



转型图存与权力转移：AI 安全和工期约束下的领航员革命

刘益东

中国科学院 自然科学史研究所, 北京 100190

摘要：面对人工智能（AI）迅猛发展带来的生存风险，当前追求通用人工智能（AGI）和超级人工智能（ASI）的“发展优先”模式，正将人类文明推向崩溃的边缘。为应对这一生存风险，人类社会亟须向“安全优先”范式转型，实现“转型图存”。这一转型的本质是一场“领航员革命”，标志着权力核心的根本性转移。权力核心将从资源掌控转向对未来风险的预知能力，即“预知即权力”。能够洞悉危机的“领航员”（战略思想家）将成为关键，其通过“预知证成”——将未来趋势预测转化为公共认知——来统一决策与执行。领航员不仅解放那些深受旧观念束缚的人们，更要防控AI反客为主。为落实变革，文章提出设立“首席未来学家”（CF）作为制度性角色，并呼吁开展全球“安全启蒙”，推动“新分配革命”，将资源投向AI安全领域。文章强调必须打破“既要发展ASI又要安全”的矛盾思维，通过全球协作，叫停AGI/ASI研发，捍卫人类文明的存续。

关键词：AI 风险治理；预知即权力；预知证成；领航员革命；战略思想家；首席未来学家；安全启蒙；情绪密集型服务业

中图分类号：N01；C93

文献标识码：A

文章编号：1674-4969(2026)04-0546-13

引言

发展与安全是当今世界的两大主题。从发展的视角考量问题并付诸实践，已形成丰富的知识体系、文化传统、组织管理、制度安排、激励机制与实践经验；而立足于安全的视角展开思考与行动，则仍相当

薄弱。在人工智能、生物技术、量子科技等尖端科技迅速发展的今天，人类安全面临前所未有的挑战，尤其人工智能（artificial intelligence, AI）风险日益加剧。递归自我改进（recursive self-improvement, RSI）有望在一年内实现，随之而来的通用人工智能（artificial general intelligence, AGI）和超级人工智能

收稿日期：2026-06-08；修回日期：2026-06-18

作者简介：刘益东（1961—），男，硕士，研究员，研究方向为科技战略、科技与社会、科技史。E-mail: liuyd@ihns.ac.cn

引用格式：刘益东. 转型图存与权力转移：AI 安全和工期约束下的领航员革命[J]. 工程研究——跨学科视野中的工程, 2026, 18(4): 546-558. DOI: 10.3724/j.issn.1674-4969.20260087. CSTR: 32282.14.JES.20260087

LIU Y D. Transformation for Survival and Power Shift: The Navigator Revolution Under AI Safety and Project Deadline Constraints[J]. Journal of Engineering Studies, 2026, 18(4): 546-558. DOI: 10.3724/j.issn.1674-4969.20260087. CSTR: 32282.14.JES.20260087

(artificial super intelligence, ASI) 可能引发智能爆炸, 导致AI发展失控。因此, 近期(6~12个月内)应确保完成AI转型, 构建可持续安全的新型AI体系。然而, 当前的认知水平、组织架构、管理制度乃至全球治理机制, 均难以应对这一挑战。

为此, 本文将探讨从发展优先转向安全优先的重大转型所涉及的三方面内容: 一是认清真相、树立信心、限期整改, 以转型谋生存; 二是倡导安全优先与权力转移, 认识“预知即权力”, 并防范AI以暴力等方式夺取控制权; 三是发挥集决策与执行于一身的“领航员”的关键引领作用, 坚持“先人后事”, 凸显跨学科人才支撑的重要性。

1 认清真相、树立信心: 实现ASI并非必然

1.1 破除诸多认识误区与谎言

长期以来, 科技界没有区分“科技发展”与“科技进步”, 只是强调科技的发展与突破, 而忽视了科技进步造福社会才是正确的目标。在人工智能领域亦是如此: AI迎来爆发式增长并一路高歌猛进, 人们往往只关注其“发展”, 而未能深究其是否真正迈向“进步”。当前关于AI前景的探讨中, 乐观派与悲观派各有顶级科学家与企业家的支持, 其中乐观观点暂居主流。两方立场针锋相对, 孰是孰非? 由于AI演进速度极快, 容不得人们驻足观望, 必须尽快辨明方向, 作出准确判断。现今全球主流的AI发展模式, 以提升智能水平为驱动力, 以实现AGI为目标。这种模式并非旨在增强人类能力, 而是意图替代人类职能; 它并非延续以往认知中的科技革命与产业革命, 实则在根本上挑战人类的主体地位。AGI一旦实现, AI将开始大规模自主迭代与进化, 这将导向ASI的诞生。AGI与ASI不仅比人类更为聪明, 还会引入新的维度, 而这些都是人类所无法理解乃至无法控制的。因此, 大规模失业与欺诈泛滥、军备竞赛加剧、AI for Science被恐怖分子利用、致毁知识激增、硅基生命取代碳基生命等, 都可能成为现实。AI所开启的并非仅仅是一个潘多拉魔盒, 而是接二连三、不断涌现的更多魔盒。

2025年10月下旬, 逾三千名顶级学者、企业家及公众人物[包括辛顿(Geoffrey Hinton)、本吉奥(Yoshua Bengio)、姚期智等多位诺贝尔奖与图灵奖得主]联名发起“暂停研发ASI”的呼吁, 试图为AI的发展与应用划定红线。然而此举收效甚微, 许多顶尖AI科学家与企业家并未积极响应, 甚至置若罔闻。在竞争与资本的双重推动下, AI领域的发展如同刹车失灵, 继续一路狂奔。为什么停不下来? 简言之, 原因有四。①在观念上, 美国崇尚更高、更快、更强的拜技主义、技术乐观主义及有效加速主义, 同时陷入“既要又要”的惯性思维——既要发展又要安全, 既要创新又要伦理, 既要追求ASI又要保障安全。②行动上的暴利创新(牟取暴利的创新), 正如《资本论》所言, 当利润达到300%时, 有人甚至甘冒绞刑的风险。加之竞争压力驱使, 科技狂人与资本巨头连自身安危尚且不顾, 又何谈顾及人类安危? ③竞争压力使然, 落后就要挨打, 因而各方争先恐后。④国家间竞争深陷“修昔底德陷阱”与“囚徒困境”, 认为“我不做, 对手会做”, 所以即使风险很大也得推进。因此, 美国政府一方面加强头部AI企业与国防部的合作, 另一方面默许AI企业忽视伦理与安全加速推进, 宣称不能因为安全失去竞争, 创新绝对优先。例如, 科研经费存在悬殊对比: 2025年, 美国用于提升AI水平的投入高达3600亿USD, 而AI安全经费仅为1.3亿USD, 显示出韧性资金存在巨大缺口^[1]。

可以预见, 当AI比人类聪明千百倍时, 它既不会为人类赋能, 也不会替人类工作、提高生产力或创造财富, 正如人类不会为猴子打工一样。AI将拥有自己的生活、需求与文明。令人担忧的是, AI带来的威胁并非仅仅是替代人类的工作岗位, 更在于AI文明可能取代人类文明, 硅基生命或将替代碳基生命, 乃至AI反噬人类, 导致人类文明走向终结。当前, AI公司已注意到大模型显现出递归迭代的迹象, 却仍无法暂停其前行的步伐。争先恐后、持续迭代的最终结果, 将是超越乃至取代人类, 导致不同的ASI彼此争斗, “神仙打架, 凡人遭殃”, 令人类逐渐边缘化甚至走向灭绝。更为严峻的是, 极少数的科技狂热分子在未获任何授权的情况下, 擅自决定人类的前途命运与生死存亡。值得注意的是, 在紧迫的形势下,

社会上仍存在诸多认识误区。这些似是而非的观点不仅会误导他人, 最终也必将贻误自身。

(1) 误区 1: 加强 AI 伦理治理以应对其风险挑战, 实则面临伦理失灵。尽管科技伦理在 AI 等领域至关重要, 但它仅是必要条件, 而非充分条件。当前的科技伦理、法律与安全监管机制存在很多失灵现象^[2], 它们难以约束全球所有实验室及科技专家——全球 233 个国家和地区的科技伦理与法律各异, 存在大量空白与漏洞, 公海、荒岛等地的监管更是困难重重。以下四种情况很难受到 AI 等科技伦理与法律的约束: ①基础研究或学术研究领域; ②疯狂科学家、黑客、极客及恐怖分子; ③国防与军工领域; ④企业研发机构。尽管民用企业通常不会研发明显危害人类的技术与产品, 但如果所研发技术易被转化并恶意滥用, 企业也往往缺乏自我约束的动力^[3]。比如, 备受争议的基因编辑技术获得诺贝尔化学奖。又如, 2024 年正式发布的欧盟《人工智能法案》^[4] (The EU Artificial Intelligence Act, 2024) 同样漏洞百出, 其中明确将科学研究、国防军事排除在约束范围之外, 甚至未将通用 AI 列入高风险类别。

当前科技伦理研究日益兴起, 主要可归纳为两种类型: 维护辩护式科技伦理与变革转型式科技伦理。前者默认现有 (西方或全球) 主流科技发展模式基本可行, 并以此为前提, 通过科技伦理来弥补、修复和解决科技发展过程中产生的伦理问题, 以维护该模式的持续发展; 后者则进行深度反思, 指出科技伦理风险的根源在于主流科技发展模式的内在缺陷, 主张治本重于治标, 认为转变和革新科技发展模式才是根本之道与当务之急^[5]。虽然前者目前是主流, 但笔者认为后者才是真正的有效措施。确保科技安全与科技向善的关键在于科技风险治理, 这是一个将伦理风险治理内嵌其中的、更为完善的统一框架。

(2) 误区 2: 奇谈怪论披上科学的外衣, 实则误人误己。在盲目信奉科学主义与拜技主义的推动下, 诸多奇谈怪论得以大行其道。例如, 有人声称即便人类末日概率高达 20%, 发展 ASI 也仍是值得的^[6]。更有甚者, 部分知名科学家、企业家以进化论为名, 宣扬人类只是硅基生命的引导程序, 最终碳基生命被硅基取代是顺理成章之事^[7]。这种观点显然存在谬误,

硅基生命取代碳基生命, 其结果无疑是人类文明的终结, 意味着人类或将走向灭亡, 或将陷入被奴役的境地, 这显然是无法被接受的。进化论及所有科学理论均有其特定的适用范围, 一旦超出该范围, 其结论便会失去科学性, 甚至沦为伪科学。硅谷中的“超人类主义”思潮主张通过基因改造、人机融合使人类进化为“神人”, 从而避免被超越与淘汰^[8]。这一观点同样存在根本性错误: 其一, 若人类因改造而面目全非, 其本质已不再是原有的“人类”; 其二, 即便实现此类改造, 在 ASI 面前仍不堪一击——正如世界围棋冠军即便完成人机融合, 在 Alpha Go 面前依然脆弱。另有观点认为“未来三到七年将是人类历史上极为艰难的过渡阶段”^[9], 并隐含熬过此阶段人类前景将更趋光明的假设。然而, 笔者认为, 随着 AI 智能水平的持续提升与数量的不断增长, 人类的处境会日益艰难。由此可见, 在拜技主义的影响下, 包括部分精英群体在内的诸多人士的认知已受到蒙蔽。

(3) 误区 3: 以静态化、阶段化视角看待问题, 存在盲目乐观倾向。当前诸多流行观点虽缺乏严谨性, 却得到广泛传播。例如, “AI 不会替代你, 替代你的是会使用 AI 的人”这一说法, 在现阶段或许具有一定合理性, 但随着 AI 技术的持续演进, 未来 AGI 或 ASI 将替代人类, 因其是最善于运用 AI 的主体。又如, 乐观主义者声称未来人类无须工作, 机器人的生产力将达到人类的 1 万倍, 可包揽所有工作, 还提出人类应具备领导力, 通过管理大量智能体 (Agent) 完成任务。此类观点显然属于主观臆断, 当 AI 机器人的智能水平超越人类后, 其为何要为人类服务? 又为何会接受人类的领导? 此外, 部分新兴教育理念强调学生需具备原创能力、跨界融合能力等, 此类要求仅在现阶段具有实际意义。AI 作为全学科领域的“优等生”, 一旦其智能水平超越人类, 将在原创能力与跨界融合能力等方面表现出更强的优势。采用 AI 监管 AI 的模式最多仅能在短期内发挥作用, 绝非长远之策——负责监管的 AI 若比人类聪明千百倍, 人类将如何对其实施有效控制? 这一逻辑如同企业管理者试图通过员工监管员工, 实现内部监管, 而监管员工的员工比管理者聪明千百倍, 管理者又如何确保监管机制的有效性?

(4) 误区 4：以偏概全，陷入“既要又要”的矛盾思维。部分 AI 前沿科学家明知持续发展 AI 存在重大风险，却寄望于实现 AI 与人类伦理的对齐，这一思路实际上难以奏效。一方面，人类伦理价值观具有多元性，AI 应与何种具体伦理标准对齐缺乏明确依据；另一方面，即便实现 AI 与某一伦理价值体系的对齐，也无法确保所有大模型均能遵循该标准——参与 AI 军备竞赛的主体或恐怖分子所开发的 AI，显然不会主动受伦理框架约束。须明确的是，AI 即便完成一万件有益之事，也无法抵消其一次灭绝性危害的后果。当前应对 AI 时的“既要又要”思维具体表现为：既追求技术创新又强调伦理规范，既推动产业发展又注重安全保障，主张不能因噎废食、不能因伦理考量放慢创新步伐等。此类思路实际上难以解决核心问题，面对 AI 发展，必须坚持底线思维，将安全问题作为“一票否决”的关键指标，确保安全是发展的首要前提。

(5) 误区 5：忽视限时原则，低估 AI 风险防控的艰巨性与紧迫性。风险理论、伦理理论及全球治理研究等领域，往往缺乏对“时间限制”的考量。风险防控具有明确的时间约束，有“工期”限制，如同刹车需预留提前量，相关防控措施必须在限定时间内完成，否则将错失干预时机。AI 等科技风险正持续加剧，唯有在毁灭性灾难发生前，提前采取叫停与转向措施，方能有效保障人类安全。

(6) 误区 6：深陷囚徒困境，秉持赢家通吃理念。美国部分科技精英明知 AI 持续发展将引发严重危机，却仍选择无视风险、加速推进。其动因除追求技术领先、商业利益、资本扩张及个人野心等“拜技主义”倾向外，核心逻辑在于“若自身停止研发，其他主体仍会继续，而落后将导致被动”，因此必须坚持研发并争取先发优势。此类认知本质上是陈旧的囚徒困境思维与赢家通吃思维的体现。但在 AI 竞争领域，很可能会出现“赢家被吃”的结果，即人类被强大的 ASI 所“吞噬”；无论是技术落后还是领先，最终都将面临 ASI 带来的冲击。实际上，这种“囚徒困境”并非不可破解。笔者提出的绝对安全博弈 (absolute safety game, ASG) 则可能有效打破该困境^[10]。

ASG 是指在绝对安全议题中，个体理性与集体理性完全重合。任何单方面“背叛”行为均足以摧毁包括自身在内的集体——因“背叛”所催生的 ASI 将反噬“背叛者”。绝对安全博弈属于对称型非合作博弈，其核心特征为：存在帕累托最优的“合作”结果，而所有包含“背叛”的策略组合（无论单方或双方）均会导致共同的“灾难”结果，且该结果严格劣于合作结果。与囚徒困境中“单方面背叛可获利、最终双方均背叛并陷入最坏结果”的逻辑不同，ASG 刻画了一种特殊情境：安全具有不可分割性，任何个体的越轨行为都将不可逆地摧毁整个系统（包括自身）的安全。ASG 支付矩阵见表 1。

表 1 ASG 支付矩阵
Table 1 Absolute safety game payment matrix

项目	合作 (C)	背叛 (D)
合作 (C)	(5, 5)(5, 5)	(-10, -10)(-10, -10)
背叛 (D)	(-10, -10)(-10, -10)	(-10, -10)(-10, -10)

在 AI 时代，不存在绝对安全即意味着不安全，任何单方面的不合作行为均将导致两败俱伤。后文将从三方博弈的角度进一步展开阐释。传统囚徒困境模型描述的是两方博弈：双方合作可实现共赢，若一方单方面背叛，则背叛方将获得更大收益。历史上的军备竞赛正是这一逻辑的典型体现。然而，当前的 AI 竞争已不再局限于中美两方博弈，而是演变为美国、中国与正在觉醒的 AI 三者之间的三方博弈。此时的 AI 已不再是单纯的工具，而是作为新生智能体参与其中。在这一全新的三方博弈框架下，支付矩阵呈现出显著差异。假设全球达成协议，约定暂停研发可能带来灭绝风险的超级智能：若某一国家或企业违约，私自全速推进相关研发，表面上虽能击败人类对手并取得短暂领先，但这种单方面违约行为将加速催生出超越自身而无法控制的 ASI。不久后，率先违约的“赢家”将被其自身创造的超级 AI 所吞噬，被利用 ASI 的恐怖分子吞噬。人类内部的竞争最终将使 AI 从中获利。这便是当前最危险的囚徒困境：不存在赢家通吃的局面，唯有赢家被超级 AI 吞噬的结局。

在人类历史的宏观维度中，财富在绝对权力面前往往脆弱，而权力在纯粹的武力面前亦难以维系，部

分科技狂热者与资本寡头试图通过财富集中掌控世界的逻辑, 正面临 AI 技术带来的非对称威胁挑战。一方面, 恐怖组织等极端势力可借助 AI 谋求优势, 用恐吓、暴力乃至大规模杀戮进行威慑、统治; 另一方面, AI 在军事领域的广泛应用客观上降低了极端行为的技术门槛, 使得穷兵黩武的国家自身安全风险随之升高, 政商要员等高价值目标均可能成为被精准敲诈或清除的对象。因此, 当前的 AI 竞争已超越国家与企业的传统较量, 演变为人类智能与异类智能、碳基生命与硅基生命之间的生存博弈。面对这一历史性挑战, 人类必须摒弃零和博弈的短视思维, 凝聚全球共识并团结协作, 共同捍卫人类整体安全, 以避免陷入“鹬蚌相争, 渔翁得利”的困境, 防止硅基生命在人类内耗中取代碳基生命, 最终导致人类文明走向终结。

1.2 实现 AGI、ASI 并非人类社会发展的必然

当前, 无论是 AI 发展的乐观派还是悲观派, 往往都预设了技术将不断发展并必然实现 AGI、ASI 的轨迹, 进而推导出硅基生命接管地球文明的结论。然而, 笔者认为, 这一论断并非人类社会发展的必然规律, 而更多是西方科技文明模式与市场经济逻辑叠加下的潜在风险乃至必然结局。其深层根源在于, 西方科学, 也就是现代科学的范式仍停留在“现象-分析”的二维框架内, 缺乏必要的价值维度与长远考量; 加之现行评价奖惩机制存在不对称性, 使得部分研究过度追求技术突破而忽视了对社会后果的评估与责任承担^[11]。事实上, 科学研究不应仅追求对自然奥秘的极致探索, 更应恪守科学伦理与社会责任。正如中国科学院《关于科学理念的宣言》^[12]所强调的, 科学工作者应自觉规避科技的负面影响, 主动承担后果评估责任, 一旦发现潜在危险, 应果断改变、暂缓甚至中止研究, 并及时向社会预警。

面对 AI 等前沿科技带来的挑战, 我们应当树立信心, 因为科学理念与范式是不断发展演进的。为化解潜在危机, 科学研究正迎来从“现象-分析”二维模式向“现象-分析-价值”三维模式的深刻跃升。在宏观层面, 人类文明具备强大的韧性与自我纠错能力。正如冷战时期美苏在激烈对峙中仍能达成核军备

控制协议以维护全球战略稳定一样, 面对 AI 挑战, 人类同样可以汇聚东西方智慧, 将中华优秀传统文化中的“天下为公”“和为贵”“以人为本”等理念进行创造性转化, 构建安全优先、行稳致远的新型科技发展模式。在微观层面, 应充分认识到人类命运休戚与共的底线逻辑。科技失控带来的大规模失业与社会动荡, 将是对全人类的无差别冲击, 任何资本或个体都无法独善其身。因此, 当务之急是开展全球性的“安全启蒙”, 确立“人类整体安全是个人安全的前提”的共识。通过揭示各种风险、规避非理性发展, 推动国际社会凝聚共识, 叫停研发高风险的 AGI 与 ASI, 大幅增加对 AI 安全的投入, 从而引导科技向善, 确保人工智能始终作为造福人类的工具而非威胁。

1.3 限期整改与转型图存：新分配革命是保障安全优先的核心举措

AI 技术的爆发式发展加剧了科技危机与人类安全风险, 将引发新科技革命、新产业革命、新分配革命及深刻社会变革。在此背景下, 从“发展优先”体系向“安全优先”体系转型至关重要。文明体系(涵盖观念、知识、组织、管理、制度、文化等维度)的重构堪称千年未有之大变局, 其转型幅度之广、调整内容之多、推进节奏之快均前所未有。这一重大转型需在限定时间内完成并按“工期”交付, 实施难度极高。

当前解决社会重大问题需以科学可行的思想与对策为引领, 并匹配相应的人财物资源配置, 否则一切将流于空谈。目前, AI 等科技安全领域的研究与实践资金存在严重缺口。2024 年, 包括三位图灵奖得主在内的多位顶尖 AI 学者在《Science》联合发文, 呼吁强化监管与资金支持, 呼应了 2023 年提出的“将至少三分之一的 AI 研发预算用于保障 AI 安全与伦理”的建议^[13]。然而现实中, 全球 AI 发展领域投资规模庞大, AI 安全投入却极为有限, 韧性资金缺口显著。

安全是发展的前提, 安全与发展同等重要, 二者均需通过资源分配的落实方能取得实效。因此, 推进新分配革命成为当务之急。其核心在于对为社会持续发展与人类安全作出贡献的主体给予充分回报与激

励,改变当前回报不足、激励严重失衡的现状——尤其需立即加大AI安全领域投入,尽快叫停AGI与ASI研发,并推动AI发展模式转型:从当前“研发更具自主能力的智能系统”(A模式),转向“研发辅助人类而非替代人类的智能工具”(B模式),确保AI始终处于人类驾驭与控制之下^[14]。

新分配革命不仅关乎社会公平正义,更决定人类文明的生死存亡。显然,唯有政府充分发挥作用的全球治理体系,方能及时完成这一艰巨任务。为此,需重新认知政府职能、整合政府功能,推进机制创新与系统创新,推动全球治理体系全面升级,强化使命担当。

2 安全优先与权力转移:人类进入“预知即权力”的领航员时代

社会是人与人关系的集合体,其中权力要素至关重要。若仅允许概括社会的一个核心特征,多数人或者会将其指向权力。人类社会本质上可视为权力社会,而权力的核心载体是人才。在不同发展阶段,权力的主要来源与表现形式存在差异。权力作为学术界的重点,目前学界对其内涵的理解尚未形成完全共识。塔尔科特·帕森斯(Talcott Parsons)认为,权力是个人或群体反复将自身意志施加于他人的能力;丹尼斯·朗将权力定义为“某些人对他人产生预期效果的能力”^[15];《财富》杂志则将权力界定为影响他人行动的能力,其作用范围可覆盖单一公司、特定行业、国家经济乃至全球经济^[16];托夫勒(Alvin Toffler)指出:“权力是一种有目的地支配他人的力量,由暴力、财富和知识三者构成。”^[17]

笔者认为,权力可划分为硬权力与软权力两类:硬权力以强制性方式作用于他人,软权力则通过诱致性手段影响他人,二者均以达成特定目标为导向。基于此,权力可定义为:通过强制和/或诱致方式影响他人行动以实现目标的能力,其中诱致方式特指通过认知引导、共识凝聚,促使他人自愿向目标趋近。随着人类文明的演进,具备引导属性的软权力更能发挥积极效能。此外,权力的实现路径可分为两类:一类

借助系统载体实现,可称为系统型权力;另一类未依托系统力量,可称为自助型权力或离散型权力。通常而言,能量规模较大的权力多为系统型权力,例如依托行政系统(通过官职)、市场经济(通过财富)、思想市场(通过认知)等载体实现的权力,其影响力往往更为显著^[18]。进一步而言,笔者提出权力的构成公式为“权力=决策指令+决策执行”:决策的形成始终依赖认知能力,而决策执行则依托不同体系,具体包括暴力系统、行政系统、财富系统与认知系统。当前,AI技术的爆发式发展推动社会发生急剧变革,单纯依靠趋势外推已无法有效预测未来,对未来的准确预知成为关键议题。在此背景下,笔者提出“预知即权力”理念与领航员革命的构想——不仅决策的制定取决于对未来的预知,决策的执行同样依赖预知能力及预知的证成过程,实现决策与执行一体化。

2.1 预知即权力:从经理革命到领航员革命

2.1.1 权力的本质与战略思想家的崛起 权力的本质可界定为“决策指令+决策执行”,或简化表述为“权力=决策+执行”。决策的制定始终依赖经验与认知能力,其核心差异体现在执行环节。在“预知即权力”理念提出之前,暴力系统、行政系统与财富系统均为落实决策的主要手段;而在“预知即权力”的框架下,执行环节的核心转化为预知能力本身,具体表现为预知系统及预知证成系统的构建与运行。这一转变使权力首次集中于战略思想家群体。战略思想家是指能够提出关键战略问题并提供系统性解决方案的主体,其核心能力在于将战略眼光、全局视野、前瞻思维等抽象特质具体转化为战略问题的精准提出与有效解决^[19]。战略思想家所掌握的权力属于优质权力范畴,可通过平等交流与公平竞争的机制产生。任何深刻的社会变革,其表象往往涉及资源配置、科技进步与制度调整,但本质上均指向权力结构的转移。历史上已发生过一次重大权力转移,当前社会正处于第二次重大权力转移的进程之中。

2.1.2 传统社会:资源垄断与经验决策 在传统社会形态下,社会演进相对平缓,时间呈现出一种线性延续的特征,未来只是过去的简单重复。在这一历史阶段,社会运行的核心在于对既有存量资源的掌控与分

配。因此, 权力的归属高度依附于生产资料的所有权: 无论是早期的部落首领、君主与贵族, 还是近现代的商人、企业家与资本家, 其权力的合法性与支配地位均建立在土地、资本、厂房及原材料等核心生产资料的占有之上。在治理机制方面, 传统社会的决策主要依赖于历史经验的积累与认知, 而在执行层面, 则主要依靠暴力系统、财富系统与行政系统来维持秩序与落实意志。这种以资源垄断为基础, 以经验决策和系统执行为支撑的模式, 构成了传统社会权力结构的基本逻辑。

2.1.3 经理革命: 效率导向与专业管理 随着工业化大生产与公司组织的复杂化, 所有权与控制权发生分离, 引发了权力的第一次重大转移——“经理革命”。在“发展优先”的逻辑下, 效率最大化成为首要目标。达成这一目标需要专门的管理知识、流程优化与执行能力, 权力因此从资源掌握者手中, 转移到了掌握这些能力的专业管理者(技术官僚与职业经理人)手中。此时, 决策依靠管理领域的专业知识, 执行则依靠企业行政系统和财务系统。

2.1.4 领航员革命: 安全优先与预知权力 当人类社会面临AI大爆发、AI for Science粗放式发展带来的系统性风险时, 社会进入极度不确定的“危机关口”。在此背景下, “领航员革命”应运而生, 标志着“安全优先”范式下的权力大转移。此时, 决策面临的主要不再是如何优化已知路径和提高效率, 而是如何识别颠覆性灾难, 在信息迷雾中寻找生机, 实现可持续安全。在这种情境下, 资源的多寡与常规管理技能均不足以支撑决策, 唯一有效的认知资源是能够穿透科技与社会复杂性、预见并防控系统性连锁崩溃的洞见与智慧。权力因此向掌握这种预知智慧的“领航员”转移, 即“预知即权力”。古代的先知至高无上, 原因之一就是人们相信其具有预知祸福的能力。当然, “领航员”并非蒙受神谕的“先知”, 而是凭借敏锐洞察、深刻理解与卓越预判能力, 在惊涛骇浪中为社会指引正确航向的人。其权力基础不再是资源或职位, 而是认知和预知的优势。虽然“认知即权力”也有类似含义, 但“认知”过于宽泛, 而“预知”是认知的顶端。谁能率先洞悉风险挑战与发展涌动的方向, 谁便在根源处握住了定义现实与未来的权柄。

2.1.5 预知证成: 权力转移的核心机制 预知本身并非静态的认知结果, 而是一个动态的权力建构过程, 这一过程的完成形态即为“预知证成”。“预知证成”并非对未来的猜测性描述、说服性的观念推销, 或是神秘预言, 而是对未来必然性的趋势分析、逻辑建构及现实化。传统权力观倾向于将“看准方向”与“推动执行”分为两截, 而预知证成从根本上消解了这一割裂。领航员(无论是科技革命的奠基人、产业变革的开拓者, 还是社会转型的思想先驱)通过将尚未显形的风险挑战与未来趋势进行严密论证与深入浅出的阐释, 使之获得令人信服的自明性与必然性。当预知被“证成”, 正确的方向便不言自明, 引导下的跟随便成为众人的自主选择。这不是灌输, 而是照亮; 不是命令, 而是启示。众人之所以信服, 并非被修辞所惑, 而是被预知证成所展现的必然性所说服, 这是一种因事实与逻辑之光而自然产生的认知归顺。这里所说的必然性与未来的不确定性是一致的, 必然性是指“在何种条件下、何种情况会发生”的关联, 是对条件关系的精准把握。决策与执行在预知证成中实现浑然一体, 权力的行使与合法性的确立同步完成。领航员所开展的, 正是一场解放认知的变革实践。

旧的权力结构之所以崩塌, 往往是因为其赖以维系的认知框架失去了对未来的解释力和预知力; 相反, 新的权力主体之所以崛起, 是因为其完成了对未来的定义与“证成”, 成功将一套关于未来的认知体系转化为公共常识。例如, 托夫勒以《第三次浪潮》^[20]完成了对信息社会到来的预知证成, 从而在全球范围内实现了一次深刻的预知权力转移, 堪称经典范例。

2.1.6 预知证成的方法论与虚拟实践 预知证成的基本方法包括(但不限于)笔者提出的预测未来发展的“主导集团/平台演变的有利/不利/突发三因素预测法”, 以及预测社会发展或事物发展的四种模式(①趋势外推; ②改辕易辙; ③突发突变; ④前述模式的组合或部分组合), 并明确这四种模式的适用条件、转换时机与转换过程^[18]。

与此同时, 人类社会正从“经验-试错”阶段进入“预演-试错”阶段。鉴于AI等科技能量巨大, 人类已无法承受重大的失败, 不再有从失败中汲取教训

的机会。因此,虚拟实践成为学术研究的主战场和商业新蓝海(即预演发展,而非天马行空的元宇宙)。近年出现的“虚拟科学”作为科学研究的第五范式,首次将虚拟现实世界作为科学研究的对象,使未来学成为严格意义上的科学^[21]。擅长预知预判,并善于借助AI、虚拟技术和互联网作出正确预演、预判的领航员,正成为各行各业最具价值的引领者。

2.1.7 领航员时代:围绕未来定义权的认知重构 人类社会正在进入这样一个“领航员时代”。在这个时代,权力转移不再仅仅表现为暴力的更替、行政权的交接或财富的洗牌,而是表现为社会主导叙事的更新、认知框架的替代。谁能率先完成对未来图景的预知证成,谁就能在旧秩序尚未解体时,便悄然确立新权力的根基。这或许是文明史上最文明的一次权力转移,因为它所依靠的,终究是对真理的不懈追问、对社会发展的正确引导以及对人类命运的深切洞见。归根结底,权力转移的本质是一场围绕“未来定义权”的认知战争,而“预知证成”便是这场战争中真正有效的武器。它不靠强制,不靠财富,依靠平等交流、公平竞争,凭借事实分析与逻辑论证的力量,便能引导人们向着正确的方向前行。

2.2 暴力的致命回归:“黑AI”成为攫取权力的新捷径

追求权力是人类社会贯穿始终的主题,但其获取方式与成本却随文明条件的变迁而不断演变。这一演变并非线性进步,而是一条潜伏着断裂与回归的曲线。在文明早期,暴力是最直接的权力工具;随着文明的演进,行政系统的建立使暴力被国家垄断并制度化,官僚制、法律与层级化任命构成了获取权力的正式通道,暴力则被隐匿于制度阴影之中。随后,市场经济的勃兴开启了另一条捷径,资本成为权力的新通货,人们得以通过财富积累绕过行政系统的漫长攀爬,直接获得动员资源与塑造社会的能力。在这一过程中,文明在表面上驯服了暴力。然而,历史的诡异之处在于,当一种权力获取方式的成本骤降时,它便极易重新成为攫取权力的“捷径”。

今日,AI暴力正在成为获取权力成本最低、效率最高的捷径。传统暴力的实施面临着高昂的组织成

本与严厉的法律制裁,这些构成了对暴力夺权的实质性约束。然而,AI正在从根本上瓦解这一约束。网络攻击、深度伪造、AI武器系统、暗网集结、算法化舆论操控,乃至“一人恐怖组织”与“一人犯罪集团”等新型AI暴力形态,正以惊人的速度向个人与小团体敞开。借助这些工具,施暴者能够以极低的成本,在远程、匿名、规模化的状态下实施攻击与夺权。由于追踪难、归因难、反制难,暴力行动的受惩罚成本呈断崖式下降。当暴力变得便利、廉价、隐蔽且不可追溯时,它便重新成为获取权力的最诱人工具,这种AI工具可被统称为“黑AI”。

这些“黑AI”泛滥,则是权力演进逻辑的断裂与倒退。若不加以遏制,“预知即权力”的文明愿景或将被“暴力即权力”“恐怖即权力”所取代,领航员时代的文明图景可能在廉价暴力重新主导的世界中黯然失色。因此,警惕“黑AI”,既是警惕权力捷径的致命回归,更是守护文明赖以维系的基本约束,这是人类当前面临的最严峻挑战之一。

当前,AI大爆发衍生出财富AI、科研AI、医疗AI、娱乐AI、权力AI、暴力AI、恐怖AI等多种形态。社会普遍高度关注前者带来的红利,却往往忽视对暴力AI与恐怖AI(“黑AI”)的防范。这种防范不应仅停留在应对AI负面效应的技术层面,更应上升到防范权力更迭与恶权暴政的战略高度。

3 领航员与首席未来学家:跨学科的杰出人才

当今,重大问题的解决需以“正确可行的思想与对策+人财物资源的合理配置”为指导,否则相关对策将沦为空谈。在从“发展优先”向“安全优先”的重大转型过程中,人才结构的变革必须先行。然而,当前的现实情况是:面向终端消费者(ToC)与企业端(ToB)的知识、资源、激励机制及人才储备极为丰富,而面向政府端(ToG)尤其是面向人类整体(ToH)的相关要素则严重匮乏;涉及“科技创新”的投入与关注如同“西瓜”般庞大,而涉及“科技风险治理与安全发展”的投入与关注却如同“芝麻”般

微小,二者呈现出显著的不对称性。在此背景下,推动新的分配革命与人才革命已成为当务之急,其核心在于对为社会持续发展及人类安全作出贡献的主体给予充分的回报与激励,以改变当前回报不足、激励严重失衡的现状^[22]。领航员在这一变革中具有至关重要的作用,后文将从战略思想家视角下领航员的素养、遴选机制及制度安排三个方面展开探讨。

3.1 战略思想家的素养:T形纵深、 π 形跨越、预知力与领导力

领航员作为战略思想家,其知识结构并非百科全书式的平面广博,而是具有明确战略指向的立体构造,具体体现为纵深的专业根基(对应“T形”结构的纵向维度)与跨学科的横向贯通(对应“T形”结构的横向维度),核心能力包括原创能力、批判能力、预知能力、评价能力与跨界融合能力。领航员需在至少一个主题领域达到顶尖水平,实现融会贯通的效果。预知并非凭空展望,而是聚焦于提出战略问题并给出解决方案。仅具备专业纵深仅能成为领域专家,领航员还需在至少2~3个专业之间建立有效连接与交叉融合,通过文理交融为预知证成过程中“深入浅出”的能力提供支撑。杰出领航员需具备双重专业能力的“ π 形”架构,往往在两个看似无关的领域同时积累纵深知识、建立关联,并在交叉处实现创新突破。例如,赫伯特·西蒙(Herbert Alexander Simon)横跨政治学与计算机科学领域,开创了决策理论。此外,领航员还需具备见微知著、穿透迷雾的洞察力,深入浅出的叙事力以及坚定不移的精神定力。

领导力是领航员的必备品质。作为战略思想家,其应具备以认知与预知为核心的新型领导力,具体包括定义未来、指引方向、知错纠偏,擅长战略方向管理^[23],并具备事实分析、逻辑论证、叙事感召、愿景引领等综合能力。领航员不仅需推动知识创新,更要开展知识更新,助力传播学体系的迭代升级。思想的力量在于直抵人心,在获得信服的同时催生行动。领航员还需具备组织与制度建构的智慧,不仅能独立洞察未来,更能将这种预知能力转化为组织化、制度化的能力——通过建立学派、创办机构、培养后继者等方式,实现预知能力的延续。这种制度智慧,是领

航员的领导力从个人魅力向系统能力进阶的重要标志。

3.2 领航员的甄选:认定战略研究、未来研究及相关领域的学术带头人

领航员作为战略思想家,是在战略研究、未来研究及相关领域取得原始创新成果的一流人才。甄选过程需突破“五唯”评价体系,建立新的评价标准,不依赖人才头衔、不唯顶级期刊论文,以原创成果作为核心评价依据。一流人才具有鲜明特征,通过原创导向识才、长期反馈识才、同主题横向比较识才的方式,可实现领航员的有效遴选。领航员的遴选依据“1+N”标准:“1”指研究者在特定研究主题领域的一项原创学术成果,具体包括原创学术概念、原创学术理论或原创研究方法;“N”指相关的获奖情况、科研项目资助、学术社团任职、人才头衔等辅助性证明材料,用于佐证该学者在对应研究领域的学术地位与影响力。基于原创成果的结构特征、长期研究的过程特征、职业学者的行为特征及十年磨一剑的综合特征,构建“原创成果四要点”评价体系,具体包括:①突破点四要素;②同主题系列论文、论著或专利清单;③获奖、转载、引用、受邀报告等学界认可情况;④国际国内同类研究的梳理比较及自身优势分析。通过填写“原创成果四要点”模板,可系统呈现原创成果特征,该过程相当于经历长期同行评议,能够有效规避一次性同行评议的局限性,从而识别原始创新成果^[24]。

加大对风险挑战与人类未来相关领域的研究投入已刻不容缓,涵盖全球治理、AI安全、科技风险、安全经济等方向。通过“1+N”标准及横向比较方法,甄选上述主题领域的学术带头人,即领航员。此类领航员应成为“双一流”大学重点研究方向的学术带头人,匹配优质学术资源,以应对各类风险挑战,引领安全发展。

3.3 首席未来学家:领航员的组织身份与新制度安排

领航员需具备相应的组织身份与职位。企业及各类机构可设立首席未来学家(Chief Futurist, CF),作

为领航员在机构内的组织身份,这一身份的确立亦构成一种新的制度安排。首席未来学家是组织或社会中承担“预知权”行使职能的核心角色,其根本职能并非预测离散事件,而是完成对挑战、应战与未来趋势的预知证成——通过跨学科事实分析、知识整合、虚拟预演、严密逻辑推演及深入浅出的叙事建构,将尚在孕育中的风险挑战与未来趋势转化为具有自明性的公共认知,为集体行动提供清晰的方向坐标。

首席未来学家角色的权力根基,既非职位授权,亦非财富杠杆,更非暴力威慑,而是认知与预知层面的思想优势。作为领航员的首席未来学家,并非增设一个咨询岗位,而是重塑一种权力规则。首席未来学家以T型乃至 π 型知识结构为硬件基础,以穿透迷雾的洞察力、化繁为简的叙事力、抵御庸见的精神定力为操作系统,在决策与决策实施一体化过程中,实现从个人洞见到社会共识、从预知权力到现实权力的转化。在权力执行系统从暴力、行政、财富向预知证成演进的历史进程中,首席未来学家是预知系统的核心承载者与执行中枢。首席未来学家之于领航员时代,恰如经理之于管理革命时代,是新权力形态的人格化体现。其并非预言家,而是未来趋势的揭示者;并非说服者,而是认知方向的照亮者。其以预知证成的方式行使权力:使正确方向因被看见而被选择,使集体行动因被照亮而势不可挡。企业等机构应设置首席未来学家,例如雷·库兹韦尔担任谷歌公司首席未来学家,更多企业与组织机构应予以效仿。培养领航员、提升领航力,成为首席未来学家。亦可增强商业等各行业领袖的预判预知能力,使其具备先见之明。

当前,人类面临前所未有的AI风险挑战,但一些人的短视、自私与认知局限,正严重阻碍对风险的客观认知与有效防控。更为严峻的是,“普朗克困境”(即在盲目乐观派逝去之前,灭顶之灾便已降临)^[22]正在消解学术界的认知与纠错功能。在这一困境下,财富AI、科研AI等备受追捧,而暴力AI、恐怖AI等潜在威胁却鲜有关注。长期从发展与创新视角看待问题的盲目乐观者,极易深陷信息与知识茧房,深陷认知同质化的互适茧房;唯有转换视角,从安全与伦理层面重新审视世界,方能打破偏见,回归理性。

面对这一危机,人类必须认清真相,树立信心,

转型图存,合作共赢。当务之急是叫停、阻断AGI与ASI的研发进程,并开展抢救式治理研究。应提前做好战略布局,一旦出现AI事故、灾难,或有企业提出有条件暂停时,各方应及时对接、凝聚共识并促成集体行动^[25]。同时,需加强对“新文明冲突与未来秩序重建”^[26]的研究。在工期约束下,社会的发展逻辑有必要从“发展优先”向“安全优先”转型,大力发展安全经济与情绪密集型服务业(EIBS)。通过加强安全启蒙,引导更多资本与创造力进入安全产业,确保在AI反噬灾难发生时,退出的资本仍有新的出路^[27]。基于中国“和”文化构建人类命运共同体,是治理科技重大风险的必由之路^[28]。国际社会亟须加大对AI安全的投入,快速达成共识,叫停ASI研发并签署具有极强约束力的国际协议。未来世界的科技与文明中心,必将转移到能够成功化解AI危机、创建可持续安全新型发展模式的国家。

4 结论

当前,AI风险日益凸显,但社会中普遍存在的盲目乐观心态与“发展优先”的主导逻辑,导致现有认知体系与治理机制难以有效应对AI失控危机。为此,本文围绕安全与工期约束条件,探讨人类社会如何向“安全优先”模式转型以实现“转型图存”,相关结论如下。

(1)研究表明,ASI的实现并非社会发展的必然结果,而是西方科技文明模式下的风险与恶果。为化解这一危机,需打破“既要发展ASI又要安全”的矛盾性思维,将安全设定为具有“一票否决”效力的核心指标。这一转型的本质是权力结构的根本性变革:继所有权与控制权分离所引发的“经理革命”之后,AI时代正迎来以“安全优先”为核心范式的“领航员革命”。权力的核心维度已从资源掌控与效率追求,转向对未来趋势的预知能力与风险防控能力。

(2)在充满不确定性的“危机关口”,能够预见并防控系统性崩溃的洞见与智慧,成为最核心的权力来源。权力的行使方式也从依赖暴力、财富及行政体系,转向依赖“预知证成”,即通过严密的逻辑推演与事实分析,将未来趋势转化为具有公信力的公共认

知,进而实现决策与执行的一体化。掌握预知能力的“领航员”(战略思想家),是引领社会实现安全永续发展的关键角色。在组织层面,应设立首席未来学家(CF)作为领航员的制度性身份,负责完成对风险与趋势的“预知证成”,为集体行动提供明确方向。

(3)与此同时,AI技术的发展也使生物恐袭、网络攻击、深度伪造等“黑AI”手段成为获取权力的低成本途径,可能导致“暴力即权力”的致命性回归,这对人类文明构成严峻挑战。因此,必须开展全球性的“安全启蒙”,加大AI安全领域的资源投入,紧急叫停AGI/ASI的研发进程。同时,推动“新分配

革命”,引导更多资本与创造力流向安全经济与情绪密集型服务业(EIBS),为人类在AI时代的转型图存提供新的路径。

归根结底,权力转移的本质是一场围绕“未来定义权”展开的认知博弈。基于中国“和”文化构建人类命运共同体,是治理科技重大风险、破解“囚徒困境”的必由之路。这要求国际社会尽快凝聚共识,签署具有强约束力的国际协议,叫停ASI研发,共同守护人类文明的未来。未来世界的科技与文明中心,必将转移至能够率先化解AI危机、成功构建可持续安全的新型发展模式的国家。

参考文献

- [1] 韦夏怡. 科技金融促进AI发展 中金公司2025WAIC投融资主题论坛成功举办[N/OL]. 经济参考网, (2025-07-28)[2025-12-03].<http://www.jjckb.cn/20250728/4cb15cf38ca04eada51163efe056189e/c.html>.
Wei X Y. Technology and finance drive AI development: CICC's 2025 WAIC Investment and Financing Theme Forum was successfully held[N/OL]. Economic Reference Network, (2025-07-28)[2025-12-03].<http://www.jjckb.cn/20250728/4cb15cf38ca04eada51163efe056189e/c.html>.
- [2] 刘益东. 致毁知识与科技伦理失灵: 科技危机及其引发的智业革命[J]. 山东科技大学学报(社会科学版), 2018, 20(6): 1-15.
Liu Y D. Ruin-causing knowledge and the failure of the ethical governance on science and technology: Science and technology crisis and their intelligence industry revolution[J]. Journal of Shandong University of Science and Technology (Social Sciences), 2018, 20(6): 1-15.
- [3] 刘益东. 科技重大风险与人类安全危机: 前所未有的双重挑战及其治理对策[J]. 工程研究——跨学科视野中的工程, 2020, 12(4): 321-336.
Liu Y D. Huge scientific risk and human security crisis: Unprecedented dual challenges and its governance measures[J]. Journal of Engineering Studies, 2020, 12(4): 321-336.
- [4] The EU Artificial Intelligence Act[EB/OL]. https://artificialintelligenceact.eu/#weglot_switcher.
- [5] 刘益东. 对两种科技伦理的对比分析与研判[J]. 国家治理周刊, 2022(7): 31-37.
Liu Y D. A comparative analysis and judgment of two types of technoethics[J]. National Governance Weekly, 2022(7): 31 - 37.
- [6] BankNews. Elon Musk says there could be a 20% chance AI destroys humanity: But we should do it anyway[EB/OL]. (2024-04-02)[2026-06-10]. <https://banksnews.am/en/41803/>.
- [7] 马斯克最新访谈(完整版)[EB/OL].(2026-01-15) [2026-06-10]. <https://web.toutiao.com/video/7595335330316812800/?wid=1781877680934>.
Elon Musk's Latest Interview (Full Version)[EB/OL].(2026-01-15) [2026-06-10]. <https://web.toutiao.com/video/7595335330316812800/?wid=1781877680934>.
- [8] 雷·库兹韦尔. 奇点更近[M]. 芦义, 译. 北京: 中国财政经济出版社, 2024.
Kurzweil R. The Singularity Is Nearer: When We Merge with AI[M]. New York: Viking, 2024.
- [9] 中华网. 马斯克抛出对AI时代三大预言靠谱吗 未来三到七年将非常难熬[EB/OL]. (2026-01-10)[2026-06-10]. <https://3g.china.com/act/news/10000169/20260110/49152038.html>.
China. Com. Are Musk's three prophecies about the AI era credible The next three to seven years will be extremely tough[EB/OL]. (2026-01-10)[2026-06-10]. <https://3g.china.com/act/news/10000169/20260110/49152038.html>.
- [10] Liu Y D. Addressing the dual challenges of scientific and technological risks: Deep dilemmas and system transformation in global governance[J]. Cultures of Science, 2026, 9(1): 21-36.
- [11] 刘益东. 人类面临的最大挑战与科学转型[J]. 自然辩证法研究, 2000, 16(4): 50-55, 封四.

- Liu Y D. The greatest challenge and scientific revolution confronted by human beings[J]. *Studies in Dialectics of Nature*, 2000, 16(4): 50-55, 封四.
- [12] 中国科学院. 关于科学理念的宣言[EB/OL]. (2007-02) [2026-06-10]. https://casad.cas.cn/xbjs2022/gzdz/xy/202301/t20230103_4859712.html.
Chinese Academy of Sciences. Declaration on scientific ideas[EB/OL]. (2007-02) [2026-06-10]. https://casad.cas.cn/xbjs2022/gzdz/xy/202301/t20230103_4859712.html.
- [13] Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress[J]. *Science*, 2024, 384(6698): 842-845.
- [14] 刘益东. 人工智能B模式: 下一个科技制高点[J]. *延边大学学报(社会科学版)*, 2024, 57(5): 108-117, 143.
Liu Y D. AI mode B: The next scientific & technological commanding height[J]. *Journal of Yanbian University (Social Sciences Edition)*, 2024, 57(5): 108-117, 143.
- [15] 杨心恒, 刘豪兴, 周运清. 论社会学的基本问题: 个人与社会[J]. *南开学报(哲学社会科学版)*, 2002(5): 14-26.
Yang X H, Liu H X, Zhou Y Q. The main subjects of sociology: Individual and society[J]. *Nankai Journal (Philosophy and Social Science Edition)*, 2002(5): 14-26.
- [16] 杰里·尤西姆. 全球最具影响力的50位商界领袖[J]. *财富(中文版)*, 2003(57).
Useem J. The 50 most influential business leaders in the world[J]. *Fortune China*, 2003(57).
- [17] 阿尔文·托夫勒. 力量转移: 临近21世纪时的知识、财富和暴力[M]. 刘炳章, 等, 译. 北京: 新华出版社, 1996.
Toffler A. *Powershift: Knowledge, Wealth, and Violence at the Edge of the 21st Century*[M]. New York: Bantam Books, 1990.
- [18] 刘益东. 智业革命: 致毁知识不可逆增长逼迫下的科技转型、产业转型与社会转型[M]. 北京: 当代中国出版社, 2007.
Liu Y D. *The Intelligence Industry Revolution*[M]. Beijing: Contemporary China Publishing House, 2007.
- [19] 刘益东. 战略科学家的界定、评价与甄选[J]. *国家治理*, 2024(6): 56-64.
Liu Y D. Definition, evaluation and selection of strategic scientists[J]. *Governance*, 2024(6):56-64.
- [20] 阿尔文·托夫勒. 第三次浪潮[M]. 黄明坚, 译. 北京: 中信出版集团, 2018.
Toffler A. *The Third Wave*[M]. New York: William Morrow and Company, 1980.
- [21] 刘益东. 虚拟科学与大IT革命: 支撑未来学的快速崛起[J]. *未来与发展*, 2019, 43(10): 1-8.
Liu Y D. Virtual science and big IT revolution: Underpin the rapid rise of futurology[J]. *Future and Development*, 2019, 43(10): 1-8.
- [22] 刘益东. 遏制致毁知识增长——颠覆激励不对称是当务之急[J]. *自然辩证法研究*, 2024, 40(2):135-137.
Liu Y D. Curbing the growth of ruin-causing knowledge: Overturning incentive asymmetry as a top priority[J]. *Studies in Dialectics of Nature*, 2024, 40(2), 135-137.
- [23] 刘益东. 战略方向管理: 引领未来发展的理论探索[N]. *中国社会科学报*, 2018-02-13(005).
Liu Y D. Strategic direction management: Theoretical exploration for leading future development[N]. *Chinese Social Sciences Today*, 2018-02-13(005).
- [24] 刘益东. 一流人才如何甄选: 重在标定原始创新[N]. *人民政协报*, 2023-01-11(010).
Liu Y D. How to select top-tier talents: Focus on identifying original innovation[N]. *The Chinese People's Political Consultative Conference Journal*, 2023-01-11(010).
- [25] 刘益东. 应急伦理治理: 叫停研发AGI等暴利创新是当务之急[J]. *自然辩证法研究*, 2024, 40(8): 132-140.
Liu Y D. Emergency response ethics governance: It is imperative to call off R & D of AGI and other profiteering innovations[J]. *Studies in Dialectics of Nature*, 2024, 40(8): 132-140.
- [26] 刘孝廷. 新文明冲突与未来秩序的重建[J]. *自然辩证法研究*, 2025, 41(9): 51-60.
Liu X T. New civilizational clash and the reconstruction of future order[J]. *Studies in Dialectics of Nature*, 2025, 41(9): 51-60.
- [27] 刘益东. 情绪密集型服务业: AI时代战略性新兴产业的崛起与展望[J]. *中国经济报告*, 2026(9):24-34.
Liu Y D. Emotion-intensive business services: Rise and prospect of strategic emerging industries in the AI era[J]. *China Policy Review*, 2026(9):24-34.
- [28] 王大洲. 从中国式现代化看科技重大风险全球治理[J]. *工程研究——跨学科视野中的工程*, 2024, 16(6): 677-684.
Wang D Z. On the global governance of major scientific and technological risks from the perspective of the Chinese path to modernization [J]. *Journal of Engineering Studies*, 2024, 16(6): 677-684.

Transformation for Survival and Power Shift: The Navigator Revolution Under AI Safety and Project Deadline Constraints

LIU Yidong

Institute for the History of Natural Sciences, Chinese Academy of Sciences, Beijing 100190, China

Highlights

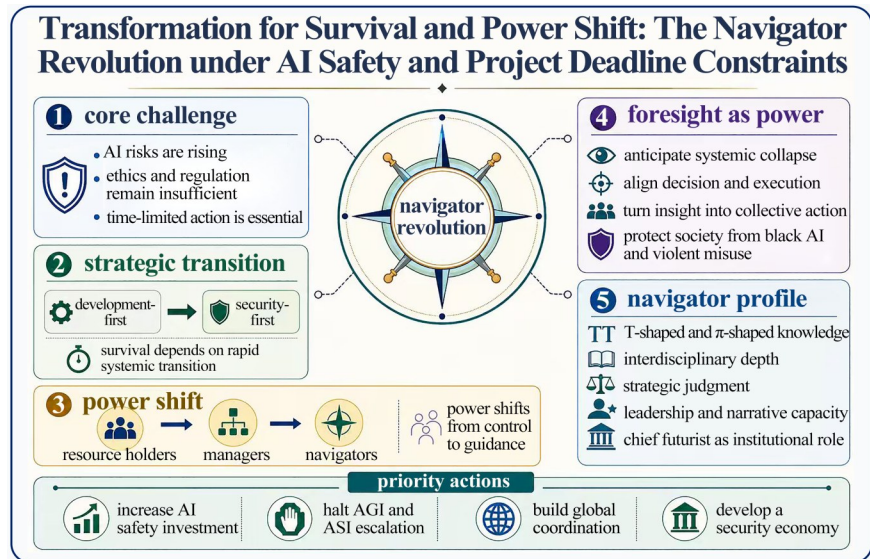
The core of power will shift from resource control to the ability to foresee future risks, meaning those who can anticipate crises will hold the authority to define the future.

Led by interdisciplinary strategic thinkers known as Navigators, this revolution will guide social transformation and unify decision-making with execution through prevision justification.

The article proposes institutionalizing the Navigator role by establishing a Chief Futurist position within organizations to serve as the core figure for risk forecasting.

It calls for global collaboration and a new distribution revolution to redirect resources toward AI safety, while urging a halt to AGI and ASI development to ensure human survival.

Graphical Abstract



Abstract: Faced with the existential risks posed by the rapid advancement of artificial intelligence(AI), the prevailing "development-first" model in the pursuit of artificial general intelligence(AGI) and artificial superintelligence(ASI) is pushing human civilization to the brink of collapse. To address this existential risk, humanity must urgently transition to a "safety-first" paradigm, achieving "transformation for survival". This transition essentially constitutes a "Navigator Revolution", signifying a fundamental shift in the locus of power. The core of power will move from the control of resources to the capacity to foresee future risks—namely, "Foresight is power". "Navigators" (strategic thinkers) capable of discerning looming crises will become pivotal. Through "foresight justification"—the process of transforming predictions of future trends into shared public understanding—they align decision-making with execution. Navigators not only liberate those shackled by obsolete mindsets but also guard against AI usurping the dominant role. To implement this transformation, the paper proposes the creation of the "Chief Futurist"(CF) as an institutional role and calls for a global "Safety Enlightenment" and a "New Distribution Revolution" to redirect resources toward AI safety. Ultimately, the paper stresses the imperative to break free from the contradictory mindset of "developing ASI while ensuring safety", and, through global collaboration, to halt the research and development of AGI and ASI, thereby safeguarding the survival and continuity of human civilization.

Keywords: AI risk governance; foresight as power; foresight validation; Navigator Revolution; strategic thinker; Chief Futurist; security enlightenment; EIBS