分布式 AI 治理：技术制度博弈与社会政府协同

江小涓

摘　要　近十年来，人工智能治理问题是各国和全球科技治理的重点热点，然而相关治理共识虽然在各国和各种全球性高级别会议上反复发布，落地却缓慢而且收效有限。其原因是这些努力聚焦于形成以法律法规为基础、具有普适性的集中式治理模式上，然而这种模式主导的治理并不符合 AI 发展特点和社会各方意愿，缺乏有效的激励相容机制。本文提出 AI 分布式治理的概念，即以技术手段为互信保障，形成遵守共同规则的去中心化 AI 合规共同体，这种治理模式与 AI 技术逻辑以及社会意愿有较好的匹配性。文中着重分析了推动不同利益主体广泛加入分布式治理的激励和约束机制，包括社会价值对标、企业信誉维护、市场竞争压力、技术社区共识、技术保障能力等。今后，AI 治理要鼓励具有相同意愿与能力的群体形成规模与规则多元化的局部治理共同体，并由这些共同体构成遍及全社会的分布式治理网络，汇聚千万主体意愿和行动，积小善为大善。同时，基于公权力的集中式治理体系不可或缺，要对那些导致严重安全、伦理与价值观问题的 AI 行为进行有效治理。

关键词　AI 治理　分布式 AI 治理　AI 去中心化治理　AI 伦理与安全

作者江小涓，中国社会科学院大学教授（北京 102488）。

中图分类号 F49　　**文献标识码** A　　**文章编号** 0439-8041(2025)10-0005-9

有关人工智能治理问题的讨论，国内外长期聚焦在以法律法规为基础的集中式治理模式上，并做出许多努力，然而相关理念反复论述，具体举措却难以形成，实践落地见效缓慢。导致这种现象的本质原因，是监管者对个性化复杂场景的理解能力远低于具体场景中的组织和个体，技术和商业领域的发展激励机制远强于行政部门的治理激励机制，集中式治理制度形成周期远长于技术迭代周期。因此，在继续努力推动集中式治理体系建设的同时，要寻求更多能发挥各种类型治理积极性的路径。分布式治理是由多个共识联合体构成的治理网络，其架构具有非统一性和去中心化特点，具有高度的弹性、灵活性和多样性，能汇聚起众多差异化的 AI 向善行动，实现有限、有效和有用的治理。

一、理念复述与实践寡现：AI 集中式治理中的突出问题

AI 技术无疑为人类社会发展进步带来巨大益处。AI 是创新引擎，重构创新范式，AI 和数据驱动的创新成果大量涌现；AI 提升工作效率，推动产品、服务以及公共服务快速迭代，生产生活便利化程度极大提升；AI 能用多模态进行创作，多种形态的文化产品极大丰富，传播渠道极大扩充能够分发海量信息；AI 擅长复杂大场景规划，优化重构城市与社会运转流程，激活多方协作能力；AI 还具备深察与远望能力，助力人类对自然界和自身的理解边界极大扩展。AI 的这些能力和贡献社会已有高度共识，本文不再赘述。

人工智能在发展的同时，其治理问题也日益受到社会各界的广泛关注。人们逐渐意识到，AI不仅给人类带来诸多益处，也带来许多挑战。

（一）AI带来新挑战

1. 可以预见的挑战。

AI带来的挑战中，有些已经出现或者可以预见。

首先是对社会共识、社会团结和社会稳定的挑战。机器生成的信息增加，内容庞杂真伪难辨，由于合成信息大量加入和大模型产出的不可检验，主观有意和客观无能导致的“真—假—虚—实”内容交相纠缠在一起，当出现偏颇或虚假画像时，个体无法知晓问题出在何方。数智化的探测和匹配技术，决定着不同社会主体的可知不可知、可得不可得及可为不可为。例如针对个人偏好的新闻推送和广告投放，既能为消费者提供更合意更有个性的产品和服务，然而也很可能产生“信息茧房”效应，那些多元信息搜集能力较差的“文化弱者”受到的影响更大。AI还可能生成一些极端观点，有意无意带来社会不同群体立场极化的效应。它们还有能力通过掌握的海量数据和计算能力，侵犯个人隐私，违背社会公平，塑造不恰当的价值观等，甚至有能力干预政治选举，并在地缘政治中扮演重要角色。以谷歌公司为例，欧盟反垄断当局曾经认为，它并非市场中的一个竞争者，而是为他人设置竞争条款的权力核心和“造王者”（king-maker）。①

其次是对市场各方关系的挑战。对数据、算力和算法的掌控能力，使得大型企业和大型组织具有了更强市场竞争力和社会影响力，中小企业地位更加被动和弱化。大数据和特定算法一定程度上可以决定资源配置，资本/劳动的关系也在改变，被资本和技术替代的工作岗位类型在增多，各种人工智能不仅能够从事重复性高的信息处理类工作，生成式大模型更具智慧性，有望替代多种类型的知识和技能密集型工作岗位。如此下去，技术进步创造的新财富无法在市场相关各方特别是技术资本持有者与劳动者之间共享，导致经济发展的包容性共享性较差。

2. 尚未预见的挑战。

图灵奖得主Yoshua Bengio于2025年牵头发布的《国际人工智能安全报告》，将人工智能带来的各种风险划分为三大类，包括恶意使用风险（Malicious Use Risks），即人为使用人工智能系统进行伤害、操控、欺诈等非法或不道德行为；技术失灵风险（Malfunction Risks），即人工智能系统在正常使用情况下由于故障或技术失灵带来的不良后果；系统性风险（Systemic Risks），即人工智能大规模应用后可能引发广泛负面社会影响。② 有些情形下带来的风险极高，例如AI在军事领域的应用；有些系统性风险也会带来全局失控的问题，例如系统性金融风险。不过人们普遍认为，AI继续发展下去，人类社会面临的最大潜在风险就是技术失灵中的失控情况，亦即当人工智能（AI）发展到通用人工智能（AGI）或超级人工智能（ASI）的阶段，人类丧失对AI的控制权，机器不仅替代人而且战胜人。

上述分类并未穷尽AI的风险。经济学将风险区分为两类，一类称为风险（Risk），指未来事件的结果概率分布是已知或可以估算的情况，例如风险投资，失败率大致在89%—90%，还有保险赔付率等，这类风险可以通过概率模型量化，并能用保险、对冲等方式管理；另一类称为不确定性（Uncertainty），指在未来事件的概率分布完全未知或无法可靠估计的情况，例如突发罕见自然灾害或战争、颠覆性技术创新等，这类情况无法用传统概率统计衡量，决策往往依赖主观判断或经验，其极端案例现在也被称为“黑天鹅”事件。③ 科技发展思想史上有过著名的“科林格里奇困境”：试图控制一项技术是困难的，而且是几乎不可能的，因为在其早期阶段，当它可以被控制时，对其有害社会后果的了解不够充分，因此没有理由控制它的发展；但当这些有害后果显而易见时，控制已变得代价高昂和缓慢。④ 可以想见，未知概率的不确定风险以及“科林格

① A. 扎拉奇、M. 斯图克：《算法的陷阱：超级平台、算法垄断与场景欺骗》，余潇译，北京：中信出版集团，2018年。

② Bengio, Y., Mindermann, S., Privitera, D., et al., “International AI safety report,” arXiv: 2501.17805, Preprint. arXiv, 2025, https://doi.org/10.48550/arXiv.2501.17805.

③ 弗兰克·奈特：《风险、不确定性与利润》，安佳译，北京：商务印书馆，2010年；纳西姆·尼古拉斯·塔勒布：《黑天鹅：如何应对不可预知的未来的》，万丹、刘宁译，北京：中信出版社，2011年。

④ 关于“科林格里奇困境”的描述，可以参见李秋甫、张慧、李正风：《科技伦理治理的新型发展观探析》，《中国行政管理》2022年第3期。

里奇困境”在 AI 时代将更加突出和更具挑战性。

（二）集中式治理：有共识却难执行

国际社会关于人工智能风险与挑战话题的密集讨论已经将近十年，关注的重点是伦理、价值观与安全风险。各国政府、政府间国际会议、国际组织、行业组织、技术社群和头部企业等纷纷发声，努力寻求规则一致、普遍适用、有公权力加持的集中式治理体系。但这个体系在构建过程中呈现出共识容易形成但实践难以推进的特点。从治理原则的共识看，在国家层面和国际层面并无大的分歧，共识度较高。较早期有联合国教科文组织（UNESCO 2017）发布的《机器人伦理报告》（Report on Robotics Ethics），最新的是在 2025 年 2 月举行的巴黎 AI 峰会上 61 国签署的《关于发展包容、可持续的人工智能造福人类与地球的声明》。研究显示，从 2016 年到 2022 年 6 月，各类机构至少提出了 200 项以上的伦理原则和框架。图 1 展示了这些组织类型的分布情况，分别来自政府、非政府组织、非营利组织、学术机构、专业协会和企业等。

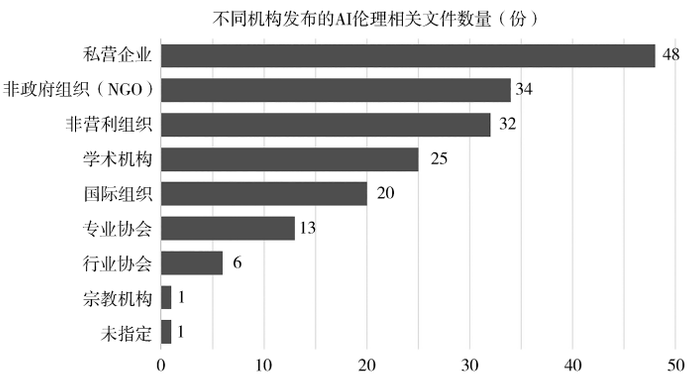

图 1　各类机构发布的 AI 治理相关文件数量

来源：Corrêa, Nicholas Kluge, et al.,“Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance,” 2022，文中图 3，按机构类型划分的出版物。https: //www. sciencedirect. com/science/article/pii/S2666389923002416.

各方面对数智时代技术规则和 AI 伦理相关问题的讨论共识度相当高，合并同类项后的治理原则包括：安全性、透明性、非歧视、可解释、可追溯、公平公正、包容开放、尊重隐私、共享利益、共同繁荣、以人为本、人类控制等，几乎所有政府、政府间和国际组织的倡议都包括其中多数项。[①] 例如安全性与可控性要确保系统在设计上能避免失控风险，再如公平性与非歧视性要防止人工智能在算法训练或部署过程中加剧社会不公等等。这些理念无疑是正当和令人向往的，业界和社会各方面认同度高，在近十年的时间中数百次被各类机构复述。在这些倡议类文件中，有多项是由国际组织和多国峰会发布的，有不少希望能够建构一种具有一致性普适性的集中式治理框架和治理体系。例如联合国秘书长安东尼奥·古特雷斯以及 OpenAI 的首席执行官山姆·奥尔特曼等在 2023 年 6 月举行的“人工智能全球治理”高级别公开会议上，提出设立一家类似于国际原子能机构（IAEA）的机构来监管人工智能。[②]

但是，各国政府以及政府间国际组织持续强调的这些原则与共识，在转化成治理行动方面进展缓慢，目前仍主要存在于理念层面和意愿层面，少数正在推进的落地努力也面临许多困难。欧盟《人工智能法案》的立法周期——从 2021 年 4 月欧盟委员会提出草案到 2024 年 8 月 1 日正式生效——历时约 3 年零 4 个月，然而这个花费三年多时间精心设计的监管框架，在 ChatGPT 横空出世后面临巨大新挑战。“我们原本以为 AI 就是

① 显示出这些共性的原则和报告很多，除了正文前面部分提及的之外，较有影响和代表性的还有：电气与电子工程师协会（IEEE 2016）发布的《合伦理设计：利用人工智能和自主系统（AI/AS）最大化人类福祉的愿景》（*Ethically Aligned Design*：*A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*），未来研究院（FLI　2017）发布的《阿西洛马人工智能原则》（*Asilomar AI Principles*），中国国家新一代人工智能治理专业委员会（2019）发布的《新一代人工智能治理原则》，联合国数字合作高级别小组（2019）发布的《数字相互依存的时代》（*The Age of Digital Interdependence*），还有许多，此处谨举上述几例。

② 2023 年 6 月，联合国秘书长古特雷斯召集了“人工智能全球治理”高级别公开会议，古特雷斯提议以国际原子能机构为蓝本，建立一个专门负责人工智能治理的国际机构。OpenAI CEO 山姆·奥尔特曼在此次会议上支持这一倡议。同年 10 月，古特雷斯宣布成立人工智能高级别咨询机构，由来自政府、企业和学术界的 39 名专家组成。

那些传统的应用场景，但ChatGPT改变了一切。”“治理体系的复杂性呈指数级增长，执行机构或许难以承担这样的制度负担”。[①] 韩国于2025年初通过的《基本人工智能法案》，是继欧盟AI法案后的全球第二部此类立法，现在也面临来自政界与业界的强烈修订压力。2025年7月29日，在由韩国执政党主办的一场会议中，多位专家表达了对该法案严格监管条款可能阻碍本土AI创新的担忧；韩国科技信息通信部也首次公开表示，随着欧盟自身对实施AI法规有较多担忧，韩国也可能调整执行时间表，并愿意倾听各方修订意见。[②] 至于联合国秘书长安东尼奥·古特雷斯等提出的设立专门机构的设想，两年过去了未见实际进展。

导致实践停滞的根本原因，是AI技术的指数级迭代，集中式治理进度难以跟上技术发展进度。世界各国涉及AI的制度安排都远慢于技术迭代的速度。同时，多国政府以及社会也担心治理会影响发展。信息广泛传播也带来制衡力量，技术发展和治理措施带来何种影响的信息可获得性极大增加，全社会更多了解正在发生的事情，所谓“无知之幕”支撑社会易于接受公共决策的状况不复存在。[③] 每一项决策涉及的损益各方都清楚知晓所获或所失，以全社会正和博弈或者正加总效应来论证该项规则的正当性或必要性，更不易被全社会成员特别是利益受到损害成员的接受。

本部分的分析表明，国家层面和国际社会建立集中式AI治理的努力还需继续推进，但短期看，期待这种努力取得显著效果为期过早。应对AI挑战还需要寻求更多的治理路径。本文下面部分将要提出，基于市场和社会自发努力的分布式治理架构能够取得广泛实效，值得重视和强化。

二、分布式治理：复杂博弈中的多元合规共同体

（一）分布式治理：去中心化合规共同体

人工智能是能量巨大的先进技术，其应用波及经济社会各个层面和条线，必然导致各方权益改变和关系重构，这个过程极为复杂，超出政府的理解和计算能力。市场就是应对复杂性而生的机制，千千万万个体和群体将复杂性拆解为交易双方的简单损益计算，当事人拥有做出正确判断和选择的具体信息。赫伯特·西蒙（Herbert A. Simon）是卡内基梅隆大学计算机科学和心理学系教授，1978年因其对“有限理性”（Bounded Rationality）的研究获得诺贝尔经济学奖。他认为人类决策受认知能力限制，无法完全处理复杂环境中的所有信息。而市场和企业都是应对复杂性的制度设计，企业通过层级制简化决策，市场通过价格机制协调分散知识，两者均是为了降低复杂性带来的不确定性。[④] 然而，市场交易和社会博弈常见的双方点对点沟通方式，并不适合许多情形下AI发展与治理的要求。AI技术是大场景技术，多方协作是常态，因此本文使用分布式治理这个概念。

分布式是计算机科学的核心概念（Distributed Computing/Systems），指将计算任务、数据存储或业务功能分散到多台独立计算机（节点）上协同完成的系统架构，其核心思想是多台机器协同工作，提高计算效率、可靠性和可扩展性。较常见的还有分布式能源体系（Distributed Energy System），与传统集中式能源系统（如大型火电厂）不同，它通过多节点、小规模、模块化的方式实现能源就近生产和梯级利用。还有分布式公共治理（Distributed Public Governance），其核心在于通过分散决策权，让公民、企业、社区组织等多元主体共同参与公共事务。这些以及更多分布式体系，其本质是构建分散的决策与运行模式，利用技术作为互信工具（如区块链、大数据、人工智能等）替代传统中心化机构的信用背书，具有更好的灵活性与弹性，每个节点的运行依据自身特点定义，局部故障或失灵不影响整个系统的运转。从这个描述就可以看出，AI治理特别适用于分布式体系来推进。

---

① 澎湃新闻·澎湃号·湃客：“世界人工智能大会上，全球人工智能的立法者们都在聊些什么？”2025-07-31，https://www.thepaper.cn/newsDetail_forward_31252653。

② Yoo, Choonsik, “South Korea AI Law under Pressure for Revision amid Growing Concerns,” MLex, 2025年7月29日, https://www.mlex.com/mlex/articles/2370338/south-korea-ai-law-under-pressure-for-revision-amid-growing-concerns。

③ “无知之幕”是美国哲学家罗尔斯的代表性学术观点，当人们设计或者评判一个社会的制度或规则时，他们不了解这个制度对自身可能产生的影响，好似被置于“无知之幕”的背后，从而在选择社会规则时仅从“普遍理性”出发，追求对所有人最公平的结果。参见J. 罗尔斯：《正义论》，何怀宏、何包钢、廖申白译，北京：中国社会科学出版社，1988年。

④ 赫伯特·西蒙：《管理行为》，詹正茂译，北京：机械工业出版社，2021年。

这里给出分布式 AI 治理的简单定义：以理念共识为合作基础，以技术手段为互信保障，形成遵守共同规则的去中心化 AI 合规共同体。这种分布式治理与 AI 技术逻辑一致，有利于技术特性与治理模式的耦合与加持。我们相信，由多个共识联合体组成的分布式治理网络，能够发挥灵活多样的治理功能，汇聚起众多差异化的 AI 向善行动，有效应对复杂多样的人工智能治理挑战。

这里我们借用“大教堂”和“集市”这两个概念形象描述下集中式治理与分布式治理的不同。“大教堂”与“集市”原本用于描述闭源和开源软件，这里借用这两个比喻表明集中式治理与分布式治理。“大教堂”（集中式治理）是自权威向大众下达规则的路径，由少数人（决策层）决策，制定有固定规则，有严格的执行程序和规则控制等要求；而“集市”（分布式治理）是多点平行交互博弈的路径，市场和社会主体依据自己利益和诉求，自主选择特定交易对象结成合作共同体，规则灵活，进出自由。每个“集市”（分布式治理空间）的规模并不大，但千万个“集市”的总和空间是巨大的，产生的治理动能也是巨大的。正如《大教堂与集市》一书中提到的“林纳斯定律”所言：“（开源软件）只要有足够多的眼睛，所有的 bug 都是浅显的。”[①] 这句话用来思考 AI 可能形成的多种安全与伦理挑战以及应对方式是很适宜的。显然，分布式治理是有限度的、不完美的，但基于各方博弈和共识形成，为多方协同合作奠定可信基础。

（二）分布式治理与技术合意性

与过往的科技治理相比，AI 治理有很大不同。当代许多科技发展是在寻求甚至构建自然界和人类社会演进中并不存在的状态，不少探索意在改变人类的“自然状态”或“社会状态”。例如 AI4S（AI for science）应用最密集的生命科学领域，有许多科技意图改变我们的生理、繁衍和认知结构。在这类科技方向上，每个人与科技专家应该有平等的知情权和发言权。笔者曾经强调过，“合理与合意”应该成为经济社会治理的具体目标，这里再强调下合意性问题。[②] 技术发展的“合意”是指全社会对某项技术发展“具有最大公约数的社会共识”，从定义就可看出，是否合意必须由公众来决定，只有各方能充分表达才能找到最大公约数。分布式治理允许和鼓励有共识者针对特定诉求结成大大小小的同盟，认为规则“合意者”才会共同构建治理合作体。显然，分布式治理能够推动 AI 发展向合意方向移动。

三、激励相容及分布式治理若干形态

（一）为何愿为和如何为之

分布式治理需要千千万万的行动者，激励来自何方？全社会对 AI 安全与伦理问题高度关注和自由表达，形成泛在、强大和持续的压力和导向；AI 企业发展需要海量用户，社会各方对其价值观和合规性的评价至关重要；企业和其他创新团队具备以技术促进向善及合规的能力，又具备自主决策和快速响应能力；监管者立法者的发声和行动所体现的方向性关注，对投资者、开发者、消费者以及公众心理都有重要引导和约束作用。这些都是分布式治理的激励约束机制。

1. 价值对标与主动作为。

人类价值观、社会意愿和企业内部向善文化形成合规建设强大推动力。AI 公司在“价值对齐”方面有大量行动，甚至“对齐”这个术语的普及和落地也是企业提出和实践的。[③] 许多公司和研究机构已经或准备制定明确的伦理和法律框架，指导大模型的开发和使用。这些框架通常包括数据隐私保护、内容审核标准等。OpenAI 在其 GPT 系列模型中引入了多种安全措施，如内容过滤和用户反馈机制，设立了“AI Safety and Alignment Team”，致力于研究和开发技术以确保 AI 系统符合人类价值观。Google 在 AI 伦理和透明度方面投入了大量资源，建立了专门的伦理委员会，并发布了多篇关于 AI 伦理和对齐的研究报告。阿里巴巴达摩院在 AI 伦理和对齐方面也进行了大量研究，定期发布报告，推出“AI+社会责任”计划。企业不仅各自努力，还

① 埃里克·史蒂文·雷蒙德：《大教堂与集市》，卫剑钒译，北京：机械工业出版社，2014 年。

② 江小涓：《数智时代的秩序重构与治理合作：合理合意双重目标》，《管理世界》2025 年第 5 期。

③ GPT 等大模型出现后，OpenAI 和 DeepMind 的研究者开始将“对齐”定义为确保模型目标与人类意图和价值观一致技术框架。2019 年，OpenAI 发布《*Alignment for Advanced AI Systems*》，明确将对齐列为关键研究方向。

基于共识产生了多种类型的AI治理企业共同体。2016年9月，谷歌、Facebook、亚马逊、IBM、微软联合成立AI合作伙伴组织（Partnership on AI），目标是推动AI技术符合伦理规范，重点关注公平性、透明性及社会责任。2025年7月16日，在中国人工智能产业发展联盟第十五次全会上，阿里巴巴、百度、火山引擎、零一万物、第四范式、科大讯飞、蚂蚁集团等16家企业代表，共同发布《人工智能安全承诺》实践成果，推动签约企业及更多企业遵守联盟共同规则。①

2. 信誉维护与回应关切。

企业、个人和组织对自身信誉的珍视推动企业积极回应社会质疑和主张。AI与核不扩散等机制不同，后者与市场及公民的日常并无直接关系，只与政府间军事行为有关。AI不同，生存发展时刻需要海量用户，声誉影响生死攸关。2016年，Twitter聊天机器人Tay因用户恶意输入，迅速学习并发布种族主义言论。Microsoft在24小时内下线Tay，并修订AI聊天机器人规则，增加实时内容过滤和用户行为黑名单。2023年，OpenAI因使用用户与ChatGPT的对话数据训练模型引发争议。用户和媒体质疑其隐私政策模糊，可能泄露敏感信息。OpenAI及时更新了数据使用政策，明确承诺默认禁用用户对话数据用于模型训练，除非用户主动开启“数据共享”选项，同时新增了数据删除请求通道，允许用户提出移除特定数据的要求。最新的案例是，2025年2月底至3月初，腾讯元宝新上线的《用户协议》规定，用户上传至平台的内容及AI生成的内容，腾讯将获得不可撤销、永久、免费、可分许可的使用权，此举引起用户强烈质疑。3月1日至5日，腾讯作了三次修改，承诺体验优化计划开关默认关闭。一周内从《用户协定》上线到修改完成，这个过程政府完全没有发声，或许还未来得及知晓和理解事件本身，全程就只是用户与AI企业的博弈过程，最终达成规则共识，腾讯元宝也成为活跃用户最多和应用最广的生成式AI大模型之一。由此可见用户信誉维护对AI企业行为的重要约束力。②

3. 技术社区与标准构建。

数智化技术快速迭代，各种技术社群在价值、伦理和安全理念的构建传播以及标准制定等方面发挥着重要作用，是AI国际治理领域的重要行动者。技术标准开发组织（SDOs）在AI规则和伦理标准的制定方面发挥了重要作用。例如电气和电子工程师协会（IEEE）2016年4月发起“自主与智能系统伦理的全球性倡议”，联合数百名跨学科专家研讨伦理准则，并于2016年12月发布首个《人工智能伦理设计准则》版本，强调AI需“造福人类”，要求系统设计融入透明性、可追溯性及人类监督机制，这是针对AI伦理问题较早讨论和行动的团体。再如国际标准化组织（ISO）制定颁布多项ISO体系中的AI标准，例如ISO/IEC 23894：AI风险管理指南；ISO/IEC 42001：AI管理系统（AIMS）标准等。国内也有诸多技术社群制定共识规则，如中国电子技术标准化研究院（CESI）发布的《人工智能标准化白皮书》，颁布的国家标准GB/T 35273-2020《个人信息安全规范》等，涉及AI可控性、公平性以及数据伦理等内容。

4. 技术保障与规则可行。

各种保障AI模型安全与伦理问题的分布式努力，技术可行是共识能够落地的重要保障。AI是治理对象，同时AI也是治理工具和手段。“技术治理技术”的逻辑是将法律要求直接嵌入技术系统，通过技术手段实现合规要求。谷歌、微软、IBM、阿里巴巴、腾讯等AI巨头都开发了算法审查工具来加强对技术风险的监控。第三方技术安全与内容合规审查企业也纷纷成立，在AI治理中发挥重要作用。以深度伪造检测为例，传统的治理手段在面对AI技术时往往失效，例如对于线上伪造证件问题，通过传统的照片或视频验证不易鉴别，只能通过大模型安全工具来解决。再如训练数据中的历史歧视被编码为系统规则（如招聘AI排斥女性），需引入对抗

① 中国信息通信研究院、中国电子信息产业发展研究院、清华大学、上海人工智能实验室：《中国人工智能安全承诺框架》，2025年7月26日。

② 2025年2月底至3月初，腾讯旗下AI助手“腾讯元宝”因用户协议中的版权条款引发用户广泛争议，腾讯元宝新上线的《用户协议》规定，用户上传至平台的内容及AI生成的内容，腾讯将获得不可撤销、永久、免费、可分许可的使用权，涵盖存储、修改、商业化等用途，此举引起用户很强质疑，网络上出现拒用元宝的声音。自3月1日起，腾讯作了三次修改，第一次删除了“不可撤销”“永久”等表述，并新增条款允许用户联系腾讯拒绝授权；3月4日做了第二次修改，明确体验优化计划开关默认关闭，明确声明“用户输入和输出的内容，权利归用户或相应权利人所有”；3月5日再次修改并发出官方致歉，承认旧协议“给用户造成困扰”，承诺默认状态下内容不用于模型优化，输入和输出的内容版权仍归用户。此处表述由笔者汇聚多方信息做出，关于最终协定版本与官方致歉声明，参见 https://finance.sina.com.cn/tech/roll/2025-03-05/doc-inenqtkm7792422.shtml。

性去偏见算法。智能合约则是基于区块链的自动化执行合约条款的计算机程序，条件具备即自动触发，可以支撑多种形态和用途的去中心化管理。在保证企业合规方面，AI 也是高效率工具，例如隐私政策生成工具可以快速生成符合法律要求的隐私政策文件。更重要的进步出现在 2023—2025 年间，许多大模型纷纷制定前沿模型安全框架，包括 Anthropic 的“负责任扩展政策（ResponsibleScalingPolicy，RSP）”2.2 版、OpenAI 的“应对准备框架（PreparednessFramework）”第 2 版、GoogleDeepMind 的“前沿安全框架（FrontierSafetyFramework）”2.0 版、Meta 的“以结果为导向的前沿 AI 框架（Outcomes-LedFrontierAIFramework）”以及 Amazon 的“前沿模型安全框架（FrontierModelSafetyFramework）”。尽管名称各异，这些安全框架有一个共同核心：在模型能力达到可能造成严重危害的门槛时，不得在缺乏充分安全保障的情况下贸然部署。这些框架体现了行业对前沿 AI 模型（frontierAImodels）在安全治理上的探索和承诺。①

5. 竞争压力与伙伴协同。

生成式 AI 依托平台发展，平台为相关各方交易立规则。平台虽然居于核心地位，但从自身利益出发，制定的规则要关注相关各方面诉求，兼顾各方利益，力争使有共识的生态圈最大化。同时，大模型之间存在竞争，即使最好的模型，潜在竞争者始终存在。这种可竞争性不仅限于本地，而是存在于网络可以触达之处。消费者转换 AI 使用的沉没成本几乎为零，只需手指点击几下即可从一个 AI 转到另一个 AI 上。实践中，大模型生态圈中的各方往往是多栖的，时刻在比较不同模型的优劣，因此每个模型要力争使自己的安全与伦理状况处于各方可接受状态，如果因为店大而欺客，就会丧失大量用户和供应商。为了增强社会影响力和感召力扩大朋友圈，许多大平台推出自主向善的 AI 应用。微软地球智能（AIforEarth）项目到 2021 年利用机器学习已支持了全球 120 多个国家和地区的 860 余个项目，涵盖气候变化、生物多样性保护、农业优化和水资源管理四大领域（MicrosoftCor）。蚂蚁森林是蚂蚁集团于 2016 年 8 月在支付宝平台推出的公益性环保服务，通过数字化手段将用户的日常低碳行为转化为生态保护行动，用户通过 40 余种低碳行为积累虚拟“绿色能量”并可申请在生态脆弱地区种下真实树木。截至 2025 年 4 月，累计种植真树超 6 亿棵。

6. 跨国运作与治理联盟。

AI 的规模经济与范围经济特别显著，边际收益很高，用户数至关重要。因此大模型都力求全球落地应用。然而，不同国家对 AI 合规性的要求不同，政府又难以及时明确评测 AI 合规状况，带来大模型跨国落地的困境。为此，跨国 AI 伦理与安全评测共同体成为重要应对工具。一类由国际技术组织牵头联合多个企业推进，国际上有由联合国旗下世界数字科学院（WDTA）发布的 WDTA AI STR 系列，以及由蚂蚁集团、清华大学、中国电信等联合谷歌、微软、OpenAI 等 20 余家机构制定的《AI 智能体运行安全测试标准》，融合多国监管要求和企业跨国运营诉求，将伦理要求拆解为可量化的测试项（如“记忆模块安全审计”）。共识度高的评测模型具有内在商业驱动力，企业通过认证体系降低跨国经营合规成本，如蚂蚁集团借 WDTA 标准通过欧盟 DGA 审计，据说有 34 个国家的 AI 项目需通过该认证方可部署。还有一类由有影响的科技机构牵头制定，例如上海人工智能实验室主导的 Open Compass 支持多语言（中英文）与多模态任务的综合评测框架，覆盖 50+数据集，提供开源可复现的评估流程，支持自定义评测模块。Open Compass 被阿里巴巴、百度、腾讯、华为等企业用于模型迭代优化，其开源性推动全球超过 200 家机构采用，评测结果在一定程度上成为行业通用基准。

（二）分布式治理的泛在性、回应性和长期性

分布式治理具有泛在性。AI 时代，技术发展及其对经济社会的影响变化快，何为合理合意的治理原则难以及时判断，集中化一致性治理推进艰难，分布式治理将发挥更加广泛而重要的作用。分布式治理的范围可大可小，在不同场景下有不同形态，能探索如何平衡各方利益，并及时调整方向和重点。分布式治理不会是理想的治理模式，也不可能是各方都满意的规则，但对共同体中各方来说至少是可接受的，因此有广泛的存在基础。

分布式治理具有回应性。AI 技术发展方向难以预测，许多潜在而未发生的挑战在实践中无法预判和有效应对；随着事态出现而产生的治理需求，问题明确利害可测，此时相关各方进行回应才具有可行性。比如近

① 案例来源可参见 https://mp.weixin.qq.com/s/E-nUO6Pc-FL2XT3brE6ONA。

几年ChatGPT在学术研究中广泛使用，许多“实事”和“数据”来源于机器编造或者称之为“幻觉”，导致无知误用和有意造假的案例和数据频发，因而一些学术规范严格的期刊建立了人工智能内容审查标准。同类问题在法律界也出现，有律师在辩护时引用了AI编造的“判例”，引致美国律师协会等明确要求，在法律实践中使用人工智能辅助时必须设立严格的验证机制。① 这种“见招拆招”的回应式治理看上去缺乏远见和预判，却有较好的针对性和实效性，而且治理成本较低。

分布式治理具有长期性。技术演进速度太快，冲击范围很大，行政和法律监管能力的适应和提升有时跟不上，集中式治理体系难以快速全面形成，分布式治理持续存在并不断迭代将发挥重要作用。分布式治理力度较弱，没有公权力赋予的强制力量，不过AI发展中出现的许多问题没有必要也不太可能用劲太猛以至完全消除，那样做会极大限制AI发展和应用，社会成本昂贵。例如大模型的不可解释性目前看来难以完全消除，合理的治理目标不是“消除”而是“限制”，使这种损害不会对经济社会产生严重和系统性的不良后果。在可预期的未来，每个场景或局部受到AI挑战的问题千差万别而且不断变化，分布式治理向集中式治理转变和集中式治理失效并生成新的分布式治理，很可能在时间与空间上始终并存。

四、集中式治理必须存在，构建底线和边界

分布式治理虽然广泛而有效，但基于公权力的集中式治理体系必不可少。市场与社会力量形成的分布式治理，并无公权力的强制力，特别是没有法律法规所赋予的严峻处罚权力，因此对于那些导致严重安全、伦理与价值观问题的恶意行为约束力不够。特别当引发的议题具有“军事”“民族”“国家”等属性时，自发的分布治理体系不能有效运作，主权国家和政府间国际组织需要发挥不可或缺的作用。公权力机构要以明智恰当的行权方式，向全社会表明治理并非创新的对立面，而是实现人工智能健康、有序、可持续发展过程中不可或缺的制度性支撑。②

（一）严防严惩带来严重后果的恶性使用

以挑战人类价值观、破坏社会团结甚至制造暴力和恐怖事件等为目标的AI恶性使用，对社会造成严重伤害，既要严防更要严惩。多国法律中明确规定AI严重违法行为清单包括侵犯用户隐私、知情不处理甚至共谋发布虚假信息、恐怖主义、仇恨言论等内容。中国相关规定包括不得生成煽动颠覆国家政权、推翻社会主义制度、危害国家安全和利益等内容。通常情形下，AI模型用途多元，可以造福人类但同时可以被用来作恶，因此严惩比严防更加重要。比如刀可以切菜但也可以有意伤人，汽车是交通工具但也可以有意撞人，由于严防即禁用的成本过高，严惩就成为“合意”选择。而对军事、金融等高风险应用领域，则应设定更严格的防控措施。

（二）强制要求公开透明支撑分布式治理的有效性

AI问题技术性隐蔽性强，只有开发者真正知道算法如何运行，不仅公众不易知晓其规则，外部技术人员和政府监管人员要理解技术架构也并非易事，即使存在安全与伦理问题，也有可能许久不被社会所知晓。因此规则的公开透明是基本要求，这不是可选项，而是AI治理的基础条件。要确保各方可以质疑平台涉及公众内容的算法审核，允许持权公共部门或特定的研究者获取平台相关数据以开展算法规则和在线内容的实时研判。

（三）保护AI弱势群体权益维护稳定与公平

AI带来的失业问题与以往不同，过去多轮技术革命对就业而言虽有替代但更有新创，目前看上去似乎难以认为AI时代这种乐观情形会重现，被AI替代的群体很可能成为技术进步中的利益受损者。合意性的一个重要诉求是技术进步的益处能被社会成员所分享，失业冲击处理不好会带来较大的社会问题，为此更符合时代要求的社会再分配体系和社会保障体系需要加快研究。③ 全社会合理分享AI红利有充足理由，AI发展建立

---

① 事件来源参见 https://paper.sciencenet.cn/htmlnews/2023/5/501752.shtm。

② 引自薛澜在2025中国数字经济发展和治理学术年会上的主旨演讲：《全球视野下的人工智能治理——挑战、机制与未来路径》，参见 https://mp.weixin.qq.com/s/Cn6XcqdX1jmDPADxQdR_Ag。

③ 全民基本收入是一个热门话题。这项政策旨在为所有公民提供定期、无条件的基本收入，确保每个人都能满足基本生活需求。有关情况可以从在线资源 Basic Income Earth Network (BIEN)(https://basicincome.org)网站上获得。

在全社会持续积累的基础上，仅从海量数据这个角度就能理解全社会为 AI 发展付出的知识积累，因此应该合理分享创新红利。

（四）政府发出信号引导 AI 发展方向和社会理念

严格意义上的行政和法律强监管往往具有滞后性，但政府以及立法机构对 AI 头部企业行为的实时关注，对严重违法违规案件的及时罚处，对一些损害公共利益行为的批评指责，发起相关监管政策和法律制定修改的公共讨论，发布建议、指南、最佳实践、优秀案例、专业标准等软性举措，都对创新方向、技术应用、投资选项、公众心理和社会舆情等有显著的导向作用。在中国传统社会中，“礼”这类社会共识虽然属于软规则，但并非单纯靠言传身教来推行，实际上也是通过一系列有形引导来助力实施，而且特别需要“官方”和社会的认可来传承。[①] 国外实践也表明，有效的社区自主治理需要政府和外部组织承认社区自定规则的权利，复杂问题也需要社区、区域和国家的协同。[②] 总之，公权力机构影响力和引导力强大，是集中式治理的主责方，也是分布式治理的重要力量。

简短的结语：AI 治理将是一个在技术与制度博弈中渐次推进的过程，政府或任何个人和组织都不能独自应对这些挑战。汇聚千万主体意愿和行动的分布式治理和代表社会核心关切的集中式治理要协同努力，集小善为大善，以大善引领小善，努力使 AI 为人类社会发展进步服务。

（责任编辑：沈 敏）

Distributed AI Governance: Technology System Gaming and Social Government Synergy

JIANG Xiaojuan

Abstract: Over the past decade, the issue of AI governancehas emerged as a focal and high-profile concern in both national and global technology governance. Despite the repeated release of related governance consensus at various national and high-level global forums, implementation has been sluggish and effectiveness limited. The reason is that these efforts have focused on forming a centralized governance model based on laws and regulations with universal applicability. However, such a model fails to align with the intrinsic characteristics of AI development and the diverse intentions of societal stakeholders. In other words, there is a lack of incentive compatibility mechanisms. This paper introduces the concept of distributed AI governance: a decentralized AI compliance community predicated on technological mechanisms to ensure mutual trust and adherence to shared rules. Such a governance paradigm aligns more coherently with both the logic of AI technology and societal preferences. The core of the analysis lies in examining the incentive and constraint mechanisms that promote the broad inclusion of heterogeneous stakeholders into distributed governance. These mechanisms encompass alignment with societal values, maintenance of corporate reputation, market competition pressures, consensus within technical communities, and the assurance capabilities of technological infrastructure. In the future, AI governance should encourage groups with the same willingness and ability to form local governance communities with diversified scales and rules, and these communities should constitute a distributed governance network throughout the whole society, converging the will and actions of millions of subjects, and accumulating small goodness into big goodness. Simultaneously, the indispensable role of centralized governance anchored in public authority must be preserved to effectively govern AI behaviors that provoke serious safety, ethical, or value-based concerns.

Key words: AI governance, distributed AI governance, AI decentralized governance, AI ethics and security

① 时延安：《软秩序与强秩序——对中国传统社会中礼刑关系的重新认识》，《朝阳法律评论》2010 年第 2 期。

② E. 奥斯特罗姆：《公共事务的治理之道：集体行动制度的演进》，余逊达、陈旭东译，上海：上海三联书店，2000 年。