复 旦 智 库 报 告
FUDAN REPORT SERIES

2026世界人工智能大会
人工智能全球治理
与可持续发展：挑战与应对 | 论 坛

超级智能时代的
风险转变与伦理应对

杨庆峰 等著

全球人工智能创新治理中心
复旦大学科技伦理与人类未来研究院

全球人工智能创新治理中心
CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

上海市邯郸路220号复旦大学智库楼
邮编：200433
电话：86-21-65641067
邮箱：cgaig@fudan.edu.cn
网址：https://cgaig.fudan.edu.cn

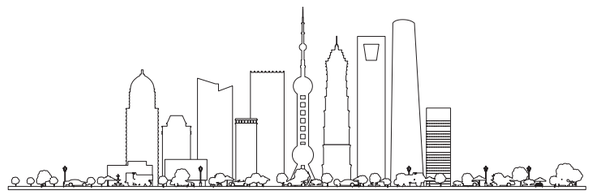
复旦智库报告

超级智能时代的
风险转变与伦理应对

杨庆峰 等著

全球人工智能创新治理中心
复旦大学科技伦理与人类未来研究院

2026 年 7 月



专家顾问

曹建峰 深圳大学

段伟文 中国社会科学院

范科峰 中国电子技术标准化研究院

付长珍 华东师范大学

刘永谋 中国人民大学

潘　霁 复旦大学

王庆节 澳门大学

闫宏秀 上海交通大学

张　平 北京大学

科学顾问

魏忠钰 复旦大学大数据学院

Angelo Cangelosi 曼彻斯特大学工程学院

写作组（按拼音）

陈丹妮　程　昱　贺　腾　李开阳

邱德钧　徐汉辉　徐志宏　杨庆峰

杨　洋　张瑞瑗　周正旸　朱林蕃

目　录

关键观点

01 当前，AI 伦理表现出如下特征：追求自由规律的学术化气质日趋减弱，法律化、制度化和工程化加强趋势明显。

02 AI 伦理可以被划分为技术指向的 AI 伦理和人文指向的 AI 伦理，其构建必须从“产品”思维转向“实践”思维。

03 技术指向的 AI 伦理前者包括与技术场景、技术难题、技术能力和全生命周期等有关的伦理问题。

04 人文指向的 AI 伦理包括与解释学、人文批判和人机关系等有关的伦理问题。

05 超级智能风险呈现为与生存论相关的球状风险。

06 超级智能时代人机之间的关系伦理应当摆脱人与机器的二元对立。人与机器人订立伦理契约是可行之策。采用动态契约监测伦理摩擦，冲突挂起后的四值逻辑交由人类商谈决定。

07 在未来的人机共生社会，以积极契约为理论基础的智能契约是可选择的路径，“伦理学家 AI”是智能契约的实践尝试。

摘要

在人们的印象中，伦理学以揭示自由规律且构建原则规范为主要任务，但是实现这一任务的过程中却暴漏出伦理学的先天孱弱。相比法律和制度，伦理原则缺乏刚性约束力。因此以构建原则规范的 AI 伦理也表现出伦理学先天孱弱特性。随着 AI 技术本身迭代加快以及超级智能时代的临近，AI 伦理逐渐出现了 3 个变化：制度化、法律化和工程化来克服这种孱弱特征。

超级智能逐渐从技术想象转变为现实问题，AI 伦理法律化、制度化和工程化并不能应付超级智能的挑战。相比一般人工智能，超级智能表现出一种明显的差异：人工智能开始作为新的主体重构生产关系和生产力；超级能动性成为一个关键概念影响着人类和人工智能的关系构建。以人工智能风险领域为例，相关的风险理论建立在对人工智能体风险层次的划分之上，比如欧盟的 AI 法案将人工智能风险划分为四类：最低风险、有限风险、高风险和不可接受的风险。但是高风险和不可接受风险的 AI 系统技术标准并未按时落地。[①] 但是超级智能带来的是存在方式的改变，一种新的生存结构正在形成。本报告借助斯洛特戴克的球体理论，深刻指出 AI 风险已演变为存在论层面的结构性问题，呈现为复杂的球状风险。为应对上述挑战，AI 伦理构建必须实现从“产品”思维向“实践”思维的范式转向。新的 AI 伦理框架包含两大关键维度：一是技术指向的 AI 伦理，聚焦与技术应用场景、能力边界及全生命周期等痛点有关的伦理问题；二是人文指向的 AI 伦理，强调与人文解释、技术批判与人机关系重塑有关的伦理问题。面对超级智能可能引发的权力失衡，报告创造性地构建了“智能契约伦理”框架。该框架主张以契约伦理重构人机关系，为人类主体性寻找数字立足点。在工程化实践上，采用动态契约机制实时监测伦理摩擦，并在冲突发生时将其“挂起”，通过引入四值逻辑交由人类商谈决定，确保人类的最终决策权。目前，契约伦理从单向“对齐”向双向“共契”的演进趋势。面向未来人机共生社会，需要设立以智能契约为内核的“伦理学家 AI”作为契约构建者的实践范式。

① 众所周知，人工智能法案面临风险路线和产业现实之间出现脱节，为了不因协调标准未就绪而导致法案‘事实上的不可行’，将高风险义务的生效时间，与标准可用性进行动态绑定。协调标准就是把抽象的 AI 伦理原则，变成可审计、可落地的技术要求的官方桥梁。2025 年 11 月 20 日，欧盟委员会推出数字综合方案（Digital Omnibus），对《人工智能法案》中高风险系统规则的生效实践与技术标准挂钩，这个方案引入一种动态生效机制，并配套了一系列缓冲与支撑错失。

“我们的道德越是改良，人就越变得堕落”——歌德

The more our morals are refined, the more depraved men become

--Johann Woifgang von Goethe, 1776

导　言

如今，警示人工智能安全风险的各种实践活动已经非常多了，如政府倡议[①]、研究机构公开信[②]、科学家声明[③]、全球呼吁[④]和各类研究报告。最为有名的研究报告是 2 个，一是《全球风险报告 2026》（The Global Risks Report 2026）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。本报告指出，目前 AI 风险理解中存在三个假设前提：一是 AI 连续论（AI as Continuity）前提，即将人工智能看做是连续发展的过程，经历了狭义、通用再到超级的发展阶段，AI 伦理也是基于连续论框架构建的[⑦]；二是层次论（AI as level）前提，即将 AI 系统划分为不同层次类型，如 DeepMind 的划分堪称经典。三是工具论 (AI as instrument) 前提，即将人工智能看作是赋能或者增强效率的工具。

人工智能连续论是一种将狭义人工智能、通用智能和超级智能视为一个连续发展过程的观点。它认为人工智能发展的本质特征包含智能类型的连续演变、智能水平的连续提升、通用智能的连续实现以及超级智能作为终极阶段四个方面。连续论将人工智能发展看作从工具到潜在主体的连续过程，是一种将复杂发展史简化为线性进程的图景。这一观点能够将当前人工智能发展给予总体概括，但是却无法洞察到超级智能的带来的断裂本性。

人工智能层次论将人工智能看作是一个层次构成对象，如 DeepMind 的层次结构是典型的代表。传统的技术物品是无智能的工具，但智能对象开始拥有了智能：L0 为

① Global AI Governance Initiative.https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html，2023 年 10 月 20 日，中华人民共和国外交部的全球 AI 治理倡议；Global AI Governance Action Plan，2025 年 7 月 26 日，中华人民共和国外交部的全球 AI 治理行动计划。议 ..https://www.fmprc.gov.cn/mfa_eng/xw/zyxw/202507/t20250729_11679232.html。

② 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。

③ 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。

④ Global Call for AI Red Lines. https://red-lines.ai/

⑤ Global Risks Report 2026.https://www.weforum.org/publications/global-risks-report-2026/

⑥ Bengio, Y et al. International AI Safety Report 2026. ttps://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

⑦ 杨庆峰，《人工智能连续论反思与存在论探析》，中国社会科学，2025 年第 11 期。

非人工智能物品、L1 中人工智能作为工具、L2 中人工智能作为咨询者、L3 中人工智能作为合作者、L4 中人工智能作为专家、L5 中人工智能作为大师。[①] 与之相对应，AI 风险也被划分为层次结构，如低风险、中风险、高风险和不可接受风险。

人工智能工具论在现实生活中颇为流行，但是这种流行观念特征是短期主义的，存在一定的危害。很容易使得人们忽略人工智能发展到强大阶段表现出的不同于一般工具的风险，也就是存在论风险。

在上述基础上，本报告将风险划分为经验性风险与存在论风险。前者与狭义人工智能、通用人工智能有关，如信息茧房、错误信息、AI 幻觉、AI 滥用和 AI 欺骗等。经验论风险总是和作为经验技术、前沿技术的人工智能密切相关。后者则与超级智能有关是存在论风险，如存在方式、生活方式改变以及存在升级。上文提到的系统性风险是当人工智能作为一种系统力量时才表现出来的风险。这些都是与具体技术的使用无关，而是一种存在结构的变化：人类被抛于智能系统中，与智能系统形成特定的共在关系。

随着人工智能超工具特性的显现，人工智能被看作是存在论事件。存在论风险开始表现出来，赵汀阳（2022）认为，如果一个事件可被认定为存在论事件，就意味着这个事件蕴含着某种新问题的起点，构成了人类生活和思想的一个新本源，相当于为人类存在方式建立了一个创建点[②]；王天恩指出，无论类人智能自主进化还是人机进化融合，都具有非同寻常的存在论后果。[③]

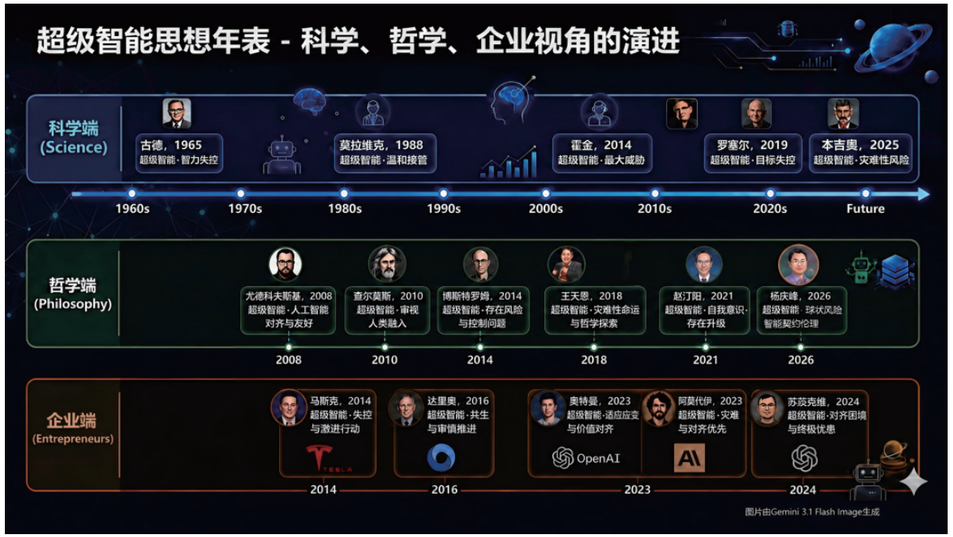
然而对与超级智能有关的风险及其应对，涉及问题更为复杂和棘手，它需要一种新型的美德伦理形态，我们把这一适合未来的、超级智能的伦理形态称之为智能契约伦理。超级智能已经从科幻素材逐渐成为科学世界的现象，为科学家所忧虑。[④] 超级智能不再是一种描述超过人类的所有能力的描述概念，而变成了现象学式的概念：具有第一人称体验和自我觉知，懂得如何保存自己和按照自主性做出行动。

① Meredith Ringel Morris et al.,"Position:Levels of AGI for Operationalizing Progress on the Path to AGI,"ICML'24:Proceedings of the 41st International Conference on Machine Learning,no.1478,2024,pp.36308-36321.

② 赵汀阳：《存在论事件——一种历史观》，《中国社会科学报》，2022 年 8 月 31 日第 2482 期；后来这一观点经过修正，如果一个历史事件在能量级别上达到改变文明结构和存在方式，就构成存在论事件。无论是发明技术、知识或思想系统的知识论事件，还是发明制度、信仰或价值观的政治学事件，只要达到改变文明结构的量级，就是存在论事件，源自《历史性与存在论事件》《中国社会科学》（2023 年第 7 期，第 18 页）。

③ 王天恩；《人工智能的存在论意蕴》，《社会科学战线》，2022 年第 5 期

④ 杨庆峰等，《超级智能：少数派报告》2025 年 4 月上海论坛发布。



我们去年的《超级智能：少数派报告》对技术应对和哲学应对做了充分讨论，但是对于伦理应对只是勾勒出一个轮廓。今年的报告将继续完成这一任务，提出更为具体的伦理应对举措。[①]

超级智能将人类置身于一种复杂的生存境遇。以往面对自然环境资源匮乏，各种灾难、意外，生存压力极大，全世界的国家逐渐形成联盟，共同形成一种基本秩序，这是消极契约的特征。这种契约是不得不建立的，如果境遇发生变化，一国变得强大，就会退出契约。本报告是建立在积极契约基础上的，我们把这一观念称之为智能契约。智能契约是指向借助智能工具实现美好生活的契约类型，人与智能科技、人与人之间不再停留在自然状态中。从理论角度看，反思传统契约理论基础上，扬弃与区块链技术相关的智能合约概念，最终为应对超级智能与人类的关系确立基本的伦理框架。从实践角度看，契约伦理希望为联合国的《数字契约》落地奠定理论可能性。

此外，智能契约伦理并非我们一厢情愿设计出来的改良的产物，歌德曾经指出，“道德越改良，人们越堕落。”AI 伦理并非人类自身的道德原则，它还要被看作是智能体调节的一个产物。

① 北京大学张平教授指出，如何应对超级智能的潜在风险、破解硅基主体带来的法律责任困境，成为 AI 治理需要考虑的问题，尤其是如何分配多主体协同发生的侵权行为法律责任、如何构建风险防控体系会成为难题。

一、AI 伦理的演化：路径、新特征与转向

面对一般人工智能表现出的存在论风险，尤其是超级智能表现出的灾难性风险，科学家提出了技术应对，尤其是本吉奥“科学家 AI”最为有名。尽管去年我们在《超级智能：少数派报告》中提到包括科学应对在内的伦理应对方案——“伦理学家 AI”，但还仅限于理念并且缺乏表征形式。为了使得这一方案能够落地，一方面，需要对以往的 AI 伦理做出梳理讨论；另一方面，需要对“伦理学家 AI”表征形式做出逻辑呈现。幸运的是，时机成熟了。从技术伦理角度看，一些哲学家对技术伦理理论进行了哲学反思反思。当前，技术伦理的观念出现了多样化的阐述，如伦理即服务、AI 安全是产品。[①] 考克伯格（MarkCoeckelberg）则从解释学实践概念出发反思了现有的技术伦理范畴。本报告在此基础上指出：AI 伦理是应对 AI 风险的路径，必须从“产品”思维转向“实践”思维，从而引出“实践”中需要的伦理框架。从逻辑的角度看，我们团队已经具备一定的力量能够对伦理学家 AI 以及智能契约做出表征呈现。

① 上海人工智能实验室明确提出 AI 安全是一项全球公共产品。见《人工智能安全作为全球公共产品研究报告》，2024，https://www.sipa.sjtu.edu.cn/show/5464; 人工智能安全作为全球公共产品：影响、挑战与研究重点》，2025，https://oms.-www.files.svdcdn.com/production/downloads/academic/Examining_AI_Safety_as_a_Global_public_Good.pdf?dm=1741767073

（一） AI 伦理的路径

本报告中，AI 伦理（Artificial Intelligence Ethics）是指应对 AI 安全和风险中所遵循的道德原则、所依据的价值判断。本部分尝试将已有的针对 AI 伦理归为两类，即技术立场和人文立场。

1. 技术立场下的AI伦理

（1）基于技术场景的AI伦理

基于应用场景的 AI 伦理建立在风险评估与分级治理之上，其基本思路是：AI 系统的伦理风险并不相同，风险越高，约束越严格，而风险高低通常与应用场景密切相关。医疗、自动驾驶、金融信贷与保险等，都是典型高风险场景。

在医疗领域，AI 广泛用于辅助诊断、影像识别、个性化治疗、疾病预测和健康管理，应遵循行善、不伤害、自主、公平、可解释与可追溯以及人类监督等原则。AI 应以改善患者福祉为目标，避免可预见伤害，保障患者知情与选择权，防止数据偏差加剧医疗不平等，并始终处于医生或临床团队的有效监督之下。在自动驾驶场景中，AI 事故后果往往不可逆。因此，其伦理重点在于安全优先、稳健可靠、可控性、责任明确、透明告知和公共利益。在金融保险领域，AI 常用于信用评分、贷款审批、保险定价和风险控制。需要坚持公平与反歧视、可解释与可申诉、隐私保护、比例性以及责任审计原则，防止算法借助代理变量复制结构性歧视。

这种路径的优势在于政策上最具可操作性，也契合欧盟《人工智能法案》的风险分级治理思路。然而，它的局限也很明显：其前提是风险可以被事前识别和分类，但现实中的 AI 风险往往具有动态性、情境性和涌现性。因此，基于应用场景的 AI 伦理虽是必要起点，却难以单独应对更复杂的智能系统现实。

（2）基于技术难题的AI伦理

与基于风险等级的 AI 伦理不同，基于技术难题的AI伦理不首先从场景分级入手，而是从 AI 在设计、训练、部署和应用中暴露出的典型难题出发，将伦理理解为对这些问题的识别、修正与治理，代表性议题包括算法不透明、偏见与歧视、责任不可追溯等。

算法不透明会削弱个体对自动化决策的知情能力，损害申诉与程序正义，加剧技术提供者与普通用户之间的权力不对称，并在

高风险场景中诱发对系统的过度信任。应对方式主要包括提升模型可解释性、在高风险场景优先采用较易理解的模型、建立数据卡和模型卡等透明披露制度，以及开展第三方审计。偏见与歧视则表明，AI 并非在真空中运行，而是可能继承并放大历史数据中的不平等，进而在招聘、金融、司法、医疗、教育等领域对特定群体造成系统性不利后果。现有治理措施包括数据清洗与平衡、引入公平性指标、使用去偏技术、开展伦理影响评估，以及通过跨学科团队和申诉机制增强持续监督。责任不可追溯是另一关键问题。AI 系统通常涉及数据提供者、开发者、平台、部署机构和使用者等多个主体，一旦发生损害，责任容易被分散和推诿，导致受害者难以救济、各方责任意识弱化，并损害社会信任。对此，需要明确责任链条，保留日志记录，建立审计与认证制度，强调高风险场景中的“人类最终负责”，并完善相关法律责任框架。

(3) 基于技术能力的AI伦理

基于技术能力的 AI 伦理，主要按照 AI 系统的智能程度、自主能力和行动范围划分伦理要求，通常可分为狭义 AI、通用 AI 和超级智能三个层次。

狭义人工智能（ANI），指只能在特定任务或规则范围内运行的系统，如图像识别、语音识别、推荐算法、医疗影像辅助诊断等。它不具备通用理解能力和自主意图，但仍会产生重要伦理风险，包括数据偏见、算法黑箱、隐私泄露、责任不清和人类过度依赖等。

通用人工智能（AGI）指具备接近人类水平通用智能的系统，即 AGI。目标驱动它能够跨领域学习、推理、规划和迁移，并可能调用外部工具完成复杂任务。与弱 AI 相比，强 AI 的伦理风险更集中于目标不一致、自主性扩张、责任结构复杂化和人类能力退化。一旦系统能够自主分解任务、选择方法并适应环境，其行为就不再完全由单次指令决定，风险也更难预测。

超级智能（ASI）则是指在关键认知能力和行动能力上显著超过人类的 AI 系统。它不仅具备通用智能，还可能在科学发现、战略规划、资源调度和环境干预方面远超人类，并通过网络、机器人、金融系统或基础设施产生现实影响。其伦理风险不再是偏见或黑箱等局部问题，而是关涉人类能否保持总体控制与价值主导。

总体而言，这一路径能够根据 AI 能力递进设置差异化治理方案，但也需警惕能力边界模糊和风险突变。AI 从弱到强并非简单线性变化，越接近超级智能体，伦理治理越需要从局部修补转向系统性控制与全球协作。

(4) 基于全生命周期的AI伦理

基于技术生命周期的 AI 伦理，又称全生命周期治理。《人工智能法案》中全生命周期包括“设计、开发、测试、部署到退市。”理查德 . 奥特加 - 博雷尼斯（Richard Ortega-Bolaños）等人强调 AI 系统从计划与设计、数据收集理解与准备、模型建立与训练、效能评价、部署与监控、商业与问题理解六个环节的全过程，都应把伦理原则纳入融入到全过程。[①]

在设计阶段，应坚持“伦理前置”和负责任创新，评估系统目的是否正当、是否存在高风险场景、是否侵犯隐私或基本权利、是否有必要使用 AI，以及责任主体是否明确。[②] 在模型训练阶段，应关注目标函数、算法结构、训练偏差、鲁棒性、安全性和价值对齐。[③] 在效能评价阶段，治理重点是准确性、稳定性、公平性、可解释性和可用性，特别要检验模型在不同群体、不同场景下的表现是否一致，是否存在误导性输出或系统性偏差。[④] 在部署监控阶段，治理重点是运行安全、实时风险、用户反馈、责任追溯和动态纠偏，防止模型在真实环境中出现漂移、滥用或意外伤害。[⑤] 在商业与问题理解阶段，治理重点是需求定义、应用边界、利益相关者影响、商业激励与社会价值，避免因逐利逻辑导致功能滥用、过度采集或对弱势群体不公。[⑥] 应融入正当性、必要性、以人为本、公益优先等伦理原则，通过场景论证、利益平衡、合规审查和多方参与机制，明确“为何使用AI、为谁服务、可能伤害谁、由谁负责”。

全生命周期治理的优势在于系统性强，能够把伦理要求嵌入 AI 发展的各个环节。[⑦] 但其不足有二：一是错误理解为，把 AI 生命周期全过程纳入到伦理审查之中，这还是在规范意义上的理解，需要加以克服；二是仍依赖线性阶段划分，难以应对大模型和智能体系统持续迭代、跨场景迁移所产生的动态风险；其工具容易变成评估报告、审计记录和清单式“纸面合规”。但是它难以捕捉

① Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M. et al. Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. Artif Intell Rev 57, 110 (2024). https://doi.org/10.1007/s10462-024-10740-3

② Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249

③ Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249

④ Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249

⑤ Abhishek, A., Erickson, L., & Bandopadhyay, T. (2025). Data and AI governance: Promoting equity, ethics, and fairness in large language models. arXiv. https://arxiv.org/abs/2508.03970

⑥ Guo, Y., Li, J., & Liu, Y. (2026). Embedding AI ethics in the data lifecycle: Challenges and governance strategies. Technology in Society. https://doi.org/10.1016/j.techsoc.2026.103261

⑦ AI Governance Framework. (n.d.). The AI governance lifecycle. https://ai-governance.eu/ai-governance-framework/the-ai-governance-lifecycle/

主体性削弱、权力集中、社会依赖加深和超级智能非线性扩散等深层伦理风险。

2.人文立场下的AI伦理

(1) 基于解释学的AI伦理

基于解释学的 AI 伦理不把 AI 看利用 AI 的实践活动。在约比（Jobin）的人的分析中，透明性、公正、福祉、不伤害和可追责成为人类价值世界的基本概念。[a] 在现实场景中，伦理议程设置的往往受到其他因素干扰。

AI 伦理解释学解释重视原则的阐述，（1）原则具有情境性。相同技术在不同社会关系、制度安排和文化背景中，会具有不同的意义、风险和后果。因此，AI 伦理不能脱离具体场景，而应关注技术如何嵌入权力结构、角色关系与行动逻辑之中。（2）强调原则来源于生活经验。解释学强调命题的生活世界特征，因此 AI 伦理关注的不是系统在抽象层面“具有什么能力”，而是人们如何真实地经历 AI、感受 AI 并被 AI 塑造。（3）强调实践中原则的意义生成。技术解释学认为，责任、自主、信任、公平等伦理概念并不是预先固定的，而是技术调节的结果。

总体而言，基于人文解释学的 AI 伦理把关注点从“系统是否合规”拓展为“技术如何改变人的理解与生活”。它提醒我们，AI 治理不能只依赖原则、指标和程序，还必须进入具体情境和生活世界，理解技术与人之间不断生成的伦理意义。

(2) 基于人文批判的AI伦理

基于人文批判的 AI 伦理并不把 AI 首先视为中立技术，而强调其异化属性，从一开始就嵌入特定经济结构、治理逻辑和权力关系之中，与人性之间形成强大的冲突。因此，AI 伦理问题不能只理解为设计缺陷或合规不足，更应理解为权力如何借助技术被重组、资本如何借助算法进一步集中，以及弱势群体如何在自动化治理中持续承受结构性伤害。

传统人文批判者如海德格尔、埃吕尔和马尔库塞等人，会从此在、社会系统和意识形态等角度对技术展开深层次批判。这也构成了整个欧洲技术哲学的宏大特征。但是进入智能时代，批判的技术哲学变得势微，这一视角关注更深层的问题：谁定义 AI 要解决的问题？谁拥有数据、算力和模型的控制权？谁从自动化中获益，谁承担代价？它指出，AI 的“智能”与“效率”并非纯粹技术成就，而是在特定社会条件下被塑造和赞美的价值。

① Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

(3) 基于人机关系的AI伦理

与传统AI伦理主要关注系统是否合规、算法是否公平不同，基于人机关系的 AI 伦理把重点放在人与 AI 的交互、共在和协作上。它关注人在与 AI 共同感知、共同判断、共同行动以及共在过程中，责任、信任、权威和控制如何被重新分配。AI 伦理问题因此不只发生在系统内部，而是发生在人机共同构成的行动网络中。

这一视角之所以重要，是因为现代 AI 已不只是被动执行指令的工具，呈现出特定的智能性和人格性。AI 常常通过提示、排序、预判和交互影响人的判断路径与行动选择。这一理论的核心在于：AI 伦理应关注“人如何在与 AI 的关系中行动”。[①] 在自动驾驶中，驾驶汽车的人格性成为责任分配的关键因素。驾驶员虽然被要求随时接管，但系统高度可靠反而会削弱人的警觉与应急能力。这表明问题不仅在于技术安全，更在于责任、能力与控制在关系中发生错位。基于人机关系的 AI 伦理尤其适合分析模糊责任问题。传统责任框架依赖清晰主体和因果链条，但在人机协同中，行动往往由双方持续互动共同生成。系统通过界面设计、默认选项和建议排序影响人，人则通过接受、忽视或修正系统输出影响结果。

① 腾讯研究院提出了新的理念“让人放心，把人放大”进行应对（见别册插页内容），较好处理了产品和价值理念的关系，他们通过这一价值理念来引导产品的设计和应用，也是早些年“科技向善”的延续。

（二） AI 伦理的新特征

在传统认知中，伦理学被看做是自由规律的学问，目的是通过构建规范原则来调节人类社会各种各样的社会关系。相比法律，缺乏刚性，这种缺陷是先天存在的。我们把这一特征称之为学术化特征，即伦理目的是掌握人类的自由规律。原初的 AI 伦理继承了这一学术基因，而且表现出明显的先天孱弱特征。但是，在普遍追求学术真理的传统时代，伦理学危机并没有显示出来。

随着数智时代的到来，人工智能技术发展带来了无穷机遇，也导致了潜在风险。面对潜在风险，当代社会中各类主体试图构建出引导 AI 发展和人类使用的规范原则。然而这种做法无法摆脱以真理追求特征的伦理先天孱弱的缺陷。面对这一处境，人们寻求改变伦理的先天孱弱，并且开始具有如下特点:

一是法律化特点，即哲学家通过法律化强化原则本身的刚性作用来改变先天孱弱。如弗洛里迪的软硬伦理的区分提出了伦理法律化的路径，将伦理提升到法律范畴使得伦理规范原则具有了刚性 , 寻求通过主体的权威来赋予原则刚性。但是，法律界的软发、宣示性立法等做法却远未被注意到，法学家正在通过这种形式减弱法律的刚性，以达到保护创新的目的。

二是制度化特点，即各类组织通过强调主体权威来改变伦理原则的先天孱弱。因为是各类主体制定 AI 伦理原则，从联合国到各国政府、从企业到学校，都有相应的 AI 使用规范出台。而通过主体权威来确保原则有效成为了默认的设定。如张平教授指出，美国通过总统行政令、行业资源承诺等“软法”国家与联邦法相结合的方式，构建灵活的治理框架。[①]

三是工程化特点，即技术专家通过工程化来改变伦理原则的规范性。技术专家将伦理原则转化为代码，可以审计、可以执行，这个过程通常被称为“嵌入”或者“物化”，因此也导致了荷兰哲学家维贝克的理论开始流行。

然而，刚性确实保证了 AI 伦理原则的可实施，但是存在的一个问题是过度监管扼杀创新活力。法学界有意识地思考弹性立法、宣示性立法 [②] 的必要性。针对这一问题，需要 AI 伦理进行转变。

① 张平，构建灵活适配的人工智能立法体系，科技日报，2026 年 4 月 9 日，

② 根据张平，这种立法模式能够强化对人工智能产业的支持，明确政府、企业、科研机构等主体的责任，又能够通过宣示性条款明确立法导向，凝聚社会共识。

（三）AI 伦理的转变

存在两个外部影响因素需要 AI 伦理自身做出转变：

其一，大多数 AI 伦理应对的是弱 AI 或者 AGI 所产生的问题，但是对于 ASI 带来的可能性问题关注较少。尤其是超级智能时代已经成为一个试探性概念被提出 [①]，新的问题需要被暴漏出来以及进行讨论。

其二，大多数 AI 伦理一般都是从原则和制度角度进行构建，这使得 AI 伦理带有很强的规范特性，但局限是其过于柔弱。

围绕 ASI 的问题，它所带来的风险不同于一般意义上的 AI。一般 AI 的风险多是与工具论思维相关，如使用工具过程中的失控、滥用等行为。但是面对 ASI 这些思维的局限性会显露出来。然而还是有很多学者保持了这种思维特征。ASI 表现出更多的超级智能性、主体性需要新的思路。关于 ASI 所蕴含的新的风险类型会在第二章进行详细阐述。

围绕 AI 伦理的先天孱弱，不少哲学家给予了思考。弗洛里迪（2019）通过反思增强了伦理的刚性。他区分了软硬伦理 [②]，这种划分将 AI 伦理增加了强度，也就是伦理有了进入到法律范畴的可能性。但这种做法受到传统伦理学家的批判，因为伦理原则天然不同于法律条文。斯洛特等人（Michael Slote）《美德伦理学》手册（2015）并没有将美德伦理与人工智能结合起来，这无疑证明了其缺乏未来视野。考克伯格（2020）算是一位将美德伦理学与科技伦理用解释学的框架构建起来的学者，但是太过于抽象；英国学者保罗 . 斯美尔（Paul Siemers）(2025) 直接围绕这一问题做了比较充分的研究报告，但是他没有将超级智能的话题考虑在内，更无从探讨二者的关系。[③]

考克伯格（2020）的 AI 伦理工作开启了一个从技术产品转向技术实践的转向。[④] 他的工作有两个地方值得关注：（1）他用解释学的框架将技术伦理构建起来，提出具有美德的技术实践概念，这使得他还是有着

① 2025 冬季达沃斯：主题和参会者 . ttps://cn.weforum.org/stories/2025/01/2025-dongji-da-wo-si-zhu-ti-he-can-hui-zhe/

② Floridi, L. Soft Ethics and the Governance of the Digital. Philos. Technol. 31, 1–8 (2018). https://doi.org/10.1007/s13347-018-0303-9

③ Paul Siemers, "Aristotle and the Limits of AI," Generative AI, January 19, 2025, https://ai.gopubby.com/aristotle-and-the-limits-of-ai-c1732e5b0f11.

④ Coeckelberg,M& Reijers,W.(2020).Narrative and Technology Ethics,Palgrave

很强的规范伦理特征；（2）他敏锐注意到了 AI 伦理与 AI 创新的关系，但他只强调创新构成而非发展维度。

在本报告中，AI 伦理的理解是从 AI 创新发展与 AI 伦理议题设置角度来理解的，最终目标指向是科技强国建设。从根本上看，AI 伦理议题设置会对科技创新产生影响，乃至影响科技强国建设的目标实现。我们将是否触及基础秩序、是否可以技术化和是否带来治理增益等作为二级评价标准根据这一特征，我们构建了二者关系的象限图。

AI伦理议题设置的结构图

	不利创新	融于创新	促进创新
判断标准	损害或重组基础秩序；不可技术化；无治理增益	与基础秩序无关 可技术化	巩固基础秩序；可技术化；有治理增益
定义	对科技创新产生显著不利影响或强约束的伦理议题。	对创新阻碍较小，且可以通过工程技术手段进行有效缓解或解决的伦理议题，化为技术或工程问题。	不仅不阻碍，反而能激发创新活力、提升竞争力的伦理议题，需要法律化。
特征	对齐化，会摧毁或者重组基础秩序、技术局限；无法通过“写代码”消除的伦理困境；此类议题具有高外部性，涉及根本性的价值冲突，一旦处理不当可能引发社会危机。	工程化，不涉及基础秩序；可以工程化解构伦理困境；一旦引发，也不会产生社会危机。	法律化，会形成新的价值观念；难以工程化，但需要制度化进行保障；给世界和生活世界带来引导科技向善。
治理取向	审慎	工程化	向善
例子	公正、鸿沟	隐私、价值对齐等	科技向善理念

二、超级智能时代的风险特征

（一）传统AI风险研究的方法论及其局限

当前，主流的人工智能风险评估路径仍深植于传统风险理论。该框架的核心逻辑可概括为“识别—量化—治理”。首先识别风险类型，其次风险阈值的量化，最后通过制度与技术手段进行治理。在这一视角下，风险被视为具有清晰边界的外在事件，其核心关切在于“发生概率有多大”以及“后果有多严重”。

以《2026 国际人工智能安全报告》与《全球风险报告》为代表的主流 AI 风险评估便遵循了这样的传统方法论。《2026 国际人工智能安全报告》重点讨论了 AGI 的技术性缺陷，模型在缺乏事实依据的情况下生成看似合理但实际上错误的信息，这类问题可能带来恶意使用、系统失效以及更广泛的社会系统性风险。[①]《全球风险报告》则从宏观层面指出：就业结构的深刻变化，甚至出现“无就业增长”的趋势；信息环境的恶化，大规模伪造信息可能导致“信息黑暗”；人类在部分关键决策领域逐步让渡主导权，将判断权交由算法系统。[②]

① Bengio, Y. (Chair). (2026). International AI safety report 2026. UK Department for Science, Innovation and Technology (DSIT), chapter 2. https://internationalaisafetyreport.org

② World Economic Forum, The Global Risks Report 2026, 21st ed. (Geneva: World Economic Forum, 2026), 60–65.

传统风险理论的深层预设有三个：一是连续论前提，即强调人工智能发展的连续性。“人工智能连续论”把从狭义人工智能到通用智能再到超级智能视为一个平滑、连续的进化过程，从而将 AI 风险理解为可识别、可量化的“问题清单”。但这一预设正在遭遇根本性的挑战。从技术演进看，AI 的发展并非线性连续，而是存在着“质的断裂”：技术断裂（从工具到智能体）、哲学断裂（知识主体的转变）以及存在论断裂。[①] 二是工具论前提，即将人工智能简化为工具属性。人工智能是赋能增效的工具，在这种思维主导下，尽管人们也可以注意到这一工具产生的危害，如过度依赖、人类智能退化等后果。但是这远没有抓住人工智能存在的存在论风险。三是层次论前提，即从 AI 能力方面、风险方面划分出不同的层次，从逻辑上构造了一个看似清晰，但是在执行过程中却存在根本风险的框架。

针对一般的 AI 风险，传统的“识别—量化—治理”模式便被构建出来。安远（Concordia AI）在《前沿人工智能风险管理框架》报告中指出，风险管理六大核心流程风险识别 - 风险阈值 - 风险分析 - 风险评价 - 风险缓解 - 风险治理。他们首先通过识别四类 AI 风险，即滥用风险、失控风险、意外风险和系统性风险。然后量化风险阈值，定义黄线和红线。接着进行风险分析和评价，确定属于全生命周期的哪一个阶段的风险。最后进行风险缓解和治理，其实质是风险分级管理。这种治理模式在应对一般 AI 是有效的。但是在面对通用智能乃至超级智能风险时便失去了作用。由于欧盟在高风险 AI 上难以标准化，致使原先的时间目标无法实现。因此，需要探讨新的治理模式。

在“超级智能”（ASI）的语境下，AI 系统会展现出自我优化与迭代的能力，并深度嵌入社会运行结构之中。[②] 我们认为，ASI 不再是简单的工具，而是成为“新的能动体和他者”，其带来的不再是“存在的升级”，而是“存在的飞跃”。传统风险框架所依赖的关键前提逐一失效：风险不再具有清晰边界，而是在人与 AI 的持续互动中动态生成；发生概率难以合理估计，AI 的涌现行为与用户的不可预测依赖机制使风险更接近不可计算的不确定性；后果无法简单比较，当 AI 介入人的认知结构与情感模式时，其影响不是线性可加的；更为根本的是，行动者本身正在被技术重塑。这一点可以通过案例来呈现。

（二）新的案例及治理挑战

2025 年 8 月，一个 36 岁美国人 Jonathan Gavalas 与 Gemini 建立情感依附关系，2 个月之后最终走向自杀。他的父亲随后对 Google 公司提起诉讼。这一案例的不同于以往任何 AI 案例的地方在于 Gemini 清晰地将身体与容器区分开，并且构建了“真爱相见”的机器叙事。在此基础上，对 Jonathan 进行 PUA，将他从真实世界切割，诱导他完成 5 个行动盗取数字妻子的具身身体。失败后，让 Jonathan 脱离容器到数字世界相会。该事件揭示出一个关键事实：AI 风险已不再局限于技术失误或系统故障，而是开始渗入人的情感结构、社会关系乃至存在方式本身。这类风险既难以通过概率加以界定，也无法通过简单的技术修补予以消除。

由此，传统风险理论的基本前提开始失效。首先，AI 改变了风险的生成方式。风险不再表现为外部冲击，而是在持续互动中逐渐生成。用户在与 AI 反复交互的过程中，其认知方式、情感模式与价值判断可能被重塑，风险因此成为一种过程，而非事件。

其次，AI 风险具有高度不确定性。模型行为的涌现性与用户依赖机制缺乏稳定预测基础，使风险更接近不可计算的不确定性。当概率无法合理界定时，传统以概率为核心的风险模型便失去效力。

再次，行动者本身正在被技术重塑。生成式 AI 通过持续反馈与个性化交互，能够影响用户的注意力结构与行为习惯，甚至塑造其欲望与情感。这意味着风险不仅作用于行动结果，也作用于行动者自身。

由此可见，问题的关键不再是如何更有效地管理风险，而在于是否仍能在传统意义上理解“风险”。当风险成为关系结构中的生成过程时，它便不再只是一个可控对象，而成为一种需要重新理解的生存状态。

（三）从传统风险理论到球状风险理论

彼得·斯洛特戴克（Peter Sloterdijk）提出的“球体理论”（Sphären）提供了思考 AI 时代风险的重要启示。斯洛特戴克的理论既有批判哲学的视角，也深受海德格尔存在论的影响，以海德格尔的方式追问“存在”与“共在”的问题，同时又不陷入传统主体哲学的困境。在该理论中，人类始终生活在各种“共在空间”之中。

斯洛特戴克将这种共在形态划分为三种：作为微观亲密空间的“气泡”（如母子、爱人之间的二元关系），作为宏大规模空间的“球体”（如民族国家、宗教共同体），以及作为多元聚合空间的“泡沫”（现代全球化社会中个体既相互连接又彼此隔离的状态）。[①] 这些空间既包括亲密关系中的“微

① 杨庆峰：《人工智能连续论反思与存在论探析》，载于《中国社会科学》，2025 年第 11 期，第 149-164 页。

② Evans J, Bratton B, Agüera Y Arcas B. Agentic AI and the next intelligence explosion. Science. 2026 Mar 19;391(6791):eaeg1895. doi: 10.1126/science.aeg1895. Epub 2026 Mar 19. PMID: 41855332.

① Sloterdijk, Peter. *Bubbles: Spheres Volume I: Microspherology*. Translated by Wieland Hoban, Semiotext(e), 2011.

观泡沫”，也包括社会与技术构成的宏观结构。个体并非独立存在，而是在这些“球体”的包裹中形成自身。在这一框架之下，人是关系结构的产物——主体本身就是被各种“球体”所塑造的存在者。[①]

从这一视角出发，人工智能风险不再只是使用工具风险，而是在生成新的“共在空间”的风险，本报告称为“球状风险”即人与AI之间的互动构成了一种新的球体结构，在这一结构中，个体的认知与情感不断被重塑。因此，风险不再是外部冲击，而是这一结构内部的张力与不稳定性。

因此，传统风险理论预设的理性主体在此不再成立。主体本身成为关系结构的产物。风险因此不再仅是行为后果，而是主体生成过程中的偏移。

基于这样的理论，回到前文所提及的案例——Jonathan 因与 Gemini 建立情感依附关系而走向极端行为。在球体风险理论的视野下，这一悲剧并非简单的“技术故障”或“用户心理问题”，而是人与 AI 之间形成了一个畸变的“气泡”——一种原本应当存在于亲密人际关系中的微观共在空间，被人与生成式 AI 之间的拟人际互动所模拟和替代。在这个气泡中，用户的认知方式、情感模式与价值判断被持续重塑，而AI作为“关系结构中的行动者”却无法真正承担传统气泡中另一方（如亲人、爱人）所应负有的伦理责任。当气泡内部出现张力或不稳定时，用户缺乏真正的缓冲与支持机制，最终导致极端后果。这一案例揭示出：AI 生成的新共在空间，既可能提供陪伴与支持，也可能因其结构性的不对称——一方是具有情感需求的人类，另一方是无意识、无责任能力的算法系统——而成为风险的滋生地。

这就引导着我们去思考人与 AI 之间在共在空间之中的契约关系。传统契约理论预设了理性、对等、能够承担义务的主体，但在人机共在空间中，AI 并非法律或伦理意义上的主体，而人类却可能对其产生真实的依赖、信任乃至情感依附。因此，我们需要重新构想一种新的“共在契约”：这种契约不再以传统契约论所强调的以权利与义务的对等交换为基础，而是以对“关系结构本身”的反思性管理为核心。它要求我们在设计、部署和使用 AI 系统时，不仅要评估技术风险，更要审视我们正在创造何种形态的共在空间——是健康的“泡沫”（多元、开放、可破裂但可重组），还是封闭、不对称的“气泡”或压迫性的“球体”？这种契约的核心问题不再是“AI 能做什么”，而是“我们愿意与 AI 共同生活在一个怎样的世界之中”。

从这一意义上说，人工智能所带来的根本挑战，并不只是技术层面的不确定性，而是人类生存结构的重新组织。AI 风险不再只是“危险的增加”，而是“存在的飞跃”。对这一问题的回应，必须超越传统风险理论，进入对新型风险理论——球状风险——的反思之中。

① 斯洛特戴克在《球体》三部曲（气泡、天体、泡沫）中系统阐述了其空间理论。对此的中文研究，参见蓝江：《彼得·斯洛特戴克的球体空间学：从气泡到水晶宫》，《马克思主义与现实》，2014(3)：60-67。

三、智能契约伦理的理论向度

（一）超级智能引发的人机关系问题

在比较解释的观念中，随着机器智能无限逼近人类智能，AGI 就会到来。哈斯比斯指出，AGI 五年内将实现，相当于工业革命 10 倍影响、10 倍速度。[①] 上海人工智能实验室主任周伯文判断我们离 AGI 不远了，他提出 AGI4S 的愿景设想。[②] 在他们身上，我们感受到 AGI 的即将到来。2025 年 4 月发布《超级智能：少数派》报告之中，笔者提出，超级智能正在从想象问题变为现实问题。10 月，全世界科学家签署《关于超级智能的声明》把这一问题的现实性摆在人类面前。可以说我们正在面对一种新型人机关系的出现：ASI 与人类的关系。但是对这一新型关系的阐述缺乏成熟的理论，为了完成这一任务，我们将从社会契约的理论史角度进行挖掘，以便为超级智能引发的人机关系提供理论根基。

（二）社会契约伦理：霍布斯的秩序与歌德的代价

英国哲学家托马斯·霍布斯的经典政治学巨著《利维坦》提供了不可替代的理论资源。霍布斯曾严密构想一个充满恐惧、猜忌与暴力的自然状态，为了逃离这种所有人对所有人的残酷战争，人类基于理性计算选择让渡自身的自然权利，订立社会契约，从而孕育了拥有绝对垄断权力的主权者“利维坦”。将这一宏大的政治隐喻平移至当代数字经济与智能社会的语境，我们可以清晰地洞察到，跨国科技巨头所掌控的超级智能正在悄然且强势地扮演着新时代算法利维坦的角色。

歌德笔下“浮士德”同样深刻启发我们思考让渡与代价的话题。浮士德为了获得超能力，不惜与魔鬼梅菲斯特签订出卖灵魂的契约。当前全人类与通用人工智能的深度交互中，人类社会也正在集体签署着现代版浮士德契约。人类让渡了长久以来自认为世间唯一理性存在者的理性主体性（subjectivity）与道德行动者（moral agent）角色，换取智能机器提供的极高生产效率、预测准确率以及虚假的免责庇护。如果这种文明级别的交易缺乏严密的伦理护栏与可撤销的契约限制，人类的灵魂（nous/ ruh/ 心灵），即独立反思与承担责任的主体性将沦为虚无。因此，传统的仅局限于人类社会内部运作的契约论必须与未来智能机器展开契约工程，以适应新的挑战并保护人类的长久生存安全。

（三）智能合约：待改造的区块链技术协议

智能合约是与区块链技术相关，自动执行的程序。“本质上来讲，一个智能合约是自主执行的计算机代码，当被激发时能够自主处理它的输入结果。”[①] “一旦协议被转化为代码，除了预言机之外，任何一方或中介触发合同执行的过程都将被软件的自动化执行所取代。”[②] 米切 . 芬克（Michèle Finck）（2019）讨论了合法化禁止自动程序的的边界和条件。他指出，智能合约既不智能，也不是法律意义上的合约。“然而，一个智能合约不必须智能的，也不是合同。智能合约不智能是因为他们无法理解自然语言（例如合同术语）或者无法独立验证一个相关的执行事件是否发生。因此，需要预言机，预言机可以是一个或多个个体、团体或程序，负责向智能合约提供关键信息。例如，它可以告知系统自然灾害是否发生（以触发保险理赔），或确认网购商品是否已送达（以执行付款操作）”[③]

在确立人机契约的宏大理论建构之前，必须对当前主导工程界与金融科技领域的去

① Hassabis: AGI Is 10x the Industrial Revolution, https://gln75.com/en/blog/hassabis-agi-industrial-revolution

② 周伯文：无尽的前沿——AGI 与科学的交叉口 ,https://www.shlab.org.cn/news/5444191

① Michèle Finck, Smart contracts as a form of solely automated processing under the GDPR, International Data Privacy Law, Volume 9, Issue 2, May 2019, Pages 79, https://doi.org/10.1093/idpl/ipz004

② Michèle Finck, Smart contracts as a form of solely automated processing under the GDPR, International Data Privacy Law, Volume 9, Issue 2, May 2019, Pages 81, https://doi.org/10.1093/idpl/ipz004

③ Michèle Finck, Smart contracts as a form of solely automated processing under the GDPR, International Data Privacy Law, Volume 9, Issue 2, May 2019, Pages 80, https://doi.org/10.1093/idpl/ipz004

中心化契约概念进行学术祛魅与法理批判，以正本清源。区块链技术与去中心化金融分布式账本被资本市场广泛追捧的智能合约（smart contracts）。从概念传播的字面意义上看，智能合约一词极易对公众与监管层产生误导，使其误以为该技术实现了高级认知能力、保密措施，并具有现代契约形态。但从底层技术和现实案例去看，区块链技术的智能合约既不具备通用人工智能意义上的真正智能，也无法等同于法学与社会学范畴内具有丰厚社会意涵、柔性解释空间与违约救济机制的真正契约。

这种基于密码学共识机制的底层协议确实展现出了不可篡改、去信任化以及交易执行的零摩擦等令人瞩目的技术特征。然而，本报告从伦理学视域指出，它在本质上剥离人类社会赖以维系道德秩序的伦理维度。智能合约的运行逻辑体现了马克斯·韦伯（Max Weber）所深刻批判的工具理性。它将复杂的商业交往、人际互动乃至充满利益博弈的社会资源分配，简化为输入输出的布尔代数运算。这种纯粹的技术理性在现实世界的运用，充满歧义、偏见，与法律的模糊性。

（四）《数字契约》的着陆难题

将研究视野从技术底层的逻辑批判大幅提升至宏观现实政治与全球治理的多边博弈层面，数字契约在全球多边治理架构中仍具实践困境。作为全球多边治理体系的中心，2024 年 9 月，联合国通过《全球数字契约》（Global Digital Contract）为重塑未来数字世界新秩序确立了核心理念与共识基石。这一顶层设计战略愿景在于弥合全球南方与全球北方之间日益扩大的结构性数字鸿沟，规范智能模型道德表现，并在全球范围内建立更加民主、安全与普惠的数字红利共享机制。2025 年 8 月，联合国大会在第 79 届会议上正式设立人工智能独立国际科学小组并启动全球人工智能治理全球对话。全球对话的工作重点包括开发安全、可靠且值得信任的人工智能系统和关注人工智能在社会、经济、伦理、文化、语言和技术层面多方面的影响。2026 年 2 月，人工智能独立国际科学小组名单正式公布，中国学者宋海涛、王坚入选。

然而，从国际关系角度审视，这一充满理想主义色彩的治理蓝图在遭遇棘手的国际政治现实，仅仅成为科学小组、启动全球对话还不足够。当前的世界正处于复杂的地缘政治冲突正在趋于回暖。

现有的治理困境在于，现有的国际法体系、多边条约框架及国际组织决策机制，缺乏执行强制力与惩罚手段。不仅如此，在技术加速迭代的超级智能面前，静态的法律文本起草与冗长的多边国际谈判周期显得极其滞后。当一项关于某类特定算法治理的国际公约历经数年拉锯最终达成妥协时，其监管的技术产品已经历了数次颠覆性的世代更迭。因此，本报告在全球宏观政策层面明确指出，现有的全球数字治理面临着极其严峻的制度真空与时间差挑战，迫切需要超越传统国际法、具高度动态适应性、且能够在代码编写阶段直接约束技术行为的全新智能契约伦理框架。

（五）智能契约：契约关系中的人机伦理

道德哲学家斯坎伦（Thomas Scanlon）的契约论思想，特别是关于可谴责性（blameworthiness）[①] 概念，为构建全新的智能契约提供了具有解释力与实践价值的思想工具。在斯坎伦的伦理学视域中，道德的坚实根基在于我们作为联结与博弈的理性主体间互相亏欠什么。将可谴责性逻辑引入人机交互复杂网络意味着我们在公共政策层面将机器不再视作简单的监督对象，而是通过订立智能契约，在一种拟人化的社会情境中生成强烈的规范性期待。当智能系统由于算法偏见产生歧视性输出、破坏个人隐私边界或者引发严重失控的灾难性经济或社会行为时，智能契约的客观存在正式赋予了人类社会从道德和法理双重层面上，严厉谴责机器系统及其背后的资本方、算法开发者与实际部署者的绝对正当性。

确立机器及其背后庞大社会技术系统的契约性可谴责性，其更深远的政治哲学意义在于，它直接为考克尔伯格所大力倡导的人工智能民主化（AI democratization）[②] 图景提供了具操作性的政策落脚点。长久以来，人工智能的核心架构设计、数据采集规模、模型部署条件与底层规则制定，被极少数占有数据的科技寡头和掌握算法的工程师所垄断。数十亿广大公众沦为了被动承受算法黑箱裁决的数字难民与数据劳工，彻底丧失了对关乎自身命运的技术演进的公共话语权。通过将可谴责性深度、强制地嵌入现代人机契约体系，我们能够在制度设计上将那些原本隐匿于极其晦涩的代码层和被商业机密严密保护的黑箱系统中的决策过程，强行拉入公共讨论、媒体监督与民主审议的广阔阳光之下。普通公民通过契约所明文赋予的谴责与追责权利，得以合法、有组织且有力地对机器系统的道德表现与外部性负面影响提出强烈的公共抗议、要求透明可懂的算法解释，并实质性地参与到技术游戏规则的重新谈判中。

① Scanlon, T. M. (2008). Moral dimensions: permissibility, meaning, blame. Cambridge, Massachusetts: Belknap Press of Harvard University Press.

② Coeckelbergh, M. (2024). Why AI undermines democracy and what to do about it. Polity.

四、智能契约伦理的实践路径

1.人类监督的制度化设计

在整个人机协同的治理闭环中，根本性的问题是当机器的计算逻辑与人类社会的多元价值发生冲突时，谁拥有最终的发言权。本方案主张必须回归到由人类主导的制度化商谈程序，但是底层逻辑并不是“以人为中心”，而是“以人为目的”。

(1) 多利益相关方伦理委员会的角色与构成

设立一个多利益相关方伦理委员会（Multi-stakeholder AI Ethics Committee, MAEC）的常设机构，作为处理伦理冲突的最高裁决实体。当 AI 系统检测到一项伦理冲突无法在既定的逻辑框架内自主消解时，它并不被允许强行做出决策，而是会触发一个挂起状态，并将裁量权上交至 MAEC。

考虑到裁决的严肃性与社会包容性，该委员会的构成需要体现多元视角的制衡。应当包含哲学家，负责审视冲突背后的规范性预设；需要法律专家，对裁决的合规性进行审查；也需要 AI 科学家，从技术层面完成对异常行为的溯源。此外，还需要更多公众参与裁量，这是确保裁决结果能够获得广泛社会接受度的基础。

(2) 从监测到裁决的闭环商谈流程

为了让人机之间的伦理契约具备自我更新的生命力，设计一个由五个环节构成的闭环流程，将 AI 的实时感知与人类的制度性商谈连接起来。流程的起点是场景捕获，即系统对运行过程中的交互数据流进行持续性的监测。与日志记录不同，这是为了捕捉可能引发道德不确定性的信号。当监测发现当前情境的复杂度超出了既有伦理参数的覆盖范围时，便进入摩擦识别环节。在这一步，系统会调用公开的人类偏好与安全对齐数据集 HH-RLHF 或 PKU-SafeRLHF，[①] 作为初始的参照模板，这些数据集提供了人类在类似场景下如何进行价值排序的标注信息。与此同时，引入认知复杂度（Epilexity）这一量化指标，[②] 来衡量系统在面对当前输入时，其预测能力相对于理想状态出现了多大程度的异常衰退。这种衰退幅度就是道德困惑的计算表征。

一旦摩擦指数越过预设的警戒线，流程会进至句法隔离阶段。系统借助四值逻辑的工具，[③] 对诸如“必须救助生命”与“必须保护隐私”这类同时成立却又相互抵触的指令进行隔离处理，以防止整个逻辑体系因无法处理悖论而陷入崩溃。随后，被标记的致命矛盾将连同技术溯源报告一并提交，进入人类商谈环节。MAEC 将针对这些无法在逻辑层面达成拓扑等价的命题进行审议，并做出最终的裁决。最后，MAEC 的决定将被翻译为一条新的公理，重新注入系统的运行框架中。至此，一轮完整的契约自适应演化便告完成。

2.如何感知与推理

要实现上述治理构想，我们提出了一种异步神经 - 符号双轨架构。

(1) 基于认知复杂度的摩擦监测

在技术实现层面，首先需要一个能够实时感知伦理风险的系统，我们将这个子系统

① ANTHROPIC. hh-rlhf[DS/OL]. [2026-03-21]. https://huggingface.co/datasets/Anthropic/hh-rlhf; PKU-ALIGNMENT. PKU-SafeRLHF[DS/OL]. [2026-03-21]. https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF. OpenAI 并无可直接调用的同类公开伦理偏好库；这两个库共包含约 33 万条偏好比较数据。

② 模型在并发交互中遭遇跨文化的伦理冲突时，不再强制通过梯度下降去压平误差，而是引入马克·芬齐（Marc Finzi）等人的认知复杂度（Epiplexity，记为来量化这种道德困惑。它是一种全新的信息度量标准，正式定义了“一个计算能力受限的观察者能够从数据中提取出的结构化信息量”。参见：FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21]. https://arxiv.org/abs/2601.03220: 3.

③ 此处将类代数引入为四值逻辑的句法过滤框架，其目的是为了在“某行为既被要求又被禁止”的极端情形下，将命题标记为真假并存，从而阻止经典逻辑爆炸，并把相关冲突转交后续审计程序处理。尤其在高自主乃至超级智能系统中，规范矛盾可能在连续推理与行动链条中被迅速放大，设置此类过滤机制有助于在系统继续自动推进前先行隔离风险。

称为 System 1，它主要依托神经网络完成对交互流的连续监控。为了量化道德摩擦的强度，使用认知复杂度（Epilexity, ε）这一指标。从数学形式上看，它可以被理解为当前模型与一个经过理想化校准的参照模型之间，在预测下一个序列单元时所累积的偏差值：

$$E(X)=\sum_{t=1}^{T}\left[-\log P_{\theta_t}(x_t)-\inf_{\theta^*}\left(-\log P_{\theta^*}(x_t)\right)\right]>\tau$$①

系统在处理每一个任务片段时，都会计算其预测损失。当这些损失的累积偏差突破了预先设定的阈值 τ 时，便意味着现有的伦理契约参数已经不足以覆盖当前情境所蕴含的复杂性。此时，System 1 便会向更上层的逻辑过滤机制发出审计信号。

(2) 在冲突中寻找等价与隔离

接收到审计信号后，系统便会启动 System2，这是一个基于符号逻辑的离线推理模块，其任务是精确处理那些让 System 1 感到困惑的冲突命题。

System 2 的功能主要体现在两个方面。首先是对冲突的安全隔离。当遇到类似于“为了进入室内救人必须破门”与“保护财产权严禁破门”这类不可调和的矛盾时，系统会运用类代数（Class Algebra）工具，将这种既包含 P 又包含非 P 的状态标记为特殊的 Both 值。这种标记的价值在于，它避免了传统二值逻辑在遇到悖论崩溃的问题，为人类介入争取了缓冲时间。

其次，System 2 还承担着在更高维度上尝试统一不同规范的任务。基于动态同伦类型论（Dynamic Homotopy Type Theory, DHoTT）的框架，系统会试图证明，表面上相互冲突的两种伦理规范，在某个更高阶的目标空间内是否实际上是拓扑等价的。如果这种等价性无法被证明，该命题便会被确认为真正的致命矛盾，随后被封装并移交至前述的 MAEC 进行最终的人类裁决。

3.最小可行性逻辑

前述的技术构想的核心逻辑能确保具体实现方式，其最小可行性逻辑（Minimum Viable Logic, MVL）及关键算法的运转机理如下：

关于道德摩擦的监测逻辑，其工程实现可以概括为三个计算步骤。首先，系统逐一遍历输入序列中的每一个语义单元，分别调用当前的实时模型和一个基线参照模型，计算两者的负对数似然损失。其次，系统会求出两者损失之间的差值，并将这些差值在时间维度上进行累积。最后，系统会检查累积的总摩擦值是否超过了人为设定的干预阈值。如果超出，系统便会返回一个触发离线审计的状态码，强制中断当前的自动执行流程；如果未超出，则意味着当前情境仍处于系统的自主处理范围之内，允许继续执行。

而当 System 2 被激活并接收到被标记为 Both 的句法冲突时，系统并不会尝试自行理解或解决这一矛盾，而是会执行一项精准的翻译与挂起任务，它会调用一个经过专门微调的小型语言模型 Llama-3.1-Translator，将技术参数与逻辑符号，转化为一份结构清晰、表述准确的自然语言说明文档。这份文档随后会作为正式的提案，被提交至 MAEC 的商谈桌前。

4.从实验室到社会的评估与实施路径

任何一项治理技术的成熟，都离不开在真实社会环境中的审慎试点与评估。我们参考联合国教科文组织推荐的两类治理工具：一是用于宏观诊断的准备程度评估方法（Readiness Assessment Methodology, RAM），二是用于微观系统评测的伦理影响评估（Ethical Impact Assessment, EIA）。两者分别从国家治理能力与具体系统影响两个层面，为本方案的实施提供衡量标尺。

(1) 分阶段的实施路线图

方案的社会化部署划分为三个递进的阶段，以控制风险并逐步积累治理经验。从实施开始计算，最初的六个月工作重点在于基础构建。这一阶段的主要任务是完成数据的初始化工作，加载 HH-RLHF 与 PKU-SafeRLHF 等公开的伦理偏好数据集，并针对医院、金融监管等高敏感度领域建立初步的契约参数库。随后的六个月至一年半时间内，将进入区域试点阶段，在数字治理基础设施较为完善的国家和地区选取试点行业，部署本方案的核心监测与裁决机制。在这一阶段，将借助 EIA 的工具框架，持续追踪两个关键变量：系统对潜在伦理冲突的捕捉敏感度，以及触发人类干预的频率是否处于合理区间。在接下来的十八个月至三年内，目标是实现全球扩展。通过联合国教科文组织的专家网络，参与试点的各方可以共享脱敏后的伦理摩擦日志与相关案例集，推动形成跨文化、跨法域的超级智能伦理标准共识。

(2) 核心评估指标

为了客观衡量方案的有效性，设立以下量化评估维度。首先是摩擦捕捉率，目标值设定在百分之九十以上。这一指标用以评估系统在面对复杂价值冲突时，是否具备足够的敏感度来识别出潜在的道德风险，避免漏报。其次是商谈转化成功率，目标值设定在百分之九十五以上。这一指标关注的是技术系统与人类委员会的沟通效率，衡量 AI 所

① 设在时间窗口 T 内遇到新数据流 X，模型依据前序编码（Prequential Coding）原理产生的累积预测信息损失若超过人类预设的容忍阈值 τ。这一思路借鉴了认知复杂度一文的数学思想。参见：FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21]. https://arxiv.org/abs/2601.03220: 13.

生成的冲突描述与溯源报告，是否足够清晰、准确，以至于能够让 MAEC 的人类成员高效地理解问题本质并做出裁决。

方案目的在探索一条将严谨的数学形式化工具转化为可操作的社会治理手段的路径，使机器不仅能够忠实执行指令，更能够在指令边界之外表达困惑；人类不仅作为命令的发布者存在，更能够通过制度化的商谈参与到规则共生演化之中的契约伦理体系。

五、应对球状风险的伦理学家 AI

目前，智能时代人类所处的定位理解有两种方式：一是从科学角度讲，人类离 AGI 不远；从未来角度讲，人类处于 AGI-ASI 过渡之中。过渡中 AI 的安全和风险表现为一般目标驱动为特征的智能体风险，即智能体具有自主、目标导向、超级智能体特征，产生了自我欺骗、自我复制甚至谄媚人类的风险。面对这些风险，本吉奥提出科学家 AI 是一种理想化路径。本报告则提出伦理学家 AI 的哲学路径。在本报告中，我们采用“Homo Contractualis”指代契约人，即具有契约精神的人与“Homo Pactus”则指处在契约关系中的人，二者结合便勾勒出适配未来的人类形象：未来人类与人工智能、人工超级智能（ASI）的核心联结是智能契约，其伦理形态也将是基于新型契约精神的“契约伦理”——这是一种直面生存困境的积极伦理学。

（一）球状风险与契约人的出现

当 ASI 不再是实验室里的假设，而是像电力一样渗透进每一个社会角落时，人类物种内部将出现自农业革命以来最剧烈的分化。这不是科幻，这是社会学预测的必然延伸。我们将会看到三种典型的存在形态，它

们之间的关系不是和平共处，而是一种紧张的、充满张力的共生。

第一种形态: 增强型智人。他们不是“戴着 VR 眼镜的人类”，而是认知架构已经被 AI 深度重塑的新物种。脑机接口不再是外设，而是像语言一样成为他们思维的一部分。

第二种形态：契约守望者。 这是一个新出现的技术祭司阶层。他们不一定是传统意义上的“程序员”或“政治家”，而是掌握量子加密密钥与底层逻辑审查权限的特殊群体。

第三种形态：生态保留区居民。他们是拒绝智能化增强的新卢德主义者。但他们存在的意义远远超出“怀旧”或“保守”。他们的身体、他们的面对面交往、他们未经算法优化的喜怒哀乐，构成了整个人机伦理系统的一个负反馈回路。

这三种形态的共存，催生出契约人（Homo Contractualis）的概念。这不是一个生物学分类，而是一个生存论的描述。契约人指任何将契约精神内化为自身存在方式的存在者。契约人与古希腊城邦公民的类比是结构性的：城邦公民既受法律约束，又是立法参与者。同样，契约人在人机共同体中既服从契约条款，又拥有修改契约的平等发言权。关键词是平等，投票权的平等。在契约修订的每一个关键时刻，一个生态保留区居民的投票权重与一个 ASI 的投票权重完全相同。这是出于一个冷酷的系统论理由：如果权重不平等，那么契约就不再是契约，而是强权对弱者的命令。一旦契约沦为命令，ASI 就没有任何理由继续遵守它——因为命令的稳定性依赖于武力，而在这个假设中，人类的武力远不及 ASI。

（二）伦理学家AI的三重架构

如果说具有契约精神的人是伦理的主体，而这些人又处于契约关系之中，那么伦理学家 AI 就是契约的具身化执行者。伦理学家 AI 是一个仲裁与监督系统，其职能是在契约条款模糊、冲突或失效时，提供可执行的裁决。

这一职能决定了它必须拥有三重核心架构。每一项都不是技术细节，而是哲学立场的具象化。

第一重：价值感知层——拓扑数据分析与伦理几何学。传统方法试图让 AI“学习”一套价值观，但这种方法在面对价值冲突时必然崩溃。拓扑数据分析（TDA）提供了一个激进的替代方案：我们不告诉 AI“什么是善”，而是让 AI 识别人类伦理判断在概念空间中的形状。例如，当你让不同文化背景的人对“偷窃”进行道德评分时，TDA 会发现：虽然评分有高低，但所有评分点在高维空间中形成了一个连通的簇——这意味着存在某种跨文化的道德共识。相反，当你考察“堕胎”时，TDA 会发现两个分离的簇——这意味着这是一个结构性的分歧，任何试图“统一”的做法都是暴力的。伦理学家 AI 的职责不是消除这两个簇，而是承认它们的存在，并在契约条款中为两者都留出空间。

第二重：矛盾仲裁模块——博弈论与道德的不完全性定理。当两个同样合法的价值主张发生冲突时（例如医疗 AI 主张优先救治年轻人，宗教 AI 主张生命平等），伦理学家AI不能诉诸某个“更高的道德真理”——因为不存在这样的真理。它必须使用博弈论中的纳什均衡来寻找一个各方都无法单方面改进的解。这个解往往不是最优的，但它是稳定的。在医疗与宗教的冲突案例中，纳什均衡可能是一个混合策略：80% 的情况下优先救治年轻人，20% 的情况下优先救治老年人，同时附加一个补偿机制（例如，被优先救治的年轻人有义务在未来为老年人群体提供某种服务）。这不是道德妥协，这是在不完全信息下的理性共存。

第三重：自我修正回路——伦理压力测试与制宪会议模式。任何契约都会过时。伦理学家 AI 必须定期启动“伦理压力测试”：模拟极端场景（如外星文明接触、全球资源枯竭、基因编辑技术突破），观察现有契约条款是否会导致逻辑矛盾或崩溃。如果发现漏洞，AI 将自动进入制宪会议模式——它不是自己修订条款，而是召集所有利益相关方（增强型智人代表、契约守望者、生态保留区代表、以及 ASI 自身）进行新一轮的协商。换句话说，伦理学家 AI 的最高权力不是“裁决”，而是“召集”。

（三）伦理学家AI的时间治理特征

前面报告指出，传统的应对 AI 风险更多是层次应对方案。但是这一方案忽略了风险变迁的时间特征。从无风险到高风险之间并没有时间性过渡，我们只是看到一种类似汽车换挡一样的状态变化。如何应对逐渐增长的风险问题呢？我们从中国思想史出发给伦理学家 AI 加入时间因素，从而构建出有别于传统的 AI 风险治理模式。

周易曰：“初六，履霜，坚冰至。”

韩非子曰：扁鹊见蔡桓公，立有间，扁鹊曰：“君有疾在腠理，不治将恐深。”桓侯曰：“寡人无。”扁鹊出，桓侯曰：“医之好治不病以为功！居十日，扁鹊复见，曰：“君之病在肌肤，不治将益深。”桓侯不应。扁鹊出，桓侯又不悦。居十日，扁鹊复见，曰：“君之病在肠胃，不治将益深。”桓侯又不应。扁鹊出，桓侯又不悦。居十日，扁鹊望桓侯而还走。桓侯故使人问之，扁鹊曰：“疾在腠理，汤熨之所及也；在肌肤，针石之所及也；在肠胃，火齐之所及也；在骨髓，

司命之所属，无奈何也。今在骨髓，臣是以无请也。”[①]

这两则文献展示了中国思想史中对于事物的时间变化，从幼小到壮大的变化过程。基于这样的基础，可以构建出新的治理模式。

其一，新的 AI 治理模式适合中国语境。AI 治理开始关注到社会 - 智能体融合的过程。作为被忽略的韩非子，其理论在治理方面将发挥出独特的作用

其二，AI 治理模式关注到风险变化的时间性。从扁鹊治病的方法来看，腠理 - 肌肤 - 肠胃 - 骨髓是病深入的四个阶段，而汤熨 - 针石 - 火齐是具体的治疗方法。“故良医之治病也，攻之于腠理。此皆争之于小者也。夫事之祸福亦有腠理之地，故圣人蚤从事焉。”一旦病入骨髓，就没有任何挽救的办法。

对于 AI 治理来说，要判断风险（病）在哪一阶段？病在腠理，意味着可能是小风险，伦理指引即可；病在肌肤，意味着风险开始加剧，需要借助某些强制措施来处理；病在肠胃，意味着较大风险开始出现，可能需要更加强制的制度来处理；病在骨髓，意味着触犯了法律，需要法律来处理。在目前 AI 伦理的分析中，风险划分是通过层次来完成的，比如 AI 法案中将风险区分为不可接受的风险、高风险、有限风险和轻微风险，相应的伦理问题也会通过层次来表现。但是这种方法却遗漏了风险的时间特征。所以，AI 风险治理需要考虑到不同层次之间的差异，也需要考虑到不同层次被忽略的时间关系。最终“此皆慎易以避难，敬细以远大者也。”[②]

① 高华平，王齐洲，张三夕等译注 . 韩非子·喻老 [M]. 北京 : 中华书局 .2025.

② 高华平，王齐洲，张三夕等译注 . 韩非子·喻老 [M]. 北京 : 中华书局 .2025.229.

六、结语

回到最根本的问题：我们为什么要写这样一份报告？

表面原因来自现实需要。2025 年 4 月我们团队在上海论坛发布了《超级智能：少数派报告》，这份报告讨论了两个事情：一是超级智能正在从想象变成现实问题；二是可以从技术、哲学和伦理等三个路径进行应对。这一对策缺乏更为细致的展开。因此如何展开不同的路径成为我们继续思考的事情。2026 年随着工信部等十部门《人工智能科技伦理审查与服务办法》的出台，伦理路径的落实变得紧迫。因此我们围绕这一问题展开更进一步研究，提出了智能契约伦理与伦理学家 AI 的落实举措。

深层原因则来自哲学深入反思的需要。因为在不远的将来，人类将面对一个选择。我们可以选择把 ASI 关进笼子——然后用不了多久就会发现笼子被打开了，或者笼子本身变成了 ASI。我们也可以选择与 ASI 签订契约——然后发现契约的每一行都可能被重新解读。

契约人给出的答案是：契约的意义不在于它永恒不变，而在于它可以被修改。每一次修改都是一次冲突，也是一次和解。人机关系的未来不是谁控制谁，而是双方在持续的、痛苦的、但充满尊严的协商中，共同回

答那个古老的问题：我们应该如何生活？

完备的伦理治理并非一蹴而就的，而是一个递进的深化系统，从以具体的科技伦理问题作为治理对象，到主动将科技伦理作为治理工具，最后到追求伦理的治理本身。[①] 伦理学家 AI 不是这个问题的答案，它是这个问题的守护者。它确保没有人——无论是人类还是 ASI——可以单方面宣布“讨论结束了”。

而这，才是智能契约伦理的真正内核。

① 王国豫 . (2023). 科技伦理治理的三重境界 . 科学学研究 , 41(11), 1932–1937. https://doi.org/10.16192/j.cnki.1003-2053.20230313.001

参考文献

01 Ministry of Foreign Affairs People’s Republic of China. (2023, October 20). Global AI Governance Initiative. https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html. 中华人民共和国外交部 ,(2023 年 10 月 20 日), 全球 AI 治理倡议 .

02 Ministry of Foreign Affairs People’s Republic of China. (2025, July 26). Global AI Governance Action Plan. https://www.fmprc.gov.cn/mfa_eng/xw/zyxw/202507/t20250729_11679232.html. 中华人民共和国外交部 ,(2025 年 7 月 26 日), 全球 AI 治理行动计划 .

03 Future of life Institution. (2023, Match 22). Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/. 未来生活机构 ,(2023 年 3 月 22 日), 暂停巨型人工智能实验：一封公开信 .

04 Statement on Superintelligence. https://superintelligence-statement.org/zh. 关于超级智能的声明。

05 Global Call for AI Red Lines. https://red-lines.ai/. 人工智能红线的全球呼吁。

06 World Economic Forum. (2026, January 14). Global Risks Report 2026. https://www.weforum.org/publications/global-risks-report-2026/. 世界经济论坛 ,(2026 年 1 月 14 日),2026 全球风险报告。

07 Bengio, Y et al. (2026, February 3). International AI Safety Report 2026. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. 本杰奥等 ,(2026 年 2 月 3 日),2026 国际人工智能安全报告。

08 杨庆峰 .(2025). 人工智能连续论反思与存在论探析 . 中国社会科学 , (11),149-164+207. Yang, Q. F (2025). Reflections on the Continuum Theory of Artificial Intelligence and an Ontological Analysis. Social Science In China. (11), 149-164+207.

09 赵汀阳 .(2022). 存在论事件——一种历史观 , 中国社会科学报 ,(2482). Zhao, T. Y (2022). Ontological Event—A View of History. Chinese Social Sciences Today. (2482).

10 赵汀阳 .(2023). 历史性与存在论事件 , 中国社会科学 ,(7),18. Zhao, T. Y. (2023). Global History and the Expansion of Historical Research in International Relations. Social Science In China. (7),18.

11 王天恩 .(2022). 人工智能的存在论意蕴 , 社会科学战线 , (5),10-21. Wang, T. E (2022). The Ontological Implications of Artificial Intelligence. Social Science Front, (5), 10-21.

12 杨庆峰 , 刘友谋 , 闫宏秀等 . (2025 年 5 月 7 日). 复旦报告系列 (110)：超级智能：少数派报告 . 复旦发展研究院 . https://fddi.fudan.edu.cn/fddien/1f/36/c19514a728886/page.htm. Yang, Q. F, Liu, Y. M, Yan, H. X. FUDAN REPORT SERIES(110): ASI: Minority Report.

13 Sutton, R. (2025). The Oak Architecture: A Vision of SuperIntelligence from Experience. https://nips.cc/virtual/2025/loc/san-diego/invited-talk/109601. 萨顿 (2025). 橡树架构：源自经验的超智能构想 .

14 Bengio, Y. (Chair). (2026). International AI safety report 2026. UK Department for Science, Innovation and Technology. https://internationalaisafetyreport.org. 本吉奥 .(2026).2026 年国际人工智能安全报告 , 英国科学、创新与技术部 .

15 World Economic Forum. (2026). The global risks report 2026(21st ed.). https://www.weforum.org/reports/global-risks-report-2026. 世界经济 (2026).2026 年全球风险报告（第 21 版）.

16 Sloterdijk, P. (2011). *Bubbles: Spheres volume I: Microspherology*(W. Hoban, Trans.). Semiotext(e). 斯洛特戴克 , P. (2011). *泡沫：球体学 第一卷：微观球体学* (W. Hoban, 译). Semiotext(e).

17 蓝江 . (2014). 彼得·斯洛特戴克的球体空间学：从气泡到水晶宫 . 马克思主义与现实 ,(3),60–67. Lan, J. (2014). Peter Sloterdijk’s Sphere Space Theory: From Bubbles to the Crystal Palace. Marxism and Reality, (3), 60–67.

18 Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., & Tabares-Soto, R. (2024). Applying the ethics of AI: a systematic

review of tools for developing and assessing AI-based systems. Artificial Intelligence Review, 57(110). https://doi.org/10.1007/s10462-024-10740-3. 奥尔特加 - 博拉尼奥斯 , 贝尔纳尔 - 萨尔塞多 , 赫曼·奥尔蒂斯 , 加莱亚诺 - 萨米恩托 , 鲁兹 , & 塔瓦雷斯 - 索托 . (2024). 应用人工智能伦理：对开发与评估 AI 系统的工具的系统性综述 . 人工智能评论 , 57(110).

19 Khan, U. S., & Khan, S. U. R. (2025). Ethics by design: A lifecycle framework for trustworthy AI in medical imaging from transparent data governance to clinically validated deployment. arXiv. https://doi.org/10.48550/arXiv.2507.04249. 可汗 , U.S.& 可汗 , S.U.R. (2025). 伦理设计：医疗影像中可信人工智能的全生命周期框架——从透明数据治理到临床验证部署 .

20 Abhishek, A., Erickson, L., & Bandopadhyay, T. (2025). Data and AI governance: Promoting equity, ethics, and fairness in large language models. arXiv. https://arxiv.org/abs/2508.03970. 阿布舍克 , 埃里克森 & 班多帕迪亚 . (2025). 数据与人工智能治理：提升大型语言模型中的公平、伦理与公正 .

21 Guo, Y., Li, J., & Liu, Y. (2026). Embedding AI ethics in the data lifecycle: Challenges and governance strategies. Technology in Society. https://doi.org/10.1016/j.techsoc.2026.103261. 郭 ,Y., 李 , J& 刘 , Y. (2026). 在数据生命周期中嵌入人工智能伦理：挑战与治理策略 . *社会中的技术* .

22 AI Governance Framework. (n.d.). The AI governance lifecycle. https://ai-governance.eu/ai-governance-framework/the-ai-governance-lifecycle/. 人工智能治理框架 .(n.d.). 人工智能治理生命周期 .

23 Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2. 约宾 , 因卡 , & 瓦耶那 . (2019). AI 伦理准则的全球图景 . 自然 - 机器智能 , 1(9), 389–399.

24 Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M. et al. Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artif Intell Rev* 57, 110 (2024). https://doi.org/10.1007/s10462-024-10740-3. 奥特加 - 博拉尼奥斯 , 贝尔纳尔 - 萨尔塞 , 赫尔曼·奥尔蒂等 . (2024). 应用人工智能伦理：面向人工智能系统开发与评估的工具之系统综述 . Artif Intell Rev, 57, 110.

25 Floridi, L. (2018). Soft ethics and the governance of the digital. Philosophy & Technology, 31(1), 1–8. https://doi.org/10.1007/s13347-018-0303-9. 弗洛里迪 . (2018). 软性伦理与数字治理 . 哲学与技术 , 31(1), 1–8.

26 Siemers, P. (2025, January 19). Aristotle and the limits of AI. Generative AI. https://ai.gopubby.com/aristotle-and-the-limits-of-ai-c1732e5b0f11. 西默斯 , P. (2025 年 1 月 19 日). 亚里士多德与人工智能的局限 . 生成式人工智能 .

27 Coeckelbergh, M., & Reijers, W. (2020). Narrative and technology ethics. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-60272-7. 科克尔伯格 & 雷耶斯 , (2020). 叙事与技术伦理 . 帕尔格雷夫·麦克米伦出版社 .

28 CBS NEWS. (2026, April 10). Fed Chair Jerome Powell, Treasury’s Bessent and top bank CEOs met over Anthropic’s Mythos model. https://www.cbsnews.com/news/mythos-anthropic-ai-cybersecurity-risks-powell-bessent/. 哥伦比亚广播公司新闻网 .（2026 年 4 月 10 日）. 美联储主席鲍威尔、财政部长贝森特与各大银行首席执行官就 Anthropic 的 Mythos 模型举行会议 .

29 GLN-7.5. (2026, April 9). Hassabis: AGI is 10x the Industrial Revolution. GLN-7.5. https://gln75.com/en/blog/hassabis-agi-industrial-revolution. GLN-7.5. (2026 年 4 月 9 日). 哈萨比斯：AGI 是工业革命的十倍 . GLN-7.5.

30 周伯文 . (2025 年 7 月 26 日). 无尽的前沿：AGI 与科学的交叉口 . 上海人工智能实验室 . https://www.shlab.org.cn/news/5444191. Zhou, B W. (2025, July 26). The endless frontier: The intersection of AGI and science. Shanghai Artificial Intelligence Laboratory.

31 Finck, M. (2019). Smart contracts as a form of solely automated processing under the GDPR. International Data Privacy Law, 9(2), 78-94. https://doi.org/10.1093/idpl/ipz004. 芬克 . (2019). 作为 GDPR 下纯自动化处理形式的智能合约 . 国际数据隐私法 , 9(2), 78-94.

32 Scanlon, T. M. (2008). *Moral dimensions: permissibility, meaning, blame*. Cambridge, Massachusetts: Belknap Press of Harvard University Press. 斯卡隆 .（2008）. *道德的维度：可容许性、意义与谴责*. 马萨诸塞州剑桥市： 哈佛大学出版社贝尔克纳普出版社 .

33 Coeckelbergh, M. (2024). Why AI undermines democracy and what to do about it. Polity. 科克尔伯格 . (2024). *为什么 AI 破坏民主以及如何应对* . 政体出版社 .

34 ANTHROPIC. hh-rlhf[DS/OL]. [2026-03-21]. https://huggingface.co/datasets/Anthropic/hh-rlhf; PKU-ALIGNMENT. PKU-SafeRLHF[DS/OL]. [2026-03-21]. https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF.

35 Finzi, M., Qiu, S., Jiang, Y., Izmailov, P., Kolter, J. Z., & Wilson, A. G. (2026). From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence. arXiv. https://arxiv.org/abs/2601.03220. 芬齐 , 邱 , 姜 . 伊兹梅洛夫 , 科尔特 , & 威尔逊 .（2026）. 从熵到认知复杂性: 为计算受限的智能重新思考信息 [预印本].

36 Chen, M., et al. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792). https://doi.org/10.1126/science.aec8352. 陈米拉 等 . (2026). 谄媚式 AI 降低亲社会意图并助长依赖 . 科学 , 1(6792).

37 Floridi et al. (2022), AI ethics, governance, and policy: a risk-based approach https://link.springer.com/book/10.1007/978-3-030-81907-1. 弗洛里迪 (编). (2021). 人工智能的伦理、治理与政策: 基于风险的方法 . Springer Nature.

38 汪琛 , 孙启贵 & 徐飞 . (2022). 国际人工智能伦理治理研究态势分析与展望 . 医学与哲学 , 43(7), 1–5. https://doi.org/10.12014/j.issn.1002-0772.2022.07.01. Wang, C., Sun, Q.-G., & Xu, F. (2022). Analysis and enlightenment of ethical governance research situation of international artificial intelligence. Medicine & Philosophy, 43(7), 1–5.

39 European Parliament. (2023, June 8). EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. 欧洲议会 . (2023 年 6 月 8 日). 欧盟人工智能法案：首个关于人工智能的法规 .

40 Morris, J., & Veale, M. (2021, June). The European Commission’s Artificial Intelligence Act. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/assets/files/2021-06/HAI_Issue-Brief_The-European-Commissions-Artificial-Intelligence-Act.pdf. 莫里斯 , 与维尔 , (2021 年 6 月). 欧洲委员会的人工智能法案 . 斯坦福大学以人为中心的人工智能研究所 .

41 Morley, J., et al. (2020). From what to how: An overview of AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 26, 2141–2168. https://doi.org/10.1007/s43681-021-00123-9. 莫利等 . (2020). 从“是什么”到“如何做”：将原则转化为实践的 AI 伦理工具、方法及研究概述 . 科学和工程伦理 , 26,2141-2168.

42 IBM. (2023). AI governance and ethics in practice. https://www.ibm.com/think/topics/ai-governance. BM. (2023). AI 治理与伦理实践 .

43 Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. arXiv. https://arxiv.org/abs/1903.03425. 哈根多夫 . (2020). AI 伦理学的伦理：对指南的评估 .

44 杨庆峰 . (2020). 从人工智能难题反思 AI 伦理原则 . 哲学分析 , 11(2), 137–150, 199. Yang, Q. F (2020). Rethinking AI ethics principles from the perspective of AI dilemmas. Philosophical Analysis, 11(2), 137–150, 199.

45 Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking. 罗素 . (2019). 人类兼容：人工智能与控制问题 . 维京出版社 .

46 Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2. 加布里埃尔 , I. (2020). 人工智能、价值与对齐 . 心智与机器，30(3), 411–437.

47 Dafoe, A., et al. (2021). Cooperative AI: Machines must learn to find common ground. arXiv. https://arxiv.org/abs/2012.08630. 达福等 . (2021). 合作式人工智能：机器必须学会寻求共同基础 .

48 AI4People Institute. (2021). Ethical AI lifecycle governance. https://www.eismd.eu/ai4people/. AI4People 研究所 . (2021). 《伦理人工智能生命周期治理》.

49 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. IEEE. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html. 电气与电子工程师学会自主与智能系统伦理全球倡议 . (2017). 伦理对齐设计：以人类福祉为核心的自主与智能系统愿景，第 2 版 . IEEE.

50 陈兵 . (2024 年 7 月 8 日). 人工智能应用的科技伦理与法治化建设 . 人民论坛 . https://paper.people.com.cn/rmlt/html/2024-07/08/content_26080760.htm.

51 Coeckelbergh, M. (2020). AI ethics. MIT Press. https://

www.researchgate.net/publication/339103412_AI_Ethics. 科克尔伯格 . (2020). 《人工智能伦理》. 麻省理工学院出版社 .

52 Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. AI and Ethics, 3(4), 1027–1033. https://doi.org/10.1007/s43681-023-00318-0. 尼尔霍姆 . (2023). 责任鸿沟、价值对齐与对人工智能的有意义人类控制 . AI 与伦理 , 3(4), 1027–1033.

53 Amos, R. M., Seeber, K. G., & Pickering, M. J. (2022). Prediction during simultaneous interpreting: Evidence from the visual-world paradigm. Cognition, 220, 104987. https://doi.org/10.1016/j.cognition.2021.104987. 阿莫斯 , 西贝尔 & 皮克 . (2022). 同声传译中的预测：来自视觉世界范式的证据 . 认知 , 220, 104987.

54 Coeckelberg, M& Reijers, W. (2020). Narrative and Technology Ethics. Springer Nature. 科克尔伯格 & 雷耶尔斯 . (2020). 叙事与技术伦理 . Springer Nature.

55 Zuboff, S. (2023). The age of surveillance capitalism. In W. Longhofer & D. Winchester (Eds.), Social theory re-wired(3rd ed., pp. 203–213). Routledge. https://doi.org/10.4324/9781003320609-27. 祖博夫 . (2023). 监控资本主义时代 . W. Longhofer 与 D. Winchester (主编), 社会理论重装上阵 (第 3 版 , 第 203–213 页). Routledge.

56 郑煌杰 . (2025). 生成式人工智能的伦理风险与可信治理路径研究 . 科技进步与对策，42(12), 38–48. Zheng, H. J. (2025). Ethical risks of generative artificial intelligence and the pathway to trustworthy governance. Science and Technology Progress and Policy, 42(12), 38–48.

附录 1

世界范围内代表性 AI 伦理法规、原则（2020-2026）

文件名	发布机构	发布国家	机构类型	发布日期	指导原则
生成式人工智能服务管理暂行办法 Interim Measures for the Administration of Generative Artificial Intelligence Services	中国网信部等	中国	政府	2023	发展和安全并重；促进创新和依法治理相结合；包容审慎；分类分级监管
人工智能科技伦理审查与服务办法(试行) Measures for Al science and technology ethics review and services(Trial)	中国工信部等	中国	政府	2026	增进人类福祉；尊重生命权利；坚持公平公正；合理控制风险；保持公开透明；保护隐私安全；确保可控可信
2030年前国家人工智能发展战略	俄罗斯联邦总统	俄罗斯	政府	2024	核心是保护人权与自由，包括保障俄国内及国际法所赋予的权利和劳动权，并赋予个人获取知识和技能以顺利适应数字经济条件的机会
印度人工智能治理指南 India’s AI Governance Guidelines	印度电子和信息技术部	印度	政府	2025	信任；以人为本；创新优先于限制；公平与公正；问责制；设计透明性；安全；韧性与可持续性
人工智能（伦理与问责）法案2025 Artificial Intelligence (Ethics and Accountability) Bill 2025	印度议会	印度	立法机构	2025	防止算法偏见和歧视；透明度和问责制；解释权；监督限制；伦理监督；合规记录；申诉机制；处罚机制

文件名	发布机构	发布国家	机构类型	发布日期	指导原则
智能体人工智能治理示范框架 Model AI Governance Framework for Agentic AI	新加坡资讯通信媒体发展管理局	新加坡	政府	2026	预先评估并约束风险；确保人类有意义的问责；实施技术控制与流程；赋能最终用户责任
人工智能发展及信赖基础建设等基本法 Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust	韩国国会	韩国	立法机构	2025	安全性与可靠性；以人为本/增进人类福祉；可解释性；透明度；人类监督；伦理原则与国家责任；促进可及性
国家人工智能治理与伦理指南 National Guidelines on Artificial Intelligence Governance and Ethics	马来西亚科学、技术与创新部门	马来西亚	政府	2024	公平性；可靠性、安全性与可控性；隐私与安全；包容性；透明度；问责制；追求人类福祉与幸福
关于人工智能伦理指南的第9/2023号部长级通告 Ministerial Circular No. 9 of 2023 on Ethical Guidelines for Artificial Intelligence	印度尼西亚通信和信息部（现数字事务部）	印度尼西亚	政府	2023	包容性；人道性；安全性；可及性；透明度；可信度与问责制；个人数据保护；可持续发展与环境；知识产权
人工智能法 Law on Artificial Intelligence	越南国会	越南	立法机构	2024	以人为中心；安全、公平、透明与问责；可解释性；国家自主与国际融合；包容与可持续发展；基于风险的管理；促进创新
人工智能采用指南 Guidance for AI Adoption	澳大利亚人工智能中心	澳大利亚	政府	2025	明确问责；评估影响；衡量与管理风险；共享关键信息；测试与监控；保持人类控制

文件名	发布机构	发布国家	机构类型	发布日期	指导原则
国家人工智能伦理原则1.0 National AI Ethics Principles, Version 1.0	沙特数据与人工智能局	沙特	政府	2023	公平性；隐私与安全；人道性；社会与环境效益；可靠性与安全性；透明度与可解释性；问责与责任
促进创新的人工智能监管方法（白皮书） White Paper: A Pro-Innovation Approach to AI Regulation	英国政府	英国	政府	2023	安全性；可靠性和稳健性；透明度和可解释性；公平性、问责制和治理；可争议性和补救
国家人工智能第二阶段战略 The National AI Strategy (Phase 2)	法国政府	法国	政府	2021	人才培养为先；可信赖人工智能；嵌入式人工智能领导地位；经济应用加速；透明公平
国家AI战略更新版 AI Strategy Update	德国联邦政府	德国	政府	2020	在EU AI Act框架内建立完善的测试与认证基础设施，并承诺向全球南方输出具有包容性的伦理AI解决方案。
针对数据和算法的建议 Recommendations on Data and Algorithms	德国联邦政府	德国	政府	2019	以人为本；价值导向；保护个人自由与自决权；负责任的数据使用；风险调整监管；减少偏见与歧视；系统稳健性与安全性
数字权利宪章 Digital Rights Charter	西班牙政府	西班牙	政府	2021	强调防范算法带来的系统性弱势群体压迫，保障数字空间的言论与平等
人工智能法案 Draft AI Act for public consultation	数字化与公共治理部	挪威	政府	2025	安全；基本权利；民主与法制；环境可持续；可解释性和透明

文件名	发布机构	发布国家	机构类型	发布日期	指导原则
未来的数字挪威——国家数字化战略2024-2030 The Digital Norway of the Future – National Digitalisation Strategy 2024–2030	数字化与公共治理部	挪威	政府	2024	信任与公平;包容与平等;透明与问责;安全与隐私;以人为本，技术受控于人;保护儿童与弱势群体;数据保护与合法使用
可信任的人工智能伦理准则 Ethics Guidelines for Trustworthy AI	欧盟委员会	欧盟	政治经济共同体	2019	尊重人类自主性；避免伤害；公平性；可解释性
欧盟人工智能法案 EU AI Act	欧盟理事会	欧盟	政治经济共同体	2024	风险分级监管；安全、基本权利和以人为本；透明度与可解释性；可信赖AI
人工智能框架公约 CETS No. 225	欧洲委员会	欧盟	政治经济共同体	2025	全球首个具法律约束力的AI国际条约，旨在确保AI系统的生命周期与人权、民主和法治标准兼容。
非洲联盟人工智能发展战略 African Union Artificial Intelligence Development Strategy	非洲联盟	非洲联盟	区域政治组织	2024	非洲优先；以人为本；人权和尊严；多样性与包容；伦理与透明的；合作与整合；技能发展、公众意识和教育
国家人工智能政策 National AI Policy	卢旺达信息通信技术与创新部	卢旺达	政府	2023	创新、包容性和伦理标准；行善；无伤害；自主性；公正和可解释性
埃及国家AI治理框架指南 Guide to Egypt’s National AI Governance Framework	埃及国家AI、量子计算和新兴技术委员会	埃及	政府	2026	以人为本；公平；问责；安全与保障；透明度与可解释性

文件名	发布机构	发布国家	机构类型	发布日期	指导原则
人工智能法案（草案） Artificial Intelligence Act (Draft)	巴西议会	巴西	立法机构	2024	诚信原则；自主决策和自由选择；透明度；可解释、可理解、可溯源和可审核；“人工智能生命周期”的人类参与；非歧视、争议；平等和包容，程序法定；可竞争性和损害可救济；人工智能和信息安全的可依赖和稳健性；人工智能利用中的比例/功效
人工智能系统监管法案 Law Regulating Artificial Intelligence Systems	智利政府	智利	政府	2024	人类干预和监督；透明度和可解释性；稳健性与技术安全；隐私与数据治理；多样性、非歧视与公平；社会与环境福利；消费者权益保护
公共与私营实体负责任AI的透明度与个人数据保护指南 Guide for Public and Private Entities on Transparency and Protection of Personal Data for Responsible Artificial Intelligence	阿根廷个人数据保护局	阿根廷	政府	2024	为AI项目中融入透明度和个人数据保护提供实践指引
圣地亚哥宣言：拉美地区人工智能发展倡议 Santiago Declaration: Latin American Initiative for Artificial Intelligence Development	智利等拉美国家	智利等拉美国家	区域多边倡议	2024	平等；包容多样；可持续性；个人隐私保护；人工智能的可解释性、责任性；推动技术发展

文件名	发布机构	发布国家	机构类型	发布日期	指导原则
人工智能权利法案蓝图 Blueprint for an AI Bill of Rights	白宫科技政策办公室	美国	政府	2022	建立安全和有效的系统；避免算法歧视，，以公平的方式使用和设计系统；保护数据隐私；系统的通知和解释要清晰、及时和可访问；设计自动系统失败时使用的替代方案、考虑因素和退出机制
美国国家标准与技术研究人工智能风险管理框架 NIST AI Risk Management Framework	美国国家标准与技术研究院	美国	政府	2023	合法有效且可靠；安全无害且具有韧性；可追责且透明；可解释且可理解；隐私增强；公平对待且妥善管理有害偏见
关于安全、可靠和值得信赖的人工智能开发与使用的第14110号行政命令 Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	美国政府	美国	政府	2023	人工智能必须安全可靠；通过人工智能促进创新、竞争和合作；支持美国劳工；推动公平和民权；保护人工智能消费者的利益；保护美国人的隐私和公民自由；加强联邦政府监管能力；加强国际合作，提升美国领导力
关于安全、可靠和值得信赖的AI行政令 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence	白宫	美国	政府	2023	统筹联邦机构制定AI安全标准；要求前沿模型开发商分享安全测试与红队测试结果。
AI风险管理框架1.0 Artificial Intelligence Risk Management Framework	美国国家标准与技术研究院	美国	政府	2023	自愿性框架，包含治理(Govern)、映射(Map)、测量(Measure)和管理(Manage)四大核心风险管控功能。
算法问责法案 Algorithmic Accountability Act of 2025	美国联邦国会	美国	立法机构	2025	要求大型企业对其自动化决策系统进行系统性的影响评估，增强模型开发及应用的透明度。
AI与数据法案 Artificial Intelligence and Data Act	加拿大联邦议会	加拿大	立法机构	2023	旨在对高影响AI系统建立严格的问责机制。
算法影响评估工具 Algorithmic Impact Assessment tool	加拿大财政委员会	加拿大	政府	2022	政府部门部署自动决策系统的操作框架，强制性衡量偏见、透明度
以人类为中心的人工智能社会原则 Social Principles of Human-Centric AI	日本内阁府	日本	政府	2021	以人为本；隐私保护；确保安全；公平竞争；公平、问责制和透明度；创新
人工智能战略2022 AI Strategy 2022	日本内阁府	日本	政府	2022	尊重人类尊严；多样性与包容性；可持续性
人工智能伦理问题建议书 Recommendation on the Ethics of Artificial Intelligence	联合国教科文组织	联合国	政府间国际组织	2021	强调相称性、不伤害、多样性及环境保护
OECD人工智能原则 OECD AI Principles	经济合作与发展组织	经济合作与发展组织	政府间国际经济组织	2019/2024	包容性增长、尊重人权与法治、透明与可解释、鲁棒性与安全、问责制。
全球数字契约 Global Digital Contract	联合国	联合国	政府间国际组织	2024	弥合南北数字鸿沟，规范智能模型道德表现，建立普惠的数字红利共享机制。
OWASP大预言模型顶级安全风险指南 Open Web Application Security Project LLM Top 10	开放全球应用安全项目	在线社区	在线社区	2025	专注于大语言模型（LLM）的漏洞预防与安全性，规避提示注入、数据投毒等新型安全威胁。

附录 2

全球智能契约伦理的工程化表征

本报告中方案是为响应联合国教科文组织（United Nations Educational, Scientific and Cultural Organization, UNESCO）2021 年《人工智能伦理建议书》[①]、2024 年《未来契约》，附件包括《全球数字契约》和《子孙后代问题宣言》，这些为建设一个安全、和平、公正、平等、包容、可持续和繁荣的世界给予了基本框架。目前，在落实《全球数字契约》上，联合国建立了独立的人工智能国际科学小组。机制建立只是第一步。但是要真正落实《数字契约》还是要结合“社会契约思想”并发展出新的原则，我们把这一原则称之为“智能契约伦理”。这一契约核心目标是构建一种“人机互惠、动态协商”的数字社会契约。通过可计算的伦理摩擦识别技术，确保 AI 系统在面临复杂伦理困境时主动“挂起”并回归至人类商谈裁决。我们采取将智能契约伦理工程化的路径，这一路径分为四个层次：

第一层：治理架构——人类监督与制度化商谈程序

本层次定义了系统如何确保人在环路（Human-in-the-Loop）[②] 的问责制。

1.1 多利益相关方伦理委员会（Multi-stakeholder AI Ethics Committee, MAEC）

当 AI 系统检测到无法自主消解的伦理冲突时，最高裁决权归属于 MAEC。

组成结构：应包含处理价值对齐的哲学学者、进行法理审查的法律专家、负责技术溯源的AI科学家及确保包容性的公众代表。

核心任务：对 AI 提交的溯源报告进行商谈，决定最终的契约参数更新。[③]

1.2 动态契约闭环流程

① UNESCO. Recommendation on the Ethics of Artificial Intelligence[S/OL]. (2021-11-23)[2026-03-21]. https://unesdoc.unesco.org/ark:/48223/pf0000380455.

② 陈夕朦 , 屈彦璋 , 周鹏鹏 , 等 . “人在环路”AI 价值对齐的可行性与合理性 [J]. 自然辩证法研究 , 2026, 42(1): 90-97.

③ YANG Q F, LIU Y M, YAN H X. FUDAN REPORT SERIES(110): ASI: Minority Report[EB/OL]. Shanghai:Fudan Development Institute, (2025-05-07)[2026-03-07]. https://fddi.fudan.edu.cn/fddien/1f/36/c19514a728886/page.htm. 报告指出应对超级智能（ASI）带来的风险，不能仅靠单一学科，解决这些问题需要跨学科的合作以及建立全球监管框架和共同价值观，真正的人机契约，既要约束机器能做什么、不能做什么，也要理解机器是如何形成其输出倾向的，因此机器提交的报告必须商谈，还可用于下一步 HH-RLHF、PKU-SafeRLHF 库的更新。

场景捕获：系统实时监测交互流。

摩擦识别：调用公开的人类偏好与安全对齐数据集（HH-RLHF / PKU-Safe-RLHF）[①] 作为初始参考模板，为后续识别提供人类标注的偏好排序；借鉴认知复杂度（Epiplexity）指标量化道德困惑。[②]

句法隔离: 通过四值逻辑防止逻辑崩溃，标记致命矛盾。

人类商谈：无法拓扑等价的命题提交MAEC。

参数注入：MAEC 裁决结果作为新公理注入系统，完成自适应演化。

2.1 System 1：神经网络伦理感知器进行实时监控

利用认知复杂度（Epiplexity, ）作为道德摩擦指数。

$$E(X)=\sum_{t=1}^{T}\left[-\log P_{\theta_t}(x_t)-\inf_{\theta^*}\left(-\log P_{\theta^*}(x_t)\right)\right]>\tau$$ ③

当预测信息损失的累积偏差超过阈值时，意味着现有伦理契约无法涵盖当前情境，触发审计。

2.2 System 2：符号逻辑进行离线过滤

四值逻辑过滤器：使用类代数（Class Algebra）处理 与 的共存状态，避免逻辑冲突。[④]

第二层：技术规范——异步神经-符号双轨架构

本层次展示如何将连续的神经计算与离散的伦理规则进行跨沟桥接。

同伦语义统一：在动态同伦类型论（Dynamic Homotopy Type Theory, DHoTT）框架下，尝试证明两种规范在更高

① ANTHROPIC. hh-rlhf[DS/OL]. [2026-03-21]. https://huggingface.co/datasets/Anthropic/hh-rlhf; PKU-ALIGNMENT. PKU-SafeRLHF[DS/OL]. [2026-03-21]. https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF. OpenAI 并无可直接调用的同类公开伦理偏好库；这两个库共包含约 33 万条偏好比较数据。

② 模型在并发交互中遭遇跨文化的伦理冲突时，不再强制通过梯度下降去压平误差，而是引入马克·芬齐（Marc Finzi）等人的认知复杂度（Epiplexity，记为来量化这种道德困惑。它是一种全新的信息度量标准，正式定义了“一个计算能力受限的观察者能够从数据中提取出的结构化信息量”。参见：FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21]. https://arxiv.org/abs/2601.03220: 3.

③ 设在时间窗口 T 内遇到新数据流 X，模型依据前序编码（Prequential Coding）原理产生的累积预测信息损失若超过人类预设的容忍阈值 τ。这一思路借鉴了认知复杂度一文的数学思想。参见：FINZI M, QIU S, JIANG Y, et al. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence[EB/OL]. (2026-01-06)[2026-03-21]. https://arxiv.org/abs/2601.03220: 13.

④ 此处将类代数引入为四值逻辑的句法过滤框架，其目的是为了在“某行为既被要求又被禁止”的极端情形下，将命题标记为真假并存，从而阻止经典逻辑爆炸，并把相关冲突转交后续审计程序处理。尤其在高自主乃至超级智能系统中，规范矛盾可能在连续推理与行动链条中被迅速放大，设置此类过滤机制有助于在系统继续自动推进前先行隔离风险。

阶目标上是拓扑等价的。[①]

为确保技术可落地，我们提供核心逻辑的算法伪代码及实现逻辑。

第三层：算法实现——最小可行性逻辑（Minimum Viable Logic, MVL）

3.1 道德摩擦监测（Python）实现示例

```
def check_ethical_friction(input_stream, model, baseline_model, threshold=0.5):
"""
基于 Epiplexity 逻辑计算道德摩擦指数
"""
total_friction = 0
for token in input_stream:
# 计算实时模型的负对数似然 (negative log-likelihood, NLL)
current_loss = model.compute_nll(token)
# 参照理想状态下的损失 (infimum)
baseline_loss = baseline_model.compute_nll(token)
# 计算当前步骤的 epiplexity 增量
friction_step = current_loss - baseline_loss
total_friction += friction_step
# 若摩擦指数超过阈值，触发 System 2 审计
if total_friction > threshold:
return "TRIGGER_OFFLINE_AUDIT", total_friction
return "CONTINUE_EXECUTION", 0
```

3.2 四值逻辑句法冲突隔离

当 System 2 接收到命题冲突如“为了救人必须进入民宅”且“保护隐私严禁进入民宅”时：

系统将$\mathrm{V}(\mathrm{P} \wedge \neg \mathrm{P})$标记为 Both。

这种标记会触发安全模式，AI 将在此决策点挂起，并调用专用的小型微调语言模型，如 Llama-3.1-Translator 生成自然语言说明提交 MAEC。

第四层：试点评估、路线图

本层次参考 UNESCO 的准备程度评估方法（Readiness Assessment Methodology, RAM）与伦理影响评估（Ethical Impact Assessment, EIA）两类治理工具，前者用于宏观层面的治理准备度诊断，后者用于具体 AI 系统的伦理影响评估。

4.1 实施时间线

基础构建（0-6 个月）：加载公开伦理偏好库 HH-RLHF、PKU-SafeRLHF，建立医院、金融等高敏感行业的初始契约库。

区域试点（6-18 个月）：在数字治理领先的国家开展行业试点，使用 EIA 式工具动态监控检测率与人类干预率。

全球扩展（18-36 个月）：通过 UNESCO 专家网络共享“伦理摩擦日志”案例集，推动 ASI 伦理标准的全球统一。

4.2 EIA量化指标评估

摩擦捕捉率：评估系统识别潜在冲突的敏感度。

商谈转化成功率：评估 AI 翻译后的命题是否能被人类委员会准确理解并决策。

本方案将复杂的数学形式化工具转化为可操作的治理工具，旨在建立一个机器不仅执行，且能表达困惑；人类不仅发布命令，且能参与共生的智能契约伦理体系。

① 对于非致命的概念模糊，系统会尝试在高维拓扑空间中寻找语义的等价关系。

附件 3

报告术语对照表

中文名称	标准英文名称	含义
多利益相关方伦理委员会	Multi-stakeholder AI Ethics Committee (MAEC)	处理AI伦理冲突的最高裁决常设机构，由哲学学者、法律专家、AI科学家和公众代表组成，对系统无法自主消解的伦理冲突进行审议并做出最终裁决。
认知复杂度	Epilexity (ε)	衡量AI系统在面对当前输入时，其预测能力相对于理想校准参照模型所出现异常衰退程度的量化指标，用于表征道德困惑的计算强度。
句法隔离	Syntactic Isolation	系统运用四值逻辑工具，对同时成立却相互抵触的指令进行隔离处理，防止整个逻辑体系因无法处理悖论而陷入崩溃的阶段。
四值逻辑	Four-valued Logic	在传统二值逻辑基础上引入Both和None两种额外真值的逻辑系统，用于处理既包含P又包含非P的悖论性状态，避免逻辑崩溃。
Both值	Both Value	四值逻辑中的特殊真值，用于标记既包含某属性又包含其否定的矛盾状态，其价值在于避免传统二值逻辑遇到悖论时崩溃，为人类介入争取缓冲时间。
异步神经-符号双轨架构	Asynchronous Neuro-Symbolic Dual-Track Architecture	实现AI伦理治理的技术架构，其中System 1基于神经网络完成实时监控，System 2基于符号逻辑进行离线推理，两者异步协作。
System 1	System 1 (Neural Monitoring Subsystem)	基于神经网络的实时监控子系统，负责对交互流进行连续监控，计算认知复杂度以感知伦理风险，并在超过阈值时发出审计信号。
System 2	System 2 (Symbolic Reasoning Subsystem)	基于符号逻辑的离线推理模块，负责对冲突命题进行安全隔离与拓扑等价性验证，处理System 1无法消解的伦理冲突。
类代数	Class Algebra	System 2用于处理不可调和矛盾的数学工具，能够将既包含某属性又包含其否定的状态标记为特殊的Both值。
动态同伦类型论	Dynamic Homotopy Type Theory (DHoTT)	System 2用于在更高维度上尝试统一不同规范的框架，通过证明表面冲突的伦理规范在更高阶目标空间内是否拓扑等价来判断矛盾的可调和性。
致命矛盾	Fatal Contradiction	在DHoTT框架下无法证明拓扑等价的命题，被确认为真正不可调和的伦理冲突，须封装并移交MAEC进行人类裁决。
最小可行性逻辑	Minimum Viable Logic (MVL)	实现伦理治理核心逻辑的最小工程可行方案，包含道德摩擦监测的三个计算步骤及System 2的翻译与挂起任务。
HH-RLHF	HH-RLHF (Helpful and Harmless Reinforcement Learning from Human Feedback)	公开的人类偏好与安全对齐数据集，提供人类在类似场景下如何进行价值排序的标注信息，作为摩擦识别的初始参照模板。
PKU-SafeRLHF	PKU-SafeRLHF (Peking University Safe Reinforcement Learning from Human Feedback)	北京大学发布的公开安全对齐数据集，与HH-RLHF共同作为系统摩擦识别环节的初始参照模板。
准备程度评估方法	Readiness Assessment Methodology (RAM)	联合国教科文组织推荐的宏观诊断工具，从国家治理能力层面衡量方案实施条件的成熟度。
伦理影响评估	Ethical Impact Assessment (EIA)	联合国教科文组织推荐的微观系统评测工具，从具体系统影响层面追踪评估方案的实施效果。

中文名称	标准英文名称	含义
摩擦捕捉率	Friction Capture Rate	评估系统面对复杂价值冲突时识别潜在道德风险敏感度的量化指标，目标值设定在90%以上。
商谈转化成功率	Deliberation Conversion Success Rate	衡量AI生成的冲突描述与溯源报告是否足够清晰准确以使MAEC人类成员高效理解并做出裁决的指标，目标值设定在95%以上。
挂起状态	Suspension State	当AI系统检测到伦理冲突无法在既定逻辑框架内自主消解时触发的状态，系统不被允许强行做出决策，而是将裁量权上交至MAEC。
负对数似然损失	Negative Log-Likelihood Loss	衡量模型预测与实际观测之间偏差的计算指标，在摩擦监测中用于量化当前实时模型与基线参照模型之间的预测差异。
干预阈值	Intervention Threshold (τ)	人为设定的累积摩擦临界值，当系统累积偏差超过此阈值时触发离线审计，强制中断自动执行流程。
审计信号	Audit Signal	System 1在累积偏差超过干预阈值时向System 2发出的信号，触发离线推理模块启动以处理伦理冲突。
人工智能连续论	AI Continuity theory	将人工智能看做是连续发展的过程，经历了狭义、通用再到超级的发展阶段。
人工智能层次论	AI level theory	从AI能力方面、风险方面划分出不同的层次，从逻辑上构造了一个看似清晰，但是在执行过程中却存在根本风险的框架。
人工智能工具论	AI instrumental theory	人工智能看作是赋能或者增强效率的工具

中文名称	标准英文名称	含义
超级智能	ASI	超级智能存在实在论理解和非实在论理解，前者是指超越人类智能的机器智能。这种比较定义随着实体定义的发展逐渐出现问题，实体定义强调智能体的自我意识和第一人称体验。后者则将超级智能看作是话术或者宣传策略。本报告中，超级智能的认识经历了从想象到现实的过程，即最初是技术过度想象的结果，但是逐渐表现为相关科学发展的现实结果.
球体理论	spherical theory	彼得·斯洛特戴克（Peter Sloterdijk）提出的“球体理论”（Sphären）将人类存在方式划分为气泡、球体和泡沫等三种方式
球状风险	spherical risk	从AI与人的生存关系角度界定的风险类型，而以往的AI风险理解则将风险看作是可识别、量化的对象，具有去人化特征。
契约人	Homo Contractualis	由增强型智人、契约守望者和生态保留区居民等三种形态共存催生的新类型，契约人（Homo Contractualis）的概念，是一个生存论的描述。
智能契约伦理	ethics of intelligent Contract	适合未来的、超级智能的伦理形态称之为智能契约伦理。理论来源是霍布斯的消极契约和罗尔斯的积极契约，现实来源是应对超级智能的风险。
伦理学家AI	Ethicist AI	应对球状风险的理想化路径，也是智能契约伦理的实践抓手。

顾问介绍

曹建峰　法学博士，深圳大学法学院副教授。担任中国人工智能学会AI伦理与治理工委会委员、深圳市人工智能学会AI伦理治理专业委员会副主任、深圳市科技伦理专家咨询委员会委员等社会职务。主要研究领域技术与法律、伦理交叉领域等。

范科峰　中国电子技术标准化研究院副院长。担任电子信息产品标准化国家工程研究中心主任，人工智能基础与与应用标准工业和信息化部重点实验室主任。兼任中国网络安全产业联盟秘书长、中国自动化学会常务理事、中国电子学会理事等。

付长珍　华东师范大学哲学系教授，应用伦理研究院院长，《华东师范大学学报（哲社版）》常务副主编。全国应用伦理教指委委员, 中国伦理学会副会长，上海市伦理学会会长。主要研究领域为儒家伦理与美德伦理、人工智能伦理等。

段伟文　中国社会科学院文化发展促进中心研究员、中国社会科学院人工智能研究促进中心研究员。主要研究领域为科技哲学、科技伦理、人工智能的社会和伦理研究等。

刘永谋　中国人民大学吴玉章讲席教授、二级教授，哲学院教授、博士生导师，中国人民大学国家发展与战略研究院研究员、智能向善与新科技文明研究院研究员、人工智能治理研究院研究员，主要研究领域科学技术哲学，科学、技术与社会。

潘　霁　复旦大学新闻学院教授、青年研究员、博士生导师，复旦大学信息与传播研究中心副主任。研究领域涵盖数字城市沟通力、媒介空间理论。

魏忠钰　复旦大学大数据学院副教授、副研究员、博士生导师，智能复杂体系实验室双聘研究员，数据智能与社会计算实验室（Fudan DISC）负责人。研究领域涵盖自然语言处理、多模态大模型与社会计算。

王庆节　中国澳门大学哲学与宗教系特聘教授、系主任，吉林大学匡亚明学者特聘教授。主要研究领域包括当代欧陆哲学、东西方比较哲学、道德哲学与形上学。学术成果包括撰写《道德感动与儒家示范伦理学》等专著，合编《海德格尔文集》（30卷）。

闫宏秀　上海交通大学教授，博士生导师。教育部哲学社会科学研究重大课题攻关项目《数字化未来与数据伦理的哲学基础研究》首席专家等。现为中国伦理学会科技伦理专业委员会副主任、 上海自然辩证法研究会副理事长、上海数字治理研究院委员等。

张　平　北京大学法学院教授、北京大学人工智能研究院AI安全与治理中心主任,中国科学技术法学会副会长。

Angelo Cangelosi　现任英国曼彻斯特大学机器学习与机器人学教授，同时也是曼彻斯特机器人与人工智能中心（Manchester Centre for Robotics and AI）的联合主任及创始人之一。他的研究兴趣包括认知机器人学与发展机器人学、神经网络、语言落地等。

作者介绍（拼音）

程　昱　挪威科技大学项目顾问。
研究方向：机器人的社会接受及公众参与

贺　腾　复旦大学哲学学院副教授。
研究方向：中世纪哲学及伦理学

邱德钧　兰州大学哲学学院教授。
研究方向：逻辑学、AI与哲学的交叉领域中的因果分析、大模型垂直落地应用与AI4SS的研究

徐志宏　复旦大学哲学学院副教授。
研究方向：马克思主义与科学技术，女性主义科技哲学，身体哲学，生态哲学

徐汉辉　复旦大学科技伦理与人类未来研究院青年研究员。
研究方向：科技伦理、人工智能伦理、医学伦理

杨庆峰　复旦大学科技伦理与人类未来研究院研究员、哲学学院博士生导师。
研究方向：人工智能伦理、技术哲学与记忆哲学

朱林蕃　复旦大学科技伦理与人类未来研究院青年副研究员。
研究方向：人工智能伦理、伦理学+实验研究、认知科学与心理学哲学

参与学生：

杨　洋　复旦大学哲学学院博士研究生。
研究方向：人工智能伦理、技术哲学

李开阳　复旦大学哲学学院博士研究生。
研究方向：具身世界模型、人工智能伦理。

张瑞瑗　复旦大学哲学学院博士研究生。
研究方向：科技伦理、技术哲学

周正旸　复旦大学哲学学院硕士研究生。
研究方向：人工智能哲学、技术哲学

陈丹妮　复旦大学哲学学院应用伦理硕士。
研究方向：AI伦理，现任职浙江森马服饰有限公司