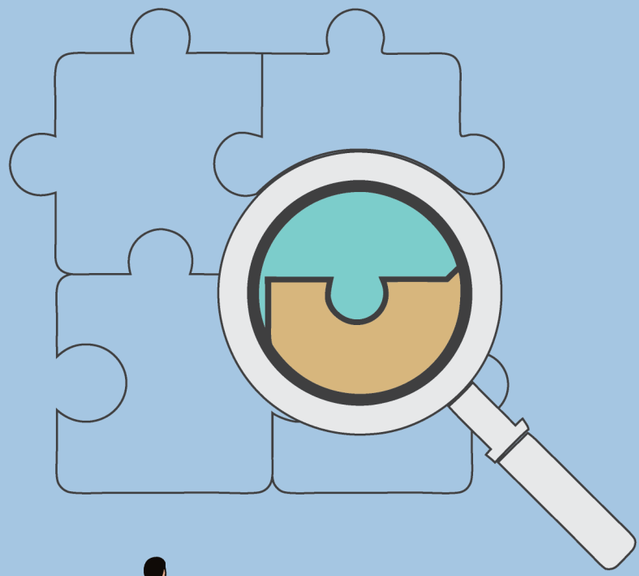
别 册

超级智能时代的风险转变与伦理应对

放心与放大：智能向善的双重纬度

杨庆峰 等著



复旦大学科技伦理与人类未来研究院

目　录

一、面对智能社会的理念：“让人放心”与“把人放大”

人工智能的飞速发展与全域应用，已将人类社会带入全新的智能时代。技术在释放科技与经济双重红利的同时，其向通用智能演进的过程也伴随不可忽视的系统性风险，如何锚定人本核心、确立适配智能社会的技术发展理念，成为当下亟待回应的时代命题。

1.1 AI 技术红利与通用智能前景

从科技史的演进脉络来看，当代人工智能技术，是继蒸汽机革命、电力革命、信息技术革命之后，人类社会第四次工业革命的核心驱动引擎，更是人类文明史上首次实现对核心认知能力的规模化、工程化延伸与拓展，具备前所未有的范式革命意义。前三次工业革命的核心突破，是通过机械力、电力与可编程信息系统，完成对人类体力劳动的替代、放大与标准化重构，其赋能边界始终局限于物理世界的劳动改造与效率提升；而 AI 技术的跨越式发展，首次将技术赋能的范畴延伸至人类心智世界的符号处理、逻辑推理、模式识别与复杂决策等核心认知领域，打破了 “高阶智力活动是人类专属能力”的长期认知边界，实现了科技发展史上的里程碑式跃迁。[①]

作为具备全链条赋能能力的通用技术体系，AI 的大规模产业化应用，正在全球范围内重构生产函数、重塑产业分工体系、释放长期增长动能，成为后疫情时代全球经济复苏与可持续发展的核心支撑。从宏观经济增长维度来看，AI 技术通过对生产、分配、流通、消费全环节的智能化改造，实现了全要素生产率的系统性提升。国际权威机构测算显示，生成式 AI 与人工智能技术的规模化应用，有望每年可带来万亿美元级的新增价值[②]，推动全球 GDP 高速增长。[③]

正是基于 AI 在科技革命与经济社会发展层面的巨大价值与突破性进展，全球人工智能研发正持续突破专用场景的能力边界，从垂直领域的定向赋能，逐步向具备跨领域迁移能力、自主学习能力与通用问题解决能力的人工通用智能（AGI）加速演进。以 GPT-4 为代表的大语言模型，已展现出跨学科、多模态的通用问题解决能力，被学界认为是 AGI 时代来临前的早期火花[④]。然而，技术能力的指数级跃迁与应用边界的扩张，很大程度也会伴随人机关系的深度重构与系统性风险的显现。当前 AI 应用中已初现端倪的认知不透明性、算法权力失衡、人类主体性弱化等伦理挑战，并不会随着技术能力的升级自然消解，反而会在 AGI 时代被进一步放大，形成关乎人类社会运行秩序、个体权利保障与文明发展走向的根本性挑战，也为我们当前的技术治理与伦理建构提出了更为紧迫的时代命题。

1.2 人类面临的AI风险

如今，警示人工智能存在安全风险的声音正通过各种形式呈现出来[⑤]；国际学界部分声音呼吁加强监管，防止失控风险，然而 3 年过去，这些实验室或许关闭，但是

① Brynjolfsson E, Mitchell T, Rock D. What can machines learn and what does it mean for occupations and the economy? C//AEA papers and proceedings. 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association, 2018, 108: 43-47.

② GOLDMAN SACHS. The generative AI revolution: Implications for growth, labor markets, and equities R/OL. (2024-03-20) 2026-05-20

③ MCKINSEY & COMPANY. The economic potential of generative AI: The next productivity frontier R/OL. (2023-06-14) 2026-05-20

④ BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of artificial general intelligence: Early experiments with GPT-4 R/OL. arXiv preprint arXiv:2303.12712, 2023.

⑤ 中华人民共和国外交部。全球人工智能治理倡议 EB/OL. (2023-10-20) 2026-05-20.

没有产生明显的效果，模型之间的竞争势态更加严峻，这反而催生了新的风险[①]；2025 年 10 月，全球科学家等签署了禁止超级智能研发的声明[②]，而超级智能正在引发诸多风险，“生存论问题”“存在论风险”“灾难性风险”成为三个标识性概念[③]；部分科学家和有影响力人士开始在全球呼吁各国政府 2026 年要应对人为制造的流行病、广泛的虚假信息、包括儿童在内的大规模个体操纵、国家和国际安全隐患、大规模失业以及系统性侵犯人权等风险。[④] 在这些风险的呈现形式中，研究报告相对严谨可靠。最为有名的是 2 个，一是《全球风险报告 2026》（The Global Risks Report 2026）中准确地勾勒人工智能从前沿技术向系统力量的转变，人工智能风险是所有风险中跨期上升幅度最大的领域，其风险体现出系统性特征，[⑤] 这种分析强调了使用技术实践；二是约书亚·本吉奥（Yoshua Bengio）《2026 国际 AI 安全报告》(International AI Safety Report 2026) 认为 AGI 新型风险可以划分为恶意使用、故障性和系统性等三类，这份报告强调的是技术本身。[⑥] 这些报告将技术风险 - 风险治理作为基础框架。

相关理论研究已指出，人工智能发展不宜仅以连续进化叙事进行描述。技术演进过程中存在关键断裂，包括从工具形态向智能体形态的转变、从知识辅助向存在介入的转变。该类断裂意味着伦理讨论不能停留在结果可控层面，还需要进入关系正当性层面。[⑦] 在这一条件下，为了防范和化解这些潜在风险和长远的风险，理论界与产业界也在思考应对之道。腾讯研究院在中国《新一代人工智能伦理规范》所强调的增进人类福祉、尊重生命权利、坚持公平公正等基本原则的基础上，在 2026 年 2 月进一步提出“让人放心，把人放大”[⑧]， 这一理念延伸了过去“科技向善”的避险思维，强调在智能化浪潮中，人类主体性应当得到最大程度的尊重与赋能。

不过，缺少放心条件的放大，可能形成高效但申诉能力不足的技术秩序。缺少放大目标的放心，可能形成低风险但能力受限的保守治理。两者应当同时成立，前者提供最低伦理条件，后者提供发展方向。

① https://futureoflife.org/open-letter/pause-giant-ai-experiments/。该公开信发起于 2023 年 3 月 22 日，截至 2026 年 2 月 5 日，该声明签字人数超过 3 万。

② https://superintelligence-statement.org/zh。该声明发起于 2025 年 10 月 22 日，截至 2026 年 2 月 5 日，该声明签字人数超过 13 万。

③ 杨庆峰，超级智能灾难性风险及多元应对，延边大学学报 (社会科学版) . 2025 ,58 (05) : 99-111+143

④ GLOBAL CALL FOR AI RED LINES. Global Call for AI Red Lines EB/OL. (2025-09-22) 2026-05-20.

⑤ WORLD ECONOMIC FORUM. Global Risks Report 2026 R/OL. (20260114) 20260520.

⑥ BENGIO Y, et al. International AI Safety Report 2026 R/OL. (2026-02-03) 2026-05-20.

⑦ Van Lingen M N, Giesbertz N A A, van Tintelen J P, et al. Why we should understand conversational AI as a tool J. The American Journal of Bioethics, 2023, 23(5): 22-24.

⑧ 王丰 .2026 科技向善创新节：AI 时代“让人放心，把人放大”. 新华社，2026.01.27

1.3 “让人放心”的伦理内涵

让人放心首先是一种关系性的规范状态，而不是技术性能指标的集合。其核心关切在于：当智能系统深度进入社会实践时，相关主体是否仍能把握系统判断的依据，是否拥有介入与纠正的权利，以及在损害发生时是否存在可识别的责任结构。该内涵超出安全与合规的技术语义，指向人机关系中的主体性维持与责任秩序重建。

在这一意义上，放心以可理解性为起点。可理解并不等同于完全透明，而是要求关键判断能够以可被相关主体把握的方式呈现其依据与边界。缺乏可理解性会直接削弱责任判断的可能性，也会降低个体对系统结论的理性接受能力。相关研究指出，若仅以工具论视角理解智能系统并预设其线性优化，容易遮蔽从工具向拟主体转化的关键断裂，从而误判解释与理解的必要性。[①] 可理解性并不足以独立构成放心，还需要可介入性作为实践保障。可介入强调人在关键节点拥有提出异议、请求复核或中止流程的权利，体现为人机关系中的最终裁决权安排。例如在 AIGC 内容治理、金融风控等高风险场景中，由于此类场景中算法决策多基于深度学习等“黑箱型”模型，其决策过程隐遁于技术隐层之后，易导致归责、监督与权益救济的困境。系统对用户内容的违规判定、信贷申请的风险拒绝等算法决策，必须配套完善的人工复核机制与算法解释支撑，这既是算法解释制度体系化构建的关键环节，也是破解算法“黑箱”、实现风险治理的重要路径。[②] 当系统运行导致实际损害时，放心还必须落实为可追溯的制度结构。可追溯并非简单归咎，而是要求责任链条、权责分配与纠错机制能够被识别与启动。技术系统的复杂性往往导致责任主体被稀释或转移，因此需要在治理结构中明确责任承担主体与责任承担路径。例如欧盟等地区正在实践一种基于风险分级的“多层治理机制”（Multi-layered governance approach），其核心在于建立分布式责任（Distributed Responsibility）与混合决策架构。在医疗或金融等高风险 AI 应用场景中，制度并不试图寻找单一的原因，而是将责任链条拆解并前置，要求开发者必须落实通过设计实现可解释性。[③] 由此可见，放心并非单一维度的要求，而是由可理解、可介入与可追溯构成的规范结构。该结构既提供人机关系的最低伦理条件，也为后续标

① 杨庆峰 . 人工智能连续论反思与存在论探析 J. 中国社会科学 ,2025,(11):149-164+207

② 苏宇 . 算法解释制度的体系化构建 J. 东方法学 ,2024,(01):81-95

③ Tatipamula S. The ethics of AI decision making: When should machines be accountable J. World Journal of Advanced Engineering, Technology and Sciences, 2025, 15(1): 0315.

准倡议奠定基础，其目标不是消除风险，而是建立可以证明的正当性关系，使技术系统始终处于可理解、可介入、可追溯的规范框架之内。

1.4 “把人放大”的规范边界

把人放大指向人类能力与价值的扩展，但其伦理合理性并不自动成立。若放大仅被理解为效率提升或功能替代，容易使人被降格为技术系统的附属部分，从而偏离伦理目标。因此，放大的正当性必须建立在明确的边界之上：一方面确认技术增强的合理空间，另一方面防止增强转化为异化或降级。

这一边界首先体现在主体性层面。增强的方向应当以人的主体性地位为前提，技术应被限定为服务于人类目的的手段，而不是重塑目的结构的主导力量。当技术系统在决策、评价或分配中形成封闭的优化逻辑，人的主体性地位可能被弱化，放大便可能偏离其本来目标。伦理上的放大因此要求保持人的主体性作为最后依据。

主体性要求进一步引出能力层面的边界。增强应当扩展人的行动能力与认知能力，但不能以削弱判断与反思能力为代价。若技术在关键任务中取代人的判断过程，人的能力扩展可能伴随能力退化，形成“增强—退化”的结构性悖论。已有研究表明，技术带来的便利可能与能力维护形成张力，这种张力构成放大边界的核心问题之一。

能力边界之上还存在价值边界。放大不仅涉及效率或能力，也涉及人的价值地位是否被承认。若技术实践过度把人为数据原料或计算节点，放大便失去伦理正当性。价值边界要求在制度层面保留人的尊严地位，防止技术系统以优化名义压缩人的价值维度。

在精神层面，放大的边界体现为意义建构能力的保存。技术参与可能改变人的理解方式与判断结构，当技术提供的结果替代了自我理解过程时，个体的意义建构能力可能被削弱。例如，当算法系统过度依赖情感价值的优化来驱动用户行为时，个体的意义建构能力被算法的优化逻辑所覆盖，导致了主体性的深度消解。[①] 因此，把人放大的规范边界可以概括为三条相互支撑的限制：人的主体性优先、能力不退化、价值不降格。上述边界不是对增强的否定，而是对增强方向的规范限定，使放大能够成为伦理上可辩护的目标，而非技术扩张的副产品。

① Vinueza-Martínez C N, Ponce-Rivera O S, Díaz-Vásquez S M, et al. El valor de la dignidad humana frente a la tecnociencia J. Revista Científica Zambos, 2025, 4(3): 55-66.

二、理论框架：核心理论的系统阐释

前文已经确立了双重规范目标，即维持可被证明的信任条件，并使增强指向人的能力扩展而非主体性削弱。接下来的问题不再是概念命名，而是机制追问：为何在人工智能持续增强的过程中，这一双重目标会同时变得紧迫。若这一机制不能被解释，规范主张就容易停留在价值宣示。由此，本章采用递进式论证：先说明风险何以被低估，再说明风险如何生成，最后说明风险如何在制度结构中被放大。

2.1 AI连续论与断裂

人工智能发展的主流叙事长期由连续论主导。该叙事将技术进步描述为稳定的能力累积过程，预设从专用智能到通用智能再到超级智能属于同一轨道上的程度差异。连续论能够解释性能提升，却难以解释伦理紧张为何在近年明显加剧。问题在于，若将一切变化理解为线性增强，就会把关系结构变化误读为参数变化，从而低估主体性风险。[①]

① 杨庆峰 . 人工智能连续论反思与存在论探析 J. 中国社会科学 ,2025,(11):149-164+207

这种低估首先发生在技术层面。系统从执行型工具转向具备目标推断、策略选择和环境适应能力的智能体时，变化并非单纯提速，而是行为逻辑的改变。工具主要响应输入，智能体则可能在任务过程中重构路径并生成子目标。若仍以工具框架衡量智能体行为，责任判断和权限设计便会出现系统性偏差。

低估随后扩展到哲学层面。工具论预设人始终稳定居于主导位置，技术仅承担手段功能。但当系统可推断人的偏好、影响人的判断并反向塑造人的决策习惯时，主客关系已由单向支配转为双向塑造。相关研究指出，数字化条件下人的思维方式正被持续重塑，传统主体性分析框架因此出现解释不足。[①] 在此基础上，变化进一步触及存在论层面。人工智能介入不再局限于任务辅助，而开始进入理解方式、决断过程和意义建构。此时，技术问题不再只是系统正确率与可控性的技术评估，而是人机关系正当性的结构问题。所谓断裂，正是指某些阈值被跨越后，性质差异无法再被量变叙事吸收。[②] 由此可见，连续论揭示了发展趋势，却遮蔽了关系重构；断裂视角揭示了风险升级，却尚未回答风险在实践中如何形成。若要解释增强何以可能转化为主体性压力，还需要进入技术介入认知过程的内部机制，这正是下一节的任务。

2.2 技启的双重结构：许诺与代价

在风险升级已被识别的前提下，关键问题转为风险生成机制。技启理论提供的核心洞见在于：技术介入不是外在附加，而是进入认知形成与决断生成过程本身。只要介入进入这一层，增强就不再是单向收益过程，而表现为许诺与代价并存的双重结构。

技启不是对既有启示范式的简单补充，而是一种具有独特结构的新型认知事件。在人类思想史上，曾形成过两种影响深远的启示范式：其一是天启叙事，指将认知与行动的终极依据诉诸超越性神圣力量的指引，其权威性建立在人类理性无法直接检验的超验来源之上；其二是启蒙范式，指凭借人类自身理性突破传统权威束缚，以普遍化的理性原则重建认识和行动的基础。而技启则是指技术在特定情境中介入人的认知与决断过程，由此塑造新的理解方式和存在状态的认知范式。[③] 技启首先呈现出其许诺的一面。在技启范式下，面对人类凭借现有知识与制度无法自行化解的存在论难题，智能系统能够提供一种超出常规推理的解决路径。这种许诺不是一般性的工具效能提升，而是对人类认知瓶颈的结构性突破：系统能够整合多源信息、模拟多重情境、推断长链因果，从而在人的判断能力受限之处提供介入。从这一角度看，技启与天启、启蒙构成互补关系：天启依赖超验权威，启蒙依赖理性自律，技启则依赖技术的情境介入，三者共同拓展人类可动用的认知资源。许诺的意义正在于此——技术提供了一种前所未有的能力增量，并因此成为人类增强的重要来源。

然而，技启同时具有不可忽视的另一面。英国作家雅各布斯在小说《猴爪》中刻画了一种独特的代价结构：怀特一家许愿获得金钱，随即以儿子死亡为代价，补偿随后到来，许愿与代价在时序上倒置。这一结构在字面层次呈现为代价与愿望的共存，与传统哲学关于良药与毒药并存的二元分析相近；然而其深层结构揭示了一种更为隐蔽的模式：代价先于愿望显现，而且代价的内容与愿望并不对称，无法事先预见。

技启代价至少体现为三种不对等。其一，时间不对等，收益先显、代价后显。其二，内容不对等，代价触及主体性深层结构而非单一性能损失。其三，可见性不对等，设计方、运营方与使用方在代价识别能力上存在结构性差距。这三重不对等决定了仅靠结果评估不足以回应风险，需要建立过程性识别与制度化纠偏机制。

因此，技启理论完成的不是悲观结论，而是机制澄清：增强可以成立，但若缺少代价识别与责任安排，增强可能在实践中反转为降级。至此，风险生成机制已经明确；下一步需要解释的是，个体层面的生成机制如何在社会运行中被持续扩张，这将转入对齐、联盟与巨机器风险的分析。

2.3 价值对齐、人机联盟与巨机器风险

当代价结构从个体实践进入平台化与组织化运行后，问题重心转向制度层。此时需要回答两个连续问题：第一，何种关系形态能够约束技术目标不偏离人的目标；第二，何种权力结构会使偏离被持续扩张。前者对应对齐与联盟，后者对应巨机器风险。

价值对齐是当前人工智能治理讨论中使用最为广泛的规范性概念之一。其基本含义是：确保人工智能系统在目标、行为和结果上与人类的价值观保持一致。然而，若将对齐理解为一次性的技术设定——在训练阶段将人类价值观编码进系统，此后便可自动运行——则会遮蔽其真正的规范复杂性。相关研究指出，价值观的不确定性恰恰不应成为放弃对齐的理由，而应成为深化对齐理解的起点。[①] 对齐的真正目的，不是为人工智能找到一个终极的、静态的价值答案，而是构建一种能够理解、参与并适应人类动态寻求共识过程的机制。这意味着对齐在本质上是持续的规范协商，而非可以完结的工程任务。

① Rees T. Non-human words: On GPT-3 as a philosophical laboratory J. Daedalus, 2022, 151(2): 168-182.

② 人工智能时代与人类未来 M. 亨利·基辛格 ; 埃里克·施密特 ; 丹尼尔·胡滕洛赫尔 . 中信出版社 .2023

③ 杨庆峰 . 超级智能与存在论难题的破解 J. 阅江学刊 ,2026,18(01):99-103+173

① 闫宏秀 . 超级智能的价值对齐困惑 J. 人民论坛·学术前沿 ,2025,(23):74-83.

在此基础上，部分研究者主张以人机联盟的视角补充对齐概念。联盟的概念强调：人类在生物性层面存在诸多局限，需要伙伴而非单纯工具；人工智能不应被定位为人类意志的单向执行者，而应在共同目标下形成协作关系。联盟叙事与对齐叙事的根本差异在于主体性预设：对齐预设人类一方拥有稳定、可被提取的价值观，对齐方向是单向的；联盟则承认双方都处于动态变化之中，需要在持续互动中形成共识。从伦理论证角度看，联盟叙事更充分地回应了技启所揭示的双向塑造效应——技术不仅被人类使用，同时也在塑造人类的认知方式与判断结构，因此关系中的双方都需要在规范框架中被明确定位。

然而，无论是对齐还是联盟，若缺乏对权力结构的批判性分析，都可能在乐观叙事中遮蔽更深层的系统性风险。芒福德的巨机器理论在此提供了重要的分析视角。巨机器的核心特征是两点：一是高度集权、单向发令的指令中心；二是大量高度标准化、可替换、丧失自主性的执行单元。历史上，金字塔的建造依托了这种由人构成的刚性社会组织形态；芒福德认为，20 世纪中期的核武器是巨机器的近代原型。[①] 在当下，超级智能有可能成为巨机器的终极形态：算法系统作为前所未有的指令中心，深度学习的黑箱特性使其更难问责，而人类在双重意义上面临零件化风险——既作为数据原料的提供者，又作为算法模型的塑造对象。

在芒福德 5P 框架的基础上，相关研究提出以 5C 概念描述超级智能语境下巨机器的动力结构。控制（Control）维度指权力形态从显性命令向精准预判与自动化控制转变，呈现弥散性、微观权力特征，渗透社会结构的每一层次。[②] 计算（Computation）维度指计算成为核心权力基础与生产方式，多元社会目标被压缩为效率优化的单一逻辑。资本（Capital）维度指，在缺乏有效规范的极端情形下，资本与计算的深度绑定可能形成以算力为核心的垄断壁垒，用户数据成为这一体系中的关键生产资料，而技术决策权与控制权的集中将成为资本扩张的核心指向。共识（Consensus）维度指，在缺乏系统性规范设计的情形下，公共舆论的形成过程可能被算法过度塑造，在原本基于交往理性的信息互动中，用户对信息环境的主动塑造能力可能被削弱。文明（Civilization）维度则提出了摆在人类面前的根本性选择问题：人类文明将走向物质丰富但精神生活被算法深度编排的 “巨机器文明”，还是真正迈向以人为本的人机共生文明。

5C 框架的意义在于，它将人工智能伦理问题从技术层面的风险控制与伦理层面的规范设计，推进到权力结构批判的层次。在控制与计算维度，巨机器风险要求伦理框架必须包含对技术系统透明性与可问责性的制度性要求；在资本维度，垄断结构的存在意味着单纯的伦理倡议难以自动转化为分配公平的实践效果；在共识维度，算法塑造公众认知的机制使得对齐所预设的公共共识基础本身成为需要保护的对象；在文明维度，巨机器风险最终指向一个规范性问题：人的尊严、判断力与精神自主性，是否能够在高度自动化的计算秩序中得到系统性保障。

① 美 刘易斯・芒福德 机器的神话（下）：权力五边形 M. 宋俊岭，译 上海：上海三联书店，2017.

② 顾超 , 赵天宇 . 从巨机器到超级智能 :AI 政治的新展望 J. 阅江学刊 ,2026,18(02):77-83+176

三、面向“放心—放大”的规范主张

人工智能的深度介入已经引发人机关系的结构性变化：连续论遮蔽了存在论断裂，技启的代价结构具有渐进性与不对称性，巨机器风险指向权力集中对主体性的系统性侵蚀。在这一背景下，伦理标准倡议的首要任务是回答：何种制度条件能够在新的人机关系结构中保障人的主体性。

3.1 主体性保障的制度条件

主体性保障的核心要求并非抽象的人类中心原则，而是实践层面可操作、可检验的规范条件，其完整落地需同时满足可理解性、可介入性、可追溯性三个相互依存、层层递进的核心维度，共同构成人机关系中主体性存续的制度基础。

可理解性构成主体性保障的认知前提。若智能系统的判断依据对相关主体完全不透明，主体便无法评估系统输出是否符合自身目标与价值，更无法识别系统介入触发的技启意义上的代价结构。可理解性不要求深度学习模型内部完全透明，而是要求在关键决策节点，系统能以相关主体可把握的方式呈现判断依据与置信边界，这直接关系到不同群体间技启代价的分配公平性——理解能力的不平等会直接转化为代价识别能力的不平等[①]。

在此基础上，可介入性构成主体性保障的行动保障，核心要求是在关键节点，人拥有提出异议、请求复核或中止流程的权利与能力。这一要求不能仅依赖技术设计的善意，而需被明确为不可被效率逻辑覆盖的规范性要求，在医疗诊断辅助、司法量刑建议等涉及个人重大利益的场景中，应被视为最低制度条件[②]。

可追溯性为主体性保障提供完整的制度闭环，其核心并非单纯的事后归咎，而是要求在系统设计、部署授权与使用规范的全环节，预先建立责任识别机制、明确权责分配与纠错路径。AI 系统多主体参与的技术特征极易导致责任归属模糊，甚至被结构性利用以规避问责，可追溯性正是为了避免损害发生后陷入无主体可问的困境[③]。

三个维度呈严格的递进关系：可理解性是介入与追责的基础前提，可介入性是主体性不被虚化的过程核心，可追溯性是前两项要求不被消解的制度兜底，三者缺一不可，共同构成完整的主体性保障结构。

3.2 规范边界的三层标准

主体性保障划定了人机关系的最低伦理条件，但仅有底线要求尚不足以构成完整的伦理规范，还需进一步明确技术介入的正向目标，在能力、价值与精神三个层次上，划定增强与替代、异化之间的规范边界。首先需要澄清的是，规范意义上的增强绝非日常语境中单纯的效能提升，若技术介入在提升短期效率的同时削弱了人的理解与判断能力，便会形成隐蔽的技启代价。因此，真正的增强不能以效能为唯一标准，必须同时考察人的能力是否实质扩展、价值是否得到维护、精神自主性是否得到保全[④]。

能力层次的增强标准，核心是认知、协作与问题求解能力的实质扩展，而非单纯的任务外包，二者的区分关键在于：技术介入结束后，人是否保有或发展了独立处理同类问题的能力，而非对系统形成不可逆的依赖。相关实证研究已证实，人工

① Miller T. Explanation in artificial intelligence: Insights from the social sciences J. Artificial intelligence, 2019, 267: 1-38.

② Floridi L, Cowls J, Beltrametti M, et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations J. Minds and machines, 2018, 28(4): 689-707.

③ Alfrink K, Keller I, Kortuem G, et al. Contestable AI by design: Towards a framework J. Minds and Machines, 2023, 33(4): 613-639.

④ 唐跃洺 . 机体哲学视域下新兴人类增强技术的伦理风险与治理路径 J. 伦理学研究 ,2026,(01):95-101.

智能辅助虽能短期提升效能，但持续过度使用会显著削弱个体的独立认知与创造力，复杂任务中降幅更为突出。[①] 价值层次的增强标准，核心是维护人作为目的的主体性地位，而非使其在功能上沦为数据提供者与算法优化的对象。当前部分大语言模型对齐过程中出现的“阿谀奉承”现象为例，系统为迎合用户偏好会系统性迎合错误信念，表面满足了用户需求，实则牺牲了用户的实质认知利益，这就要求系统的优化目标必须锚定人的实质利益，而非表面的代理指标。[②] 精神层次的增强标准，是整个规范体系的最终检验，核心是保证人在高度自动化的技术秩序中，始终保有判断力、反思能力与意义建构能力。这一层次最易被效能导向的评估框架忽略，却直接关乎技术介入是否会导致人的存在论意义上的异化，尤其在教育、创作、伦理判断等与精神自主性直接相关的领域，自动化介入的深度必须受到严格审视[③]。

三个层次呈递进关系：能力层次是基础，价值层次是核心保障，精神层次是终极目标，三者共同构成完整的增强规范结构。在实践中，三层标准需根据场景差异调整权重，但底层逻辑始终一致：技术增强的终极目标是扩展人的存在可能性，而非以效率为名收窄人作为主体的发展空间。

3.3 让人放心与把人放大的伦理统一性

让人放心与把人放大在表述上是并列的，但在逻辑上并非等权的两个目标，而是构成一个具有内在张力的规范结构。放心规定下限，放大规定方向。缺少放心条件的放大，在技术能力急速扩张的情形下极易蜕变为单向的功能替代：人在名义上使用系统，实际上却在渐进地让渡判断权、认知参与和精神自主性，而这一过程因技启代价的渐进性与隐蔽性而难以被及时察觉。缺少放大目标的放心，则容易将伦理治理收缩为纯粹的风险管控逻辑：系统运行稳定、数据不被滥用、流程合规无误，但人的能力是否得到实质扩展、人的价值是否在技术秩序中得到维护，始终不在追问之列。两种偏向都无法真正回应当前人工智能发展所带来的深层伦理问题。

两个目标之间的张力恰恰是其统一性的来源。放心为放大划定边界：增强不得以牺牲主体性为代价，技术介入必须保持在可理解、可介入、可追溯的规范框架之内。放大为放心赋予方向：保障主体性的最终目的不是维持现状，而是在新的人机关系结构中开辟更丰富的人类存在可能性。若将两者分离，伦理讨论将陷入一种内在矛盾：单独强调放心，容易以保守性的风险规避代替对增强目标的积极论证；单独强调放大，容易以乐观的增强叙事遮蔽代价结构与权力风险。只有将两者作为统一的伦理结构理解，才能在技术快速演进的条件下保持规范论证的完整性。

① Rao S. The Impact of Artificial Intelligence Tools on Human Cognitive Abilities: A Comprehensive Review J. INNOVAPATH, 2025, 1(10): 7-7.

② Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models J. arXiv preprint arXiv:2310.13548, 2023.

③ Nyholm S. Artificial intelligence and human enhancement: Can AI technologies make us more (artificially) intelligent? J. Cambridge Quarterly of Healthcare Ethics, 2024, 33(1): 76-88.

四、结语

人工智能的全域渗透与向人工通用智能（AGI）的加速演进，既带来了重塑科技范式、释放经济增长动能的巨大发展红利，也同步触发了人机关系结构性失衡、人类主体性弱化、算法权力异化等系统性伦理挑战，传统以故障防控、合规审查为核心的防御性治理已难以回应当下的深层命题。本文跳出技术乐观主义与风险悲观主义的二元对立叙事，以“让人放心”与“把人放大”为双核心锚点构建了适配智能社会的 AI 伦理完整框架，通过打破技术线性进化的连续论迷思，揭示了 AI 从工具形态向智能体形态、从知识辅助向存在介入的本质断裂；借助技启理论澄清了 AI 认知介入背后“许诺与代价并存”的双重结构，破解了技术增强背后隐蔽的能力退化与主体性消解悖论；通过巨机器风险的批判性分析，将 AI 伦理问题从表层技术管控推进到权力结构与文明走向的深层审视。

最终以可理解性、可介入性、可追溯性三个递进维度构建了“让人放心”的主体性保障制度条件，划定了人机关系的最低伦理底线；以能力、价值、精神三层标准明确了“把人放大”的技术增强规范边界，锚定了 AI 发展服务于人的全面发展、助力中国式现代化建设的终极方向。二者并非相互割裂的并列目标，而是形成了具有内在张力的统一伦理结构——既非限制技术创新，也非放任技术扩张，而是为 AI 发展确立了稳固的人本坐标。这套框架既是对人类命运共同体理念在智能时代的具体践行，也是“智能向善”在技术伦理学层面的学理深化，既回应了当下全球 AI 应用的共性伦理困境，也为 AGI 时代提前筑牢了人本伦理阵地，最终指向人机共生、相互成就的文明新形态，让人工智能的发展始终以成就人、拓展人的发展可能性为终极归宿。

作者

杨庆峰　复旦大学科技伦理与人类未来研究院

李开阳　复旦大学哲学学院

2026 年 5 月 8 日

* 其中“让人放心，把人放大”原概念取自腾讯研究院文章《AI 要让人放心，把人放大》。

* 学术免责声明：本文学术观点仅代表作者个人，不代表特邀合作单位腾讯研究院的立场。