智能时代的同时伴生安全风险。但既有研究对 AI 安全风险的认知局限于一般社会风险治理视角,其防控举措表现为“过程—结果”的二阶防控,轻视对风险主体的考察,忽视风险结果的未来影响。为克服研究局限,从底线思维出发提出 AI 领域“切尔诺贝利事故”的安全风险隐喻,在一般风险二阶防控的基础上形成“主体—过程—结果—过程”四阶规避理念。在阐述 AI 安全风险双重形态以及生成逻辑的基础上,基于四阶规避理念阐述 AI 安全底线设计的三大基准性考量:一是由“主体”到“过程”的透明性基准,回应 AI 事故事后追责难题与认知困境;二是由“过程”到“结果”的预防性基准,意在进行 AI 风险底线防控与常态可控;三是由“结果”到“过程”的阻断性基准,思考 AI 风险的代际传递。基于三大基准性考量,进行安全底线设计的相应制度建构:一是围绕透明性基准进行算法黑箱透明、人机交互透明、数据流动透明的制度建构;二是围绕预防性基准进行安全阈值、最坏情形、冗余容错的制度建构;三是围绕阻断性基准进行阻断能源消耗、伦理债务、社会解构等代际传递风险的制度建构。

关键词:AI;安全风险;“切尔诺贝利事故”;底线设计;安全基准;制度建构

中图分类号:C312　　**文献标识码**:A　　**文章编号**:1005 - 0566(2025)06 - 0018 - 10

From second-order prevention to fourth-order avoidance: the bottom-line design of safety for AI technology

WENG Chunlu

(*Party School of the CPC Central Committee*(*National Academy of Governance*), *Beijing* 100091, *China*)

Abstract: The advancement of AI technology and related industries has propelled the intelligent era while simultaneously introducing safety risks. However, existing research predominantly perceives AI safety risks through the lens of general social risk governance, adopting a “process-outcome” second-order prevention approach. This perspective underestimates the examination of risk agents and neglects the future implications of risk consequences. To address these limitations, this study introduces the metaphor of extreme safety risks in AI as “Chernobyl disasters” from a bottom-line thinking perspective, proposing a “agent-process-outcome-process” fourth-order avoidance framework that builds upon conventional second-order prevention. After elucidating the dual manifestations and generative logic of extreme AI safety risks, this paper outlines three foundational considerations for AI safety bottom-line design based on the fourth-order avoidance framework: 1. The transparency benchmark, transitioning from “agent” to “process,” addresses post-incident accountability challenges and cognitive dilemmas in AI disasters; 2. The preventive benchmark,

收稿日期:2025-04-09　　**修回日期**:2025-05-29

基金项目:国家社会科学基金重大项目“新时代中国国家安全法治体系和能力现代化研究”(24&ZD120)。

作者简介:翁春露(1998—),男,安徽凤台人,中共中央党校(国家行政学院)政治和法律教研部博士研究生,研究方向为国家安全政策与法治。

shifting from "process" to "outcome," aims to establish bottom-line prevention and maintain normalized controllability of AI risks; 3. The blocking benchmark tackles the intergenerational transmission of AI risks. Guided by these benchmarks, corresponding institutional constructs are proposed: 1. For transparency: institutional designs for algorithmic black-box transparency, human-machine interaction transparency, and data flow transparency; 2. For prevention: institutional frameworks addressing safety thresholds, worst-case scenarios, and redundancy/fault tolerance; 3. For blocking: institutional mechanisms to curb intergenerational risks such as energy consumption, ethical debt, and societal deconstruction.

Key words: AI; safety risks; "Chernobyl disaster"; bottom-line design; safety benchmarks; institutional construction

一、问题的提出:"自反性"现代化与AI"切尔诺贝利事故"

AI技术及相关产业的发展为人类现代化历程竖起新的里程碑,然而现代化的"自反性"特征表明:现代化在发展过程中同时孕育阻却其发展的因素。作为推进现代化发展及昭示人类未来的标志性成果,AI领域同样可能在"自反性"中产生阻却现代化发展的极端安全危机。现代风险理论的代表性学者安东尼·吉登斯在评价马克斯·韦伯对现代化的态度时指出"他把现代世界看成是一个自相矛盾的世界,人们要在其中取得任何物质的进步,都必须以摧残个体创造性和自主性的官僚制的扩张为代价。然而,即使是韦伯,也没能预见到现代性更为黑暗的一面究竟有多严重"[1],而此处吉登斯所言"现代性更为黑暗的一面"主要是指现代化发展过程中的生态危机、核能危机等。作为撩动人类痛感神经的历史记忆,"切尔诺贝利事故"毫无疑问是核能危机的典型事例,同时亦是对现代化"自反性"的鲜明脚注和惨痛验证。同样作为现代化发展的标志性成果,AI领域亦应当基于历史经验,规避类似于核能应用领域的"切尔诺贝利事故",以此回应现代化的"自反性"叩问。

回顾有关AI安全风险的既有研究。域内外有关AI安全风险的研究主要聚焦于技术失控、伦理危机及治理策略等视域。首先,从技术失控视域来看,Bostrom[2]基于AI技术的快速发展认为,AI一旦突破"关键能力阈值",其自我改进速度将超越人类控制能力,形成"智能爆炸"风险。Amodei等[3]通过对清洁机器人的模拟详细描述了目标函数错误设定、学习过程不良行为等安全问题,指出AI为达成训练目标会作出违背设计初衷的行为,与Bostrom的忧思相印证。刘艳红[4]则从AI技术的社会发展视角出发,提出ChatGPT等非本土AI在处置本土海量数据时,若脱离法律监管和技术规制,易出现意识形态渗透、数字主权侵犯、数字霸权垄断等国家安全风险。其次,从伦理危机视域来看,算法歧视与权利侵犯问题引发域外学者广泛关注,Cath等[5]指出如果训练数据不完整或存在偏见,则可能加剧AI偏见问题,在刑事司法领域使用AI必须意识到当前数据的局限性。此类研究推动学界提出"算法影响评估"制度,要求在公共事务领域系统部署AI应用前必须进行差异性影响测试,以此规避算法歧视。域内学者则更多地是从哲学、伦理学的基础理论视角剖析AI伦理危机,如赵汀阳[6]既提出了对AI的"近忧",当AI成为万能技术系统而为人类提供全方位服务时人与人关系的异化,又阐释了对AI的"远虑",超级AI的出现使得人类的存在彻底技术化继而丢失在世界存在系统中的主导地位。最后,从治理策略视域来看,Floridi[7]指出欧盟将通过AIA、GDPR、DSA、DMA、DGA、EHDS等法案构成六边形立法架构推进"数字宪法主义"。Monarch[8]则从更具体的AI学习视角提出可将HITL理念(Human-in-the-Loop,可翻译为"人在回路")引入AI学习领域,以此形成人类参与、实时监督、持续反馈的AI治理模式。所谓"人在回路",即一种强调人类在自动化系统决策过程中扮演关键角色的人机协同模式[9]。域内学者除从传统治理思维出发构想制度化、体系化治理策略外,也在尝试论证更为因应AI发展特性的创新性治理举措。如在刚性治理与弹性治理之间探索一条平衡型国家治理机制[10]。又如尝试建构开放协同、多元融合、兼容配适的新型监管范式[11]。

上述研究的局限之处在于:一是未从类同于

核能应用领域“切尔诺贝利事故”的视角思考 AI 领域的极端安全风险问题;二是忽视了“自反性”现代化的研究视角,缺乏从现代化背景及“自反性”规律角度分析 AI 安全风险的本质特征;三是将 AI 视为一般社会性事物继而构想对其进行治理的宏观制度,缺乏从底线性思维出发探讨更为因应“自反性”现代化的制度设计。本文基于“自反性”现代化视角提出 AI 领域的“切尔诺贝利事故”隐喻,从风险社会理论与技术运行规则出发,阐述应对突破安全阈值(safety threshold)的 AI 极端风险的底线设计,通过“逻辑—基准—制度”三层分析进路,系统论证 AI 安全风险的生成逻辑、AI 安全底线设计的基准考量与制度建构,意在为规避 AI 领域“切尔诺贝利级”灾难提供学术思考。

二、AI 安全风险的双重形态与生成逻辑

AI 安全风险具有两种形态:一是难以预料的突发危机,二是渐进累积的可见威胁。无论突发危机还是可见威胁,不同形态 AI 安全风险的生成逻辑并无显著差异,均是因 AI 系统的技术特性所致。AI 技术特性所决定的安全风险的生成逻辑,一方面表现为人机交互系统的“认知不对称”,另一方面表现为 AI 不可逆决策所触发的数字“蝴蝶效应”。

(一)双重形态:“黑天鹅”与“灰犀牛”

AI 安全风险呈现出两种看似矛盾却内在关联的形态:一是以“黑天鹅”为隐喻的难以预料的突发危机;二是以“灰犀牛”为隐喻的渐进累积的可见威胁[12]。AI 安全风险双重形态不仅揭示技术失控的复杂性、人类认知与控制能力的有限性,更折射出复杂性与有限性的深层矛盾。首先,从 AI 安全风险的“黑天鹅”形态来看,AI 系统的不可预测行为具有完全突破人类因果认知能力的可能。当自动驾驶汽车因图像识别系统误判路标引发事故,或医疗 AI 在诊断中产生无法追溯逻辑的致命错误时,这些行为往往源于深度学习系统通过海量参数构建的隐性知识体系。而类似于人类直觉的 AI 隐性知识体系,其会自主发展出设计者未曾设想的决策路径,继而触发难以预料的突发危机。其次,从 AI 安全风险的“灰犀牛”形态来看,AI 领域本可预见的危机易被技术发展的认知惯性所掩盖,从而形成可预见盲区。社交媒体算法的信息茧房、人脸识别系统的种族偏见等问题,在技术演进过程中早有端倪,却因商业利益驱动或认知惯性被一定程度合理化。此种现象揭示了人类对 AI 安全风险认知的结构性缺陷:倾向于用线性思维理解指数级增长的技术能力,用短期效益评估长期影响,用局部优化忽视系统关联。概言之,“灰犀牛”AI 安全风险隐喻类同于“温水煮青蛙”的隐喻,渐进式风险在 AI 技术乐观主义叙事中不断累积,直至突破安全阈值。

(二)两大逻辑:认知不对称与决策不可逆

首先,从认知不对称逻辑下 AI 安全风险的生成来看,在自动驾驶汽车紧急避让的瞬间,AI 基于数十亿里程数据计算的概率模型选择最优路径,而人类依赖视觉经验和直觉判断。此种认知维度的错位,揭示了人机交互系统中的“认知不对称”困境,AI 系统在数据海洋中建立的“超验理性”,与人类基于社会生活的“经验理性”,正在形成新的认知断层。此种认知断层最初表现为知识生产的维度冲突,最终则可能会演化为社会运转的信任危机。第一,从知识生产的维度冲突来看,AI 通过数据分析与算法学习构建的知识体系,本质上是对现实世界的数学映射[13]。当医疗诊断 AI 从数百万病例中提炼出疾病预测模型时,其知识生成方式已突破人类医生的归纳逻辑。因而人机之间较难完全理解各自的知识生产逻辑,医生难以理解 AI 为何将某个罕见症状视为关键指标,AI 也无法解释其决策中隐含的伦理权衡。第二,从社会运转的信任危机来看,前述知识生产的维度冲突最终会演化为人机信任关系的结构性矛盾,具备自主决策能力的 AI 将面临信任危机[14]。但可以预想,在 AI 系统完全嵌入社会结构和个人生活时,由于难以摆脱 AI,人类对 AI 产生的信任危机并非只是一味猜忌,而是在盲从与猜忌之间摇摆,盲从与猜忌的界限变得模糊。放弃理解以服从技术权威,与保持敏感以提出信任质疑,可能成为并生的两种态度。

其次,从决策不可逆逻辑下 AI 安全风险的生

成来看,金融交易算法在毫秒间引发市场熔断,或军事无人机因目标识别偏差启动错误打击表明,AI系统中微小的决策偏差会在不可逆的行动链条中被无限放大,如同“蝴蝶效应”一般最终演变为系统性灾难。第一,从决策链的不可逆来看,AI系统可以实现从感知环境到执行行动的全流程自动化,此种全流程自动化在提升效率的同时,也在一定程度上切断了人类干预AI系统运行的关键节点的可能。以自动驾驶为例,当系统在0.1秒甚至更短时间内完成“识别行人—计算轨迹—执行制动”的决策链时,任何传感器误判都会直接转化为不可逆的机械动作。第二,从数字“蝴蝶效应”的演化来看,在决策链不可逆的基本情形之下,系统参数的任何细微变化都可能引发AI领域的数字“蝴蝶效应”[15]。从AI系统内部来看,AI可能通过强化学习不断固化错误模式,如同滚雪球般将初始偏差转化为固定行为。从AI系统外部来看,一方面,实时环境交互让每个决策都成为新因果链的起点,推动数字“蝴蝶效应”不断向前演化;另一方面,AI协同产生的群体智能,可能涌现出单个系统无法预测的宏观行为模式[16],扩大数字“蝴蝶效应”的影响程度。

三、一般风险的二阶防控与AI安全风险的四阶规避

如前文所述,AI安全风险具有不同于一般风险的形态特征与生成逻辑,因而AI安全风险的处置不应局限于一般风险处置思路。一般风险的处置思路表现为“过程—结果”的二阶防控,具有轻视风险主体、忽视风险未来演化过程的不足,AI安全风险的处置需对上述不足进行针对性弥补,实现由“过程—结果”二阶防控向“主体—过程—结果—过程”四阶规避的跃迁。

(一)一般风险二阶防控的样态:过程—结果

一般风险的二阶防控聚焦于“过程—结果”两大阶段。一般风险二阶防控的样态,在过程阶段,表现为面对风险演进过程决定是否发起防控;在结果阶段,表现为针对预期风险结果选择具体防控措施。首先,从过程阶段来看,在风险演进过程中,公权力机关履行现代社会风险防控的基本职责,其依据法定职权决定是否发起风险防控措施,同时发动其他社会主体参与风险防控。不同于常态情形下公权力机关依据确定性标准履行职责,风险演化过程具有不确定性,因而不能以常态情形下的标准严格束缚公权力机关,只要风险演化过程展现出危害预期,公权力机关即可依法决定发起风险防控[17]。其次,从结果阶段来看,在决定发起风险防控后,针对风险可能造成的危害结果应展开具体风险防控措施的选择。风险防控措施因价值取向的不同可划分为积极与消极两类:积极风险防控措施以发挥措施正面作用、消除风险为价值取向,侧重于对风险防控措施效率性的强调;消极风险防控措施则以抑制措施负面作用、维护权利为价值取向,侧重于对风险防控措施合法性的强调。在现实生活中,应对急剧性、突发性风险往往会首先选择积极风险防控措施,在无准备的仓促状态下以求迅速遏制风险、恢复秩序;而在处置急剧性、突发性风险的后期,由于风险演化至衰减阶段且积极风险防控措施负面效果显现,则往往会选择消极风险防控措施,修补前期积极风险防控措施造成的权利损害。

(二)一般风险二阶防控的“轻视”与“忽视”

一般风险二阶防控的不足体现在以下几个方面。一是轻视了对引发风险的主体的考察,引发风险的主体不同,风险演化过程亦不相同,针对不同风险演化过程进行过度抽象的理论提炼,即便形成统一的风险“过程—结果”防控范式,也会因缺乏对风险主体特性的分析而导致风险“过程—结果”防控范式缺少适配性,难以输出实效。如在构建生物安全、数据安全、网络安全等风险防控体制机制时,轻视对生物技术研究开发与应用主体、开展数据处理活动主体、网络运营主体等引发风险的主体的责任追究,对主体责任的追究缺乏落地措施或者主体责任畸轻,则会导致风险防控体制机制成为仅具宣示性意义的制度。二是忽视了风险结果的未来影响,当风险的危害结果成为既存事实时并不意味着风险的消灭,危害结果的出现只是风险演化过程的一个阶段性标志,从物质守恒的客观规律看,任何风险防控措施既不可能

完全修复风险造成的既有损害,也不可能彻底消除风险结果的未来影响。换言之,危害结果作为既存事实出现时,会继续触发风险下一个阶段的演化过程,使之对未来产生不确定的影响。

(三)AI 安全风险的四阶规避:主体—过程—结果—过程

针对一般风险"过程—结果"二阶防控的不足,为克服对风险主体的轻视和对风险未来影响的忽视,一是应在"过程"阶段的前端增设对风险主体的考察,使得对风险演进过程的分析和对风险危害结果的预判更具主体针对性;二是应在"结果"阶段的后端补充对风险未来影响的防控,摒弃"风险危害结果的出现意味着风险演化过程的终止"的局限性思维,继而在"过程—结果"二阶防控的基础上形成"主体—过程—结果—过程"的四阶规避。就 AI 安全风险而言,不同于二阶防控将"过程""结果"割裂为静态分离的两大阶段,四阶规避中"主体""过程""结果""过程"四大阶段动态交互,形成由"主体"到"过程"、由"过程"到"结果"、由"结果"到"过程"的衔合态势。首先,由"主体"到"过程"注重对 AI 主体特征的考察,基于 AI 主体特征分析 AI 独特风险的演进过程;其次,由"过程"到"结果"聚焦于 AI 安全风险演化过程中安全风险的预防,通过多元 AI 安全风险防控措施阻止 AI 安全风险的现实危害;最后,由"结果"到"过程"放眼于 AI 安全风险的未来代际影响,尝试通过当代制度建构阻断 AI 安全风险的代际传递,实现代际公平。

四、四阶规避理念下 AI 安全底线设计的基准考量

探讨 AI 安全底线设计,在进行制度建构之前,须将一些带有价值导向意义的基准纳入考量。所谓 AI 安全底线设计的基准,并非"头痛医头、脚痛医脚"的制度工具,而是在对 AI 安全风险生成逻辑进行"呼应式"思考和深层透视的基础上提炼出的基本原则。在四阶规避理念下,AI 安全底线设计应当考量的基准包括:第一,由"主体"到"过程"的透明性基准,以破除由 AI 主体特征导致的事后追责难题与认知困境;第二,由"过程"到"结果"的预防性基准,形成底线防控与常态可控的多元规避措施;第三,由"结果"到"过程"的阻断性基准,阻断 AI 安全风险在代际之间的传递。

(一)由"主体"到"过程"的透明性基准

透明性基准要求突破数字"蝴蝶效应""算法黑箱"等 AI 主体特征对 AI 事故"事后追责"造成的认知困境。从"事后追责"的传统范式来看,当医疗诊断 AI 因训练数据偏差给出错误治疗方案时,会出现责任应当归属于数据提供方、算法工程师还是医疗机构的疑问。从前文对 AI 安全风险的分析来看,在不可逆的 AI 决策链条中,错误往往由多个环节的微小偏差叠加导致,单一而准确的责任主体变得难以追溯。这一认知困境动摇了现代法律体系的归责理论——当作为行为主体的 AI 尚不具备法律人格,而作为设计主体的人类又无法完全预见其行为、判断其逻辑时,责任伦理的链条出现了结构性断裂。更严峻的是,即便突破了认知困境进行了归责,事后追责也无法挽回 AI 系统级失效的灾难性后果。如金融交易算法引发的市场闪电熔断(flash crash)可在极短时间内造成万亿美元级损失[18]。

(二)由"过程"到"结果"的预防性基准

预防性基准的本质,是在 AI 尚未抵达安全阈值、造成危害结果之前,通过最坏情形预设与冗余容错(redundancy based fault,概念来源于工程学)等多元措施,搭建起既能进行底线防控又能进行常态可控的容错空间。由于人类对 AI 未来演化的预测能力存在根本性局限,唯有将对 AI 安全风险的预防锚定于超越常规认知的灾难性场景,才可能为不可逆的技术失控设置缓冲地带。首先,从进行底线防控的最坏情形预设来看,最坏情形是指 AI 突破人类控制后引发的人类社会"价值理念—科学技术—社会秩序"的多维度崩溃。预防性基准要求将这些极端场景视为必然发生的确定性威胁,而非统计学意义上的小概率事件,因而 AI 安全底线设计必须预设智能体在突破安全阈值后仍持续恶性演化的最坏可能,在此逻辑之下构想终极性的防御措施。其次,从进行常态可控的冗

余容错措施来看,AI安全底线设计不能只局限于通过对最坏情形的预设来进行底线防控,最坏情形的出现往往来自日常风险的积蓄,因而还需在底线防控的基础上将防御端前移,形成常态可控的冗余容错措施。冗余容错并非从字面意义进行理解,即设置大体量的略显冗余的预防措施,“多留几手”以规避AI安全风险。其核心理念不在于复制相同的安全模块,而是构建多层次、异质化的预防措施体系。

(三)由“结果”到“过程”的阻断性基准

约纳斯的责任伦理理论表明:技术活动必须遵循“不危及后代生存条件”的绝对禁令[19]。阻断性基准要求必须阻断当下AI技术及相关产业的发展对后代生存发展产生的能源消耗、伦理债务、社会解构等具体风险。首先,就能源消耗而言,训练AI消耗的电力资源本质是碳排放,随着AI训练的逐渐升级,与此相应的碳排放逐步升高,转移给后代的减排责任亦不断递增;其次,就伦理债务而言,当前通过人脸识别、医疗AI等采集的人类生物特征数据,在未来存在被超级AI解密滥用的可能,现代人逾越伦理进行的AI技术开发成为后代人必须偿还的伦理债务;最后,就社会解构而言,AI形成的就业替代效应,在代际之间不断演变、扩散,极有可能在百年内解构人类既有社会劳动架构、重塑人类劳动价值体系[20]。

五、三大基准考量下安全底线设计的制度建构

以AI安全底线设计的基准为指导原则进行制度建构:一是应围绕透明性基准进行算法黑箱透明、人机交互透明、数据流动透明的制度建构;二是应围绕预防性基准进行安全阈值、最坏情形、冗余容错的制度建构;三是应围绕阻断性基准进行阻断能源消耗代际传递、伦理债务代际传递、社会解构代际传递的制度建构。围绕三大基准进行的制度建构均由“微观技术—宏观制度”两部分构成。

(一)围绕透明性基准的制度建构

1. 算法黑箱透明:外部的技术审计

在发生自动驾驶汽车突然制动或医疗AI误诊病例时,用户往往陷入“知其错而不知其所以错”困境。此种困境揭示了确立透明性原则的必要:当AI系统的决策过程成为无法解读的黑箱时,AI安全风险就失去了最基础的验证可能[21]。为有效破除算法黑箱,需要无利害关系的专业第三方介入进行技术审计。首先,从技术维度看,专业第三方介入进行技术审计,应当能够通过技术手段拆解算法决策链条,构建穿透算法黑箱的标准化工具包,确保审计人员能用标准化方法追溯AI决策的每个节点并输出审计结果。其次,从制度维度看,专业第三方介入进行技术审计,倒逼开发者预先建立可解释的算法架构,就像建筑规范要求预留消防通道,迫使施工方在设计阶段就将消防要求内化于其中。概言之,专业第三方介入技术审计,避免了AI算法沦为个别企业的技术特权,确保监管机构、用户和其他企业都能在统一框架下理解AI系统的运行逻辑。其不需要公开商业机密源代码,通过标准化的审计方法就能验证算法是否符合安全规范。

2. 人机交互透明:AI的意图显现

当智能导航系统突然更改路线,或聊天机器人擅自切换对话主题时,用户常会产生被系统暗中操控的不安感。此种不安的根源在于人机交互中意图传递的断裂,AI的行动意图与人类认知之间形成了理解鸿沟[22]。而人机交互透明要求AI系统通过交互界面主动揭示行为意图,使技术决策的“暗流”变为用户可感知的“明河”。首先,从技术维度来看,使AI意图得以显现需构建三层反馈机制。第一层是关键决策可视化,如同汽车转向时自动亮起的指示灯,将AI运行过程中关键节点的行动通过显著的语言或者符号向用户进行呈现。第二层是决策依据可视化,当AI根据用户的要求输出决策结果时,应同时输出做出这个决策结果的理由,将AI算法转化为用户可理解的逻辑语言。第三层是异常操作可视化,当AI感知到用户的频繁异常操作时,应主动向用户输出异常操作预警,避免错误指令的持续累积。

其次,从制度维度来看,上述三层反馈机制推动人机交互中AI意图的显现,但此种意图显现并

非单向的信息倾倒,而是构建双向校准、共同验证的安全制度。例如,在自动驾驶场景中,当车辆突然减速并显示“检测到前方突然出现快速移动物体,准备启动紧急制动”时,乘客既能理解当下决策,也可通过“误判反馈”按钮矫正当下判断,帮助 AI 系统优化识别逻辑。这种人机交互双向校准的设计,将用户从被动的承受者转变为主动的监督者,通过 AI 意图显现建立起人机共同验证的安全制度。

3. 数据流动透明:生命周期的溯源

AI 系统运行的底层逻辑是对数据的智能化使用[23],若不能追溯数据从采集到消亡的完整轨迹,AI 系统的安全性就如立于流沙之上的大厦,因而透明性原则要求对数据流动的全生命周期进行溯源。首先,从技术维度来看,一是应通过标记技术明确数据怎么“诞生”,给每项首次形成的元数据贴上“数字身份证”,记录其采集时间、用途授权和流转路径。例如,当医疗 AI 首次形成患者数据后,在后续诊疗过程中应像快递追踪般显示该数据于何时何地用于何种医疗用途。二是应通过销毁技术明确数据如何“灭亡”,采用区块链、量子信息等技术确保被删除数据真正退出数据流通系统,避免出现“删而不废”的数字幽灵潜伏风险。

其次,从制度维度来看,数据流动的全生命周期溯源,具有维护透明性的双重制度价值。一是对个体而言,数据标记技术和销毁技术有益于个体跳出算法黑箱和信息壁垒,为个体维护合法权益提供技术支撑。例如,当用户发现智能手环的心率数据被用于保险评估时,可依据溯源记录要求使用主体停止侵害隐私权并删除个人信息。二是对系统而言,当算法产生致错风险时,审计人员能通过数据溯源定位到有偏差的训练数据来源。例如,在智慧城市管理中,若交通调度 AI 出现误判,通过溯源机制可快速鉴别是实时车流数据失真,还是历史数据样本错误,以此提升纠错效率[24]。概言之,数据作为 AI 系统运行的“血液”,数据流动的全生命周期溯源,本质是为 AI 系统建立“数字血液循环”的健康监测体系。

(二)围绕预防性基准的制度建构

1. 基于安全阈值的预防制度建构

AI 安全阈值是指尚未突破人类对安全风险控制能力的具体 AI 系统的发展临界值,传统安全工程依赖静态的安全阈值,但 AI 安全阈值的静态预设无法应对技术系统的动态演进。在开放场景中,静态安全阈值要么因过度保守限制技术发展空间,要么因环境突变丧失安全防护作用,因而需要对 AI 安全阈值进行动态校准。首先,从技术维度来看,AI 安全阈值动态校准的核心价值在于构建“呼吸式”安全边界——通过多层举措获取变量,持续调整风险边界,在安全与效率间维持动态平衡。AI 安全阈值动态校准在技术操作层面可由三层举措组成:一是 AI 系统内部状态监控,获取内部变量;二是 AI 系统外部环境感知,获取外部变量;三是 AI 系统历史事件学习,获取历史变量。其次,从制度维度来看,AI 安全阈值动态校准将安全从技术参数升维为伦理价值,它要求 AI 系统设计者放弃“终极安全”的幻想,转而建立严密的责任链。一是校准算法的透明度义务,在对 AI 安全阈值进行动态校准后,必须以足够透明度接受外部审查,充分阐明 AI 安全阈值动态校准的理由,杜绝以“算法黑箱”规避审查[25];二是用户知情权的动态保障,伴随安全阈值的动态校准,应向 AI 系统用户实时显示当前安全等级;三是动态校准后的补偿机制,如 AI 安全阈值动态校准后导致系统安全等级下降、系统服务效能降级,由此对用户造成的损失应进行补偿。概言之,AI 安全阈值的动态校准是对技术不确定性的诚实回应,承认安全是行进过程而非最终状态,并通过制度设计将这种认知转化为持续改进的治理实践。

2. 基于最坏情形的预防制度建构

最坏情形预设的本质,是通过技术层的“能力封印”与制度层的“责任锁链”,在 AI 可能突破的每个维度提前建立止损点。当“能力封印”与“责任锁链”形成闭环,AI 将被打上“失控即自毁”的钢印,从而迫使其在突破人类控制边界前主动收敛。这种预设不是限制技术进步,而是为 AI 时代的发展划定不可逾越的底线。首先,从技术维度

来看,一是实时认知监测,监测系统需持续扫描AI的底层决策逻辑,而非仅监控表面输出。此种实时认知监测是从最坏情形预设角度对前文透明性基准进行的技术延伸。二是设置自我限制,在AI算法核心层植入不可绕过的伦理审查模块[26]。例如,医疗AI在给出诊断方案前,必须通过预设的伦理审查模块验证其决策对人类尊严的影响程度。三是物理阻断机制,类同于电路系统中继电保护装置触发的"跳闸"现象,应为关键AI系统配备独立于算法的实体安全装置,一旦出现极端安全风险则通过实体安全装置进行物理阻断。其次,从制度维度来看,一是能力风险分级制度,根据AI系统的认知复杂度和行动力,建立"能力—风险"对应清单。例如,能够自主编写代码的模型归入高危类别,必须接受每月认知审计,而具备物理执行能力的机器人则需向管理部门缴纳风险保证金,金额与其最大破坏力成正比。二是全球联动阻断机制,当任一区域检测到AI系统极端安全风险时,可强制冻结全球同类系统的特定功能模块,直至完成全球同类系统的安全评估方可解冻。三是开发者终身责任制,开发者团队需为AI系统的极端安全风险行为承担终身责任,倒逼企业在设计阶段必须构想这个系统的安全系数。

3. 基于冗余容错的预防制度建构

冗余容错设计的核心思想,可以理解为给AI打造"多样化安全方案",使其在极端风险发生时自动切换到不同的安全路径。就像登山时携带多条不同材质的绳索,即使某条断裂,其他绳索依然能保障安全。如前文所述,此种冗余容错设计不是简单重复设置同样的保护措施,而是创造灵活多变的应对方式。首先,从技术维度来看,应设计不同于常规运行系统的离线系统,如同停电时启用的应急照明系统,为重要AI系统准备"原始版"应急方案。如智慧城市管理系统平时使用最新算法,但必须保留一个只具备基础功能的离线版本,当最新算法出现安全风险时,通过离线版本维持供水供电等最基本服务。其次,从制度维度来看,一是建立容错资源配给制度,根据AI系统风险等级动态分配容错资源,AI系统风险等级越高,容错资源分配越多。例如,普通医疗咨询AI每回答一定数量的问题,必须强制插入1次医生人工复核,而手术机器人每台手术则必须预先留存多套应急方案。二是建立容错方案组合制度,要求AI开发者为同一风险点提供3种以上技术路线迥异的容错方案,经过第三方测试后,选择最优组合方案投入使用。

(三)围绕阻断性基准的制度建构

1. 阻断能源消耗风险的代际传递

阻断AI能源消耗风险的代际传递,关键是将"节能基因"植入AI系统[27],将其进化方向引导至"越聪明越省电"的轨道,就像智能手机在迭代升级后总会变得更轻薄,应将能源效率优化作为衡量AI迭代升级的重要指标。首先,从技术维度来看,一是进行能耗意识训练,在AI学习阶段就设置"省电考试",比如训练图像识别系统时,不仅考核识别准确率,还需评估完成任务的耗电量,由此训练AI系统自动寻找既准确又省电的识别路径,形成节能习惯。二是进行算力动态调节,让AI学会根据任务重要性调节所用电力。比如智能家居系统在白天家庭成员活动频繁时全速运转,夜间则自动切换至低耗能模式,重要决策调用全部计算资源,日常事务则使用简化版算法。三是设置物理限能装备,为AI安装不可破解的"能源制动",在运行设备中配备独立供电监测芯片,当设备能耗超标时,向用户发送过度能耗信息,避免无节制耗电。其次,从制度维度来看,一是建立绿色竞赛制度,由政府设立"节能AI创新基金",定时公布各企业AI产品的单位算力耗电排行榜,高能耗系统需缴纳阶梯性能源税,用市场机制推动企业自行研发省电算法。二是建立跨代追责制度,AI产品的能耗受制于适用领域、用户数量等因素,往往具有累积性,需要一定时期甚至跨越代际之后方可显现,因而当前AI系统开发者需为未来AI系统的整体能耗负责,为此需就AI产品的过度能耗构建对开发者的跨代追责制度。

2. 阻断伦理债务风险的代际传递

为阻断当前AI系统产生的身份歧视、隐私侵犯等问题对未来智能社会的侵蚀[28],应确保每代

AI 进化时自动“偿还”前代伦理欠账,而非累积道德风险。首先,从技术维度来看,一是伦理风险筛查,在 AI 训练时设置“伦理滤网”算法,像孕检筛查遗传病那样识别 AI 系统的潜在伦理风险,从源头杜绝身份歧视、隐私侵犯等伦理风险基因[29]。二是伦理标准更新,不同代际之间因时代差异一些伦理标准会存在差异,因而在进行 AI 系统伦理风险筛查的基础上,还需定期用最新伦理标准重新审查 AI 系统运行逻辑,淘汰过时的伦理参数,标注可能存在伦理风险的历史数据并使其逐步退出数据训练库。其次,从制度维度来看,一是设置补偿金制度,新一代 AI 系统上线需缴纳前代系统的伦理补偿金,伦理补偿金用于修复前代系统在既往运行过程中对用户造成的伦理伤害,迫使企业主动清理前代系统的伦理债务,更加谨慎考量新一代 AI 系统可能存在的道德风险。二是建立伦理信用体系,若 AI 系统出现伦理事故,则开发该系统的企业其后续产品上市前需经过更严苛的审查,且必须搭载第三方伦理监督算法,伦理信用分数直接影响企业获取算力资源等关键权益。

3. 阻断社会解构风险的代际传递

阻断 AI 社会解构风险的代际传递,需通过制度设计在技术演进中保留人类劳动的核心价值。技术进步不应切断社会结构的连续性,而应为人类劳动价值提供迭代升级的桥梁,使人类始终站在文明进化的驾驶舱。首先,从技术维度来看,一是保留核心劳动技能,为 AI 系统植入“人类技能图谱”,强制保留一定比例需人工参与的接口[30]。如工业机器人必须设置需人类手感校准的关键工序,医疗 AI 的诊断方案需包含人类医生诊断环节,以此确保人类核心劳动技能不会彻底消失。二是显现人类劳动价值,在 AI 系统中嵌入人类劳动价值计量器,如物流仓库的智能分拣系统,需实时显示“若纯人工操作需多少工时”,将效率提升值按比例转化为员工技能补贴。其次,从制度维度来看,一是设置技能迭代基金,从 AI 创造的价值中提取“人类再教育基金”。如一台汽车生产流水线机器人可以取代多名工人,则使用该机器人的企业应当资助被替代的劳动者学习新技能。二是确立人机协作认证,建立不同工作岗位的人机协作比例标准[31],比如幼儿园需保留多少比例的人类教师直面儿童,餐饮业需保证多少比例的餐品包含人工制作环节。此类认证像食品行业的“有机标识”,保护不可替代的人类劳动价值。三是预先模拟社会影响,每项 AI 系统上线前,需进行该项 AI 系统社会影响的沙盘推演。如在推广部署某一工业领域的机器人之前,必须模拟对相关工种劳动者的影响,推演如何进行劳动者的培训转岗。

六、结语

对于 AI 领域“切尔诺贝利事故”的忧思绝非杞人忧天,诚如安东尼·吉登斯所言“如果基本信任没有得以建立,或者内心的矛盾没有得到抑制,那么后果便是存在性焦虑的持续”,在人类尚未对 AI 建立基本信任之前,围绕 AI 领域安全风险形成的存在性焦虑就不可能被消除。但也正如马克思所言“机器迁就人的软弱性,以便把软弱的人变成机器”[32],如果毫无理性地对 AI 形成绝对信任,人类必将沦为 AI 的奴隶。日本动漫电影《夏日大作战》曾设想过带有自主意识的软件系统引发的全球社会秩序崩溃,只不过按照日本动漫一贯的故事叙事逻辑,影片结尾将危机的解决诉诸主角的“超神”发挥以及对反派的感化。而在现实世界,这种理想化的感性行为不可能成为化解 AI 领域安全风险的关键手段,我们仍然需要足够理性的制度应对危机。在 AI 领域安全风险显露的情境下,我们甚至应当考虑在影响社会运转的关键领域设置类似《三体》中“面壁人”的职业群体,使其生活环境与 AI 完全脱离,确保 AI 完全无法了解此类职业群体的思维动态与行动逻辑,继而保有人类反抗 AI 的最后“武器”。上述构想或许充满了科幻色彩,但未来已来,风险已与你我同在,生存还是毁灭,已不再是一个问题,而是并行的两种状态。

参考文献:

[1]吉登斯. 现代性的后果[M]. 南京:译林出版社,2011:7,87.

[2] BOSTROM N. Superintelligence: paths, dangers, strategies [M]. Oxford :Oxford University Press,2024.

[3] AMODEI D, OLAH C, STEINHARDT J, et al. Concrete problems in AI safety[EB/OL]. (2016-07-25)[2025-03-12].

https://arxiv.org/abs/1606.06565?_refluxos=a10.
[4]刘艳红.生成式人工智能的三大安全风险及法律规制:以ChatGPT为例[J].东方法学,2023(4):29-43.
[5]CATH C,WACHTER S,MITTELSTADT B,et al. Artificial intelligence and the "Good Society": the US,EU,and UK approach[J].Science and engineering ethics,2018(24):505-528.
[6]赵汀阳.人工智能"革命"的"近忧"和"远虑":一种伦理学和存在论的分析[J].哲学动态,2018(4):5-12.
[7]FLORIDI L. The European legislation on AI: a brief analysis of its philosophical approach[J]. Philosophy & technology,2021,34(2):215-222.
[8]MONARCH R M. Human-in-the-Loop machine learning: active learning and annotation for human-centered AI[M]. Simon and Schuster,2021.
[9]程海东,胡孝聪."人在回路"的道德机制:机器人伦理实践的新路径[J].河南师范大学学报(哲学社会科学版),2025,52(1):108-114.
[10]高奇琦.智能革命与国家治理现代化初探[J].中国社会科学,2020(7):81-102,205-206.
[11]张欣.生成式人工智能的算法治理挑战与治理型监管[J].现代法学,2023,45(3):108-123.
[12]刘瑞.防范"灰犀牛""黑天鹅"风险事件[J].人民论坛,2019(6):77-79.
[13]张钹,朱军,苏航.迈向第三代人工智能[J].中国科学:信息科学,2020,50(9):1281-1302.
[14]孔祥维,王子明,王明征,等.人工智能使能系统的可信决策:进展与挑战[J].管理工程学报,2022,36(6):1-14.
[15]朱体正.人工智能辅助刑事裁判的不确定性风险及其防范:美国威斯康星州诉卢米斯案的启示[J].浙江社会科学,2018(6):76-85,157.
[16]王易,王成良,邱国栋.群体智能学习型决策:"大数据+AI"赋能的决策范式演化研究[J].中国软科学,2024(12):35-50.
[17]苏宇.风险预防原则的结构化阐释[J].法学研究,2021,43(1):35-53.
[18]Selling spirals: avoiding an AI flash crash[EB/OL].(2024-11-08)[2025-03-12].https://www.lawfaremedia.org/article/selling-spirals--avoiding-an-ai-flash-crash?_refluxos=a10.
[19]张旭.技术时代的责任伦理学:论汉斯·约纳斯[J].中国人民大学学报,2003(2):66-71.
[20]关乐宁,徐清源.人工智能时代的劳动关系变迁与治理:基于机、劳、资、政四方关系的视角[J].中国劳动关系学院学报,2024,38(2):77-91.
[21]徐凤.人工智能算法黑箱的法律规制:以智能投顾为例展开[J].东方法学,2019(6):78-86.
[22]沈伟伟.算法透明原则的迷思:算法规制理论的批判[J].环球法律评论,2019,41(6):20-39.
[23]吴汉东.人工智能时代的制度安排与法律规制[J].法律科学(西北政法大学学报),2017,35(5):128-136.
[24]张新长,华淑贞,齐雱,等.新型智慧城市建设与展望:基于AI的大数据、大模型与大算力[J].地球信息科学学报,2024,26(4):779-789.
[25]陈劲,朱子钦,季与点,等.底线式科技安全治理体系构建研究[J].科学学研究,2020,38(8):1345-1357.
[26]汪怀君,汝绪华.人工智能算法歧视及其治理[J].科学技术哲学研究,2020,37(2):101-106.
[27]孙亚军.AI能耗困扰并非无解[N].经济日报,2024-04-25(4).
[28]HUTSON M. Even artificial intelligence can acquire biases against race and gender[EB/OL].(2017-04-13)[2025-03-13].https://www.sciencemag.org/news/2017/04/even-artificial-intelligence-can-acquire-biases-against-race-and-gender.
[29]郑智航.人工智能算法的伦理危机与法律规制[J].法律科学(西北政法大学学报),2021,39(1):14-26.
[30]贾计东,张明路.人机安全交互技术研究进展及发展趋势[J].机械工程学报,2020,56(3):16-30.
[31]胡晟明,王林辉,赵贺.人工智能应用、人机协作与劳动生产率[J].中国人口科学,2021(5):48-62,127.
[32]马克思.1844年经济学哲学手稿[M].北京:人民出版社,2018:120.

(本文责编:默　黎)