人工智能全球治理态势观察

胡皓阳[1,2]，彭真[2]
1. 南京理工大学 网络空间安全学院
2. 腾讯 ESG 协同办公室

随着人工智能（AI）技术的快速发展，其带来的机遇与挑战日益显著，全球对 AI 治理的关注也逐渐加深。然而，各国在 AI 监管上存在差异，影响了国际合作的进程。美国偏向自由市场驱动，欧盟则侧重风险防控，中国注重技术安全与透明度，同时地缘政治因素也加剧了合作难度。企业层面，AI 公司以合规和信任为核心目标，致力于保障伦理与安全。与此同时，AI 治理的工程实践和技术工具逐步发展，针对安全、隐私、公平性等问题提供了解决方案。本文将分析全球 AI 治理的现状与挑战，探讨推动国际合作的策略，并深入探讨企业和技术工具在 AI 治理中的作用。

一、全球监管思路有所差异，地缘政治导致国际合作存在阻碍

全球 AI 监管存在显著差异，主要表现在监管理念、政策导向和实施机制上。地缘政治与经济竞争也加剧了这些差异，导致国际合作面临障碍，全球 AI 监管政策缺乏统一性。

（一）全球监管思路的差异性

1. 美国

美国的 AI 监管秉持自由主义理念，主张通过市场自我调节和技术创新推动 AI 发展，减少政府干预。2021 年发布的《人工智能美国国家战略》强调创新优先，避免过度监管。然而，监管机构也关注 AI 的潜在风险，特别是在隐私保护和技术滥用方面。比如，OpenAI 在开发 ChatGPT 时需要遵循美国联邦贸易委员会（FTC）的指导，确保数据透明和隐私保护。这种监管方式在支持技术进步的同时，也注重防范风险，力求平衡创新与社会责任。

2. 欧盟

欧盟则采取更为严格的监管措施，强调保护公众利益，尤其是在隐私和人权方面。2021 年欧盟发布《人工智能法案》，设立了严格的法律框架，对高风险 AI 应用，如医疗、司法提出透明度、可追溯性等要求。

3. 中国

中国的 AI 监管主要聚焦技术的安全性和可控性，强调防止技术滥用，并推动产学研合作。2022 年颁布《数据安全法》和《个人信息保护法》，进一步加强了对数据使用的管理，确保 AI 应用符合社会伦理要求。中国政府要求 AI 公司提高技术透明度和用户数据保护，确保技术应用不威胁社会稳定。例如，在面部识别技术应用中，中国要求企业在使用前必须获得用户明确同意，以加强数据隐私保护。

4. 其他国家

英国、加拿大和日本等国也在积极推动 AI 监管。英国的《AI 战略白皮书》侧重技术创新与公平治理的平衡；加拿大则通过《人工智能与数据伦理法案》加强对 AI 伦理的审查，确保算法透明和公正。2021 年，加拿大要求面向公众的 AI 系统提供决策透明度，并向用户解释其算法决策的依据，确保公众理解和控制。

5. 全球监管差异

不同国家的监管思路与政策导向差异，使得全球 AI 治理碎片化。这不仅增加了跨国企业的合规成本，还可能导致技术滥用、隐私侵犯等问题。美国在推动 AI 创新时较少监管，

而欧盟则设立了较为严格的法律框架，要求公司进行详细的风险评估。这样的差异导致企业需要根据不同地区的法律调整产品开发和运营策略，增加了企业的负担。此外，监管差异还可能增加全球范围内 AI 风险的扩散可能性。

（二）国际合作的必要性

随着 AI 技术的迅速发展，全球各国面临共同的挑战与机遇，没有哪个国家能够单独应对。因此，国际合作在 AI 治理中变得尤为重要，只有通过全球协作才能确保 AI 的健康发展，避免潜在的风险。

1. 国际合作是 AI 治理的关键

（1）全球性挑战需集体智慧：AI 的快速发展已经触及跨国界的多个领域，包括跨境数据流动、网络安全、自动化武器等。单一国家无法独立应对这些复杂的全球性挑战，因此国际合作显得尤为重要。以网络安全为例，由于 AI 技术在网络防护中的应用越来越广泛，国际间的合作对于构建有效的防御体系至关重要，2021 年 G7 峰会就曾讨论过如何通过跨国合作应对 AI 在网络安全、信息战中的潜在风险，并推动相关的国际立法和行为规范。此外，联合国提出的《数字治理蓝图》也是一个重要的国际合作框架，呼吁各国共同制定网络安全标准，并加强跨境数据治理，减少因 AI 技术滥用而可能带来的全球性威胁。

（2）促进技术普惠与共同发展：国际合作能够推动 AI 技术在全球范围内的普惠发展，特别是通过共享技术研究成果和治理经验，帮助发展中国家提升 AI 技术水平，从而缩小数字鸿沟。在过去几年中，中国通过提供技术援助、建设 AI 实验室等方式，帮助非洲国家发展 AI 技术，应用于农业、教育、医疗等领域，这不仅提升了农业生产力，还推动了当地技术人才的培养。通过这种合作，AI 技术的发展不仅仅局限于技术领先的国家，还能够惠及全球发展中的地区，促进全球科技公平。

（3）确保 AI 的安全性与公平性：AI 技术的快速发展伴随着众多不确定性，尤其在算法公平性、安全性和可靠性方面，国际合作显得至关重要。为了确保技术的正当发展，多个国家和国际组织已开始进行联合研究，共享 AI 安全和公平的最佳实践。欧盟和美国在 2021 年共同发布了关于 AI 伦理的合作声明，旨在推动双方在人工智能技术的伦理和法律问题上达成共识。这一合作特别强调了算法透明性、数据保护和避免 AI 技术歧视的问题。通过国际合作，欧盟和美国能够协调共同制定 AI 的透明性标准，避免在各自国家之间因为技术差异或不同的伦理标准导致的技术滥用。

2. 国际合作推动 AI 技术的健康发展和数据要素的跨境流动

（1）维护 AI 安全与数据保护：AI 发展带来的最大挑战之一是数据安全和隐私保护，特别是跨境数据流动。全球合作可以推动国际数据保护规则的制定，确保不同国家之间的数据政策互操作性，保护个人隐私，避免数据滥用。例如，欧盟的《通用数据保护条例（GDPR）》为全球数据治理树立了标杆，其他国家也可以通过合作对接这些国际标准。

（2）构建全球 AI 治理体系：国际组织如联合国和世界经济论坛（WEF）已开始着手推动全球 AI 治理框架的建立。2024 年 9 月，联合国秘书长高级别人工智能咨询机构发布《为人类治理人工智能》报告，呼吁全球各国共同合作，制定标准以确保 AI 发展的积极性并避免其风险。报告强调，各国需要携手合作，制定国际规则和标准，共享技术成果，提升治理能力，确保 AI 技术造福全人类。

（3）加强社会参与与公众素养：国际合作不仅仅限于政府和企业，社会各方的参与同样至关重要。全球 AI 治理体系的建设需要多元化的参与机制，包括公众意见征询、社会调查等方式。通过提高公众对 AI 技术的理解与认知，增强社会对 AI 技术安全性的认识，确保技术发展符合社会的整体利益。

国际合作在 AI 治理中扮演着至关重要的角色，它不仅能够帮助各国共同应对 AI 带来的全球性挑战，还能促进技术的健康发展和数据要素的跨境流动。通过国际合作，可以共同

制定和执行国际规则和标准，共享 AI 技术的研究成果和治理经验，提升 AI 技术的安全性、可靠性、可控性、公平性，并增进全人类福祉。

（三）地缘政治对国际合作的阻碍

在探讨 AI 的全球治理时，我们不得不面对地缘政治因素对 AI 国际合作产生的复杂影响。

1. 地缘政治对 AI 技术共享的影响

地缘政治紧张局势直接限制了 AI 技术的全球共享，特别是当国家在技术领域争夺领先地位时。例如美国对中国的科技禁令限制了高端 GPU 等关键 AI 硬件的出口，2020 年，美国将中国的几家科技公司列入“实体清单”，禁止其购买美国技术。这种限制不仅妨碍了技术的自由流动，也让 AI 领域的技术壁垒更加明显，进一步加剧了全球技术分裂，影响了全球 AI 的创新和进步。

2. 大国干预行为对 AI 伦理和法律标准的影响

在全球 AI 治理中，大国通过政策干预试图推广自己对 AI 伦理和法律的标准，导致国际标准的分歧。比如美国与欧盟在隐私保护上的差异，美国注重技术创新，而欧盟则更注重隐私权保护。欧盟的《通用数据保护条例（GDPR）》提出了严格的数据隐私保护措施，而美国的监管政策则相对宽松。这些差异不仅加剧了 AI 技术的不同应用场景，也使得国际合作在数据保护和隐私方面面临较大挑战。

3. 地缘政治风险对 AI 全球治理框架的影响

随着全球地缘政治风险的上升，国家间的信任度下降，全球 AI 治理的合作变得更加困难。俄罗斯与西方国家的紧张关系影响了它们在 AI 领域，尤其是在数据共享和算法透明度方面的合作。俄罗斯政府近年来推行“数据主权”政策，限制外国企业对国内数据的获取，这使得跨国 AI 项目的合作受限，削弱了全球 AI 治理框架的统一性。

4. AI 政策与地缘政治的交织对国际合作的影响

AI 政策的制订与地缘政治、经济竞争的交织，使得全球合作进程变得缓慢。比如印度与中国在数据治理上的分歧。印度出台了严格的数字主权政策，要求所有数据必须在本国存储并进行监管，这与中国的“网络安全法”政策形成对比。两国在数据共享和跨境数据流动上的不同政策，使得国际合作的进展面临阻力。全球范围内的 AI 治理难以统一，政策差异加剧了技术的碎片化和治理的复杂性。

5. 中美地缘科技竞争对 AI 治理合作的影响

中美在 AI 领域的竞争是当前全球 AI 合作面临的主要地缘政治挑战。美国通过对中国技术企业实施出口禁令，限制了中国企业在 AI 硬件（如芯片）领域的发展，这不仅影响了两国间的合作，还加剧了全球 AI 技术的发展分裂。美国政府在 2018 年出台了针对中国 AI 公司的限制措施，这些措施使得中国无法获取美国的高端 AI 芯片，导致中国在 AI 领域的发展受到遏制。这种地缘科技竞争不仅影响了中美之间的 AI 合作，也使得全球 AI 技术发展面临更加不确定的局面，导致国际合作陷入僵局。

地缘政治对全球 AI 合作的影响显而易见。从 AI 技术共享的限制，到大国干预行为引发的伦理和法律分歧，再到政治风险对治理框架的破坏，这些因素共同作用，使得全球 AI 治理合作进展缓慢，增加了全球技术发展的不确定性。克服这些障碍需要各国共同努力，通过对话与合作，促进全球 AI 技术的健康发展。

（四）推动国际合作的策略和建议

为应对 AI 治理中的全球性挑战，国际合作至关重要。以下是推动国际合作的具体策略和建议：

1．建立国际 AI 专家小组：成立一个由全球 AI 领域专家组成的独立国际小组，发布关于 AI 能力、机遇、风险和不确定性的年度报告，并针对新兴问题发布特别报告，为全球 AI

治理提供科学依据和建议。
2．开展多方政策对话：在联合国框架下，每年举办两次政府间和多利益相关方的 AI 治理政策对话，分享最佳实践，推动人权保护，提升国际间的互操作性与协作机制。
3．创建全球 AI 能力发展网络：将联合国及相关国际组织的能力发展中心进行联动，构建全 AI 能力发展网络，为低收入国家和地区提供专业知识、计算资源和训练数据，推动技术共享与能力提升。
4．设立全球 AI 基金：通过全球 AI 基金资助 AI 研究与应用，尤其支持贫困国家，以确保技术的公平分配，缩小全球技术差距，推动全球 AI 发展的包容性和普惠性。
5．促进全球 AI 数据共享框架：建立全球 AI 标准与数据共享平台，推动国际间的数据互通与合作，特别关注发展中国家，确保其能够从 AI 技术发展中受益，并参与全球治理。
6．在联合国设立 AI 专责机构：在联合国秘书处设立 AI 办公室，专责协调全球 AI 治理事务，推动国际合作与政策对话，确保各国在 AI 发展中的共同利益和治理责任。
7．加强社会参与与公众素养：建立多元化的社会参与机制，鼓励公众参与 AI 政策制定，通过公开征求意见和社会调查等方式，提高公众对 AI 技术的认知与理解，确保技术发展符合社会的整体利益。
8．推动 AI 在可持续发展中的应用：促进 AI 在工业创新、环境保护、资源管理、能源效率及生物多样性等领域的应用，提升医疗、教育和养老等社会福祉，确保 AI 的可持续发展造福全人类。

通过这些策略，我们可以推动全球 AI 的负责任发展，确保其符合全球公平与可持续发展的目标，并在多国协作中实现技术与伦理的平衡。

二、企业以合规与信任为主要治理目标，聚焦伦理与安全

在友商的 AI 治理实践中，普遍设立了专门的伦理委员会或 AI 治理团队，制定了明确的伦理原则，强调以人为本、数据安全、隐私保护、算法透明度与公平性等核心价值，并提倡跨部门协作与人才培养，确保技术合规与社会责任。同时，部分友商在行业标准制定和国际合作方面有所创新，尤其是在 AI 安全、可信度评测以及透明度工具的开发上，提升了治理的实操性和有效性。对于腾讯而言，可以借鉴这些经验，推动更为完善的 AI 治理框架，增强公众信任，促进技术与伦理的有机结合。

（一）OpenAI

OpenAI 下设安全与安全委员会、研究与安全副总裁办公室和已解散的 AGI Readiness 团队，主张确保 AI 技术（特别是 AGI）的开发遵循严格的安全和伦理标准，平衡技术进步与社会责任，推动跨学科合作与外部专家参与，确保技术应用符合社会伦理、法律规范，并最大化技术的社会福祉。

1. 安全与安全委员会（Safety and Security Committee）

（1）工作：

① 评估安全流程：委员会成立后的首个 90 天内，主要任务是对 OpenAI 的工作流程和保障措施进行评估，确保公司在 AI 研发过程中遵循严格的安全伦理标准。

② 引入外部专家：为了增强评估的广度和深度，OpenAI 计划邀请外部安全和技术专家，协助委员会进行全面的评估，以确保客观性和专业性。

③ 公开透明：在全体董事会审查后，委员会将以符合安全和保障的方式，公开分享所采纳的建议，展现责任感，并向公众和利益相关者传达透明度。

（2）目标：

① 确保 OpenAI 的技术开发和应用始终以安全和伦理为核心，特别是在追求更高水平的 AGI 时，平衡技术进步与社会责任，防止可能的安全隐患或社会风险。

② 促进外部专家和多方利益相关者的参与，确保 AI 的开发过程具有广泛的审查机制和社

会监督。
（3）负责人：OpenAI CEO Sam Altman
（4）负责人观点：确保在 AI 技术（尤其是 AGI）发展的过程中，技术进步与社会安全之间达到平衡，并最终确保 AI 技术能够造福人类社会。
2. 研究与安全副总裁办公室（Research and Safety VP Office）
（1）工作：
① 风险管理：负责评估和管理 OpenAI 模型可能面临的风险。确保在开发和部署过程中，AI 技术符合安全标准，以避免潜在的社会、伦理和安全问题。
② 安全研究：专注于 AI 安全领域，进行深入的安全研究，识别和缓解 AI 模型中可能出现的安全威胁，确保 AI 技术在实际应用中保持高度的可靠性和安全性。
③ 跨部门协作：提倡跨部门协作，建立和实施全面的 AI 治理框架，确保开发和应用过程符合伦理法律规范。
（2）目标：
① 在 AI 的研发过程中，尤其是涉及到高风险应用时，确保采取切实的安全措施，并评估潜在的影响，以应对可能出现的风险。
② 推动 AI 开发的伦理审查机制和透明性，增强用户对 AI 技术的信任。
③ 强化多学科合作，确保 AI 技术的开发不仅符合技术要求，同时也满足伦理、法律和社会责任。
（3）负责人：OpenAI 研究与安全副总裁 Lilian Weng
（4）负责人观点：强调 AI 系统应避免偏见，确保公平与透明；提倡引入伦理审查机制并推动跨学科合作，共同制定有效的 AI 治理框架。
3. AGI Readiness 团队（已解散）
（1）主要职责：
① 制定 AI 伦理准则：团队专注于为 AGI 的开发制定伦理框架，确保 AGI 技术的开发和应用符合伦理标准，避免对社会产生潜在风险。
② 跨部门协作：与 OpenAI 内部的多个部门协作，推动 AI 治理机制的实施，确保 AGI 技术在各种应用场景中的安全性和可靠性。
③ 参与行业标准制定：团队积极参与全球范围内的 AI 治理标准的制定工作，推动行业内的规范化发展。
④ 开展 AI 安全研究：团队进行 AGI 安全性研究，评估 AGI 技术的潜在风险，并提出相关的防范措施。
（2）负责人：Ahmad Al-Dahle
（3）负责人观点：主张 AGI 开发应遵循严格的伦理标准，确保公平、透明、安全；倡导推动跨部门协作，确保技术符合社会伦理和法律规范。
（二）Microsoft
微软下设 AI 伦理与社会影响团队（Aether）、负责任 AI 办公室（ORA）和 RAISE 团队，主张将 AI 伦理嵌入开发各阶段，构建全面治理框架，推动跨部门协作，确保技术安全、隐私保护和合规性，并强调人才培养和伦理教育支持 AI 技术的健康发展。
1. Office of Responsible AI（ORA）
（1）工作：
① 制定负责任 AI 标准：发布《微软负责任 AI 标准》，基于六项原则，为 AI 系统的开发和部署提供了框架。
② 建立治理框架：与 AI 伦理与社会影响委员会（Aether）和工程团队合作，构建了全面的 AI 治理框架，确保伦理原则在工程实践中得到有效实施。

③ 发布透明度报告：发布了首份 AI 透明度报告，概述了微软在 AI 产品部署方面的成就和挑战，强调了风险识别、安全性等问题的重要性。

（2）负责人：首席负责任人工智能官 Natasha Crampton

（3）负责人观点：主张将伦理原则嵌入 AI 开发的各个阶段；倡导开发确保安全性和隐私的工具；推动跨部门协作应对技术挑战，构建全面治理框架，确保决策过程公正透明、可验证；认为 AI 治理依赖于人才支持。

2. Responsible AI Strategy in Engineering（RAISE）

（1）成果 / 措施：

① 开发实用工具和方法：与 AI 研究、策略和工程专家协作，开发了支持 AI 安全性、隐私、安全和质量的工具和方法，并将其嵌入到 Azure AI 平台。这些工具帮助开发者在产品设计和开发过程中考虑伦理和社会影响。

② 构建治理框架：与 AI 伦理与社会影响委员会（Aether）和负责任人工智能办公室（ORA）合作，建立了一个全面的 AI 治理框架，确保伦理原则在工程实践中得到有效实施。

③ 推动跨部门协作：倡导建立跨部门的治理机制，促进技术、产品、法律等团队的协同合作，共同应对 AI 治理挑战，确保技术应用的合规性和安全性。这种协作机制有助于在整个公司范围内推广负责任的 AI 实践。

④ 加强人才培养与伦理教育：强调 AI 治理需要专业的人才支持，通过内部培训和外部招聘，提升团队的 AI 治理意识和能力，确保技术应用符合社会伦理和法律规范。提供了详细的指引手册，帮助团队在设计过程中实施负责任的 AI。

（2）负责人：首席负责任人工智能官 Natasha Crampton

3. AI 伦理与社会影响团队（Aether）（已于 2023 年 3 月解散）

（1）工作：

① 制定 AI 伦理原则：Aether 委员会组织工作组，专注于解决问题、分析和开发六项负责任 AI 原则。这些原则为微软的 AI 开发和应用提供了指导框架。

② 构建治理框架：与微软的 Office of Responsible AI（ORA）和 Responsible AI Strategy in Engineering（RAISE）团队合作，建立了全面的 AI 治理框架，确保伦理原则在工程实践中得到有效实施。

③ 开发工具和系统：与工程团队合作，定义并实施用于负责任地使用 AI 的工具和系统策略，帮助开发者在产品设计和开发过程中考虑伦理和社会影响。

④ 发布治理蓝图：发布报告《治理人工智能：未来蓝图》，分享了微软在 AI 治理方面的经验和策略，强调了科学治理人工智能的重要性。

（2）负责人：首席技术官 Kevin Scott

（3）负责人观点：主张将 AI 伦理从理论转化为实践，嵌入开发各阶段；主张构建全面治理框架，推动多部门协作确保合规与安全；倡导 AI 从业者接受持续伦理教育，确保技术遵守伦理规范。

（三）Google

谷歌下设 DeepMind、Responsible AI 和 AI 伦理委员会，主张制定明确的 AI 伦理原则，确保技术应用符合社会伦理和法律规范，避免偏见，推动跨部门协作，确保 AI 系统的透明度、公平性和可解释性，并强调技术应辅助而非取代人类。

1. Google DeepMind

（1）工作：

① 制定 AI 伦理原则：发布明确的 AI 伦理原则，强调 AI 技术应为社会带来积极影响，避免强化偏见，确保安全性，遵循隐私设计原则，保持高标准的科学严谨性，具备可解释性，并在适当时接受人类的控制。

② 成立 AI 伦理委员会：负责监督和指导 AI 项目的伦理决策，评估 AI 技术对的社会影响，确保技术开发符合伦理和法律标准。
③ 开展 AI 伦理研究：关注 AI 技术在实际应用中的伦理和社会影响，确保技术应用符合社会伦理和法律规范。
④ 推动跨部门协作机制：倡导建立跨部门的治理机制，促进技术、产品、法律等团队的协同合作，共同应对 AI 治理挑战，确保技术应用的合规性和安全性。
（2）负责人：DeepMind CEO Demis Hassabis
（3）负责人观点：AI 具有巨大潜力，呼吁在开发和部署时保持谨慎；主张建立全球统一的 AI 治理框架，确保安全性、可靠性与伦理标准；强调技术透明度和可解释性；认为 AI 应作为人类的工具，辅助而非取代人类。
2. Responsible AI
（1）工作：
① 制定 AI 伦理原则：发布 AI 伦理原则，强调 AI 技术应为社会带来益处，避免加剧不公平偏见，确保安全性，遵循隐私设计原则，保持科学严谨性，具备可解释性，并在适当时接受人类的控制。
② 建立 AI 治理框架：跨部门合作，构建了全面的 AI 治理框架，确保伦理原则在工程实践中得到有效实施。
③ 开发公平性指标工具：开发 Fairness Indicators 工具，帮助开发者评估和改进 AI 模型的公平性，确保 AI 系统在不同人群中的表现一致。
④ 开展 AI 伦理研究：积极开展 AI 伦理研究，关注 AI 技术在实际应用中的伦理和社会影响，确保技术应用符合社会伦理和法律规范。
（2）负责人：前谷歌 AI 伦理研究负责人 Margaret Mitchell 和 Timnit Gebru
（3）负责人观点：主张 AI 系统必须避免偏见，确保公平性与透明度；提倡在开发过程中引入伦理审查机制，并通过跨学科合作制定和实施有效的治理框架。
3. AI 伦理委员会
（1）工作：
① 制定 AI 伦理原则：委员会协助制定了谷歌的 AI 伦理原则，强调 AI 技术应当对社会有益，避免制造或强化不公平偏见，确保安全性，遵守隐私设计原则，保持高标准的科学严谨性，具备可解释性，并在适当情况下接受人类控制。
② 建立 AI 治理框架：跨部门合作，构建了全面的 AI 治理框架，确保伦理原则在工程实践中得到有效实施。
③ 开展 AI 伦理研究：委员会积极开展 AI 伦理研究，关注 AI 技术在实际应用中的伦理和社会影响，确保技术应用符合社会伦理和法律规范。
④ 推动跨部门协作机制：委员会倡导建立跨部门的治理机制，促进技术、产品、法律等团队的协同合作，共同应对 AI 治理挑战，确保技术应用的合规性和安全性。
（2）负责人：前谷歌 AI 伦理研究负责人 Margaret Mitchell 和 Timnit Gebru
（四）Amazon
亚马逊下设 AWS AI 服务团队，主张遵循公平性、透明性、问责制和隐私保护等核心原则，推动跨部门协作，开发治理工具，确保 AI 技术的公平、可靠和可持续性，同时倡导负责任的 AI 实践并提供教育与培训。
1. AWS AI 服务团队
（1）工作：
① 制定负责任的 AI 原则：强调在 AI 系统的开发和部署中，遵循公平性、透明性、问责制和隐私保护等核心原则，确保社会责任和伦理符合预期，造福社会。

② 开发治理工具和框架：提供多种治理工具和框架，帮助客户在其 AI 应用中实施治理措施。例如，Amazon Bedrock 护栏，旨在帮助用户在生成式 AI 应用中实施安全措施，确保其与负责任的 AI 政策保持一致。
③ 提供教育和培训：通过博客、白皮书、线上培训课程等方式，向客户和开发者普及负责任的 AI 实践，提升各方对 AI 治理的认识与实践能力。
④ 推动跨部门协作：倡导跨部门的治理机制，促进不同领域团队的协同合作，共同应对 AI 治理带来的复杂挑战，确保技术应用合规性、安全性及可持续性。
（2）负责人：前 Alexa 首席科学家 Rohit Prasad
（3）负责人观点：强调亚马逊致力于开发先进、可靠且可持续的 AI 技术；提倡始终以人为中心，确保技术公平性、透明性和可靠性，同时推动跨学科合作制定有效的 AI 治理框架，并严格遵循伦理标准。
（五）Meta
Meta 下设 Responsible AI Team、AI 伦理与社会影响团队和 AI 治理委员会，主张确保 AI 系统的公平性、透明度和安全性，强调减少算法偏见，推动跨领域合作和伦理审查机制，制定 AI 治理框架，确保技术应用符合社会伦理和法律规范，最大化社会效益。
1. AI 伦理与社会影响团队（AI Ethics and Society Team）
（1）工作：
① 研究社会影响：团队关注 AI 技术对社会的广泛影响，特别是其在不同应用场景中的伦理挑战。
② 制定伦理指南：根据不同应用场景的伦理问题，制定相应的伦理指南，确保 AI 技术应用符合社会伦理、法律及规范要求。
（2）目标：促进 AI 技术与社会伦理之间的和谐共存，减少可能带来的负面社会影响，确保 AI 技术的公平、公正、安全应用。
2. AI 治理委员会（AI Governance Committee）
（1）工作：
① 监督和指导 AI 项目：负责 Meta 内部 AI 项目的监督，确保所有 AI 项目符合公司制定的伦理标准。
② 评估 AI 技术的社会影响：委员会评估各类 AI 技术应用的社会影响，确保其符合伦理原则和法律规范。
③ 制定 AI 治理框架：该委员会负责制定并实施公司层面的 AI 治理框架，确保 AI 技术在公司内部的开发与应用严格遵守伦理标准和社会规范。
（2）目标：通过系统的治理框架，确保公司所有 AI 技术和应用在开发和部署过程中遵循负责任的 AI 原则，最大化社会效益并减少潜在的风险。
3. Responsible AI Team（已于 2023 年 11 月解散）
（1）工作：
① 专注伦理与安全标准：专注于确保公司 AI 系统在开发和部署过程中遵循严格的伦理和安全标准，致力于减少算法偏见，并提升 AI 模型的公平性和透明度。
② 减少算法偏见：推动了多项措施，解决 AI 技术在实际应用中可能存在的种族、性别和其他形式的偏见。
③ 提升透明度与可解释性：致力于提高 AI 系统的透明度，确保用户能够理解和解释 AI 的决策过程，从而增加公众对技术的信任。
（2）负责人：Margaret Mitchell
（3）负责人观点：主张 AI 治理应确保系统公平性和透明度，特别是在敏感技术领域；呼吁开发者考虑多样性并引入伦理审查机制；强调多领域专家跨学科协作，共同制定有效的治

理框架。

（六）阿里巴巴

阿里巴巴下设科技伦理治理委员会和人工智能治理与可持续发展研究中心（AAIG），主张始终坚持以人为本的价值观，推动技术可持续发展，确保 AI 技术最大化造福社会，并注重隐私与安全，探索企业自律、行业共治与监管相结合的治理模式，同时关注技术的伦理与社会责任。

1. 科技伦理治理委员会

（1）工作：

① 发布《AI 治理与可持续发展实践白皮书》和《AIGC 治理与实践白皮书》。

② 制定了六大科技伦理准则，旨在为阿里巴巴的 AI 技术开发和应用提供伦理指导。

③ 提倡以人为中心，推动技术与社会价值的对接，确保 AI 技术的发展能够最大化造福社会。

（2）负责人：阿里巴巴 CTO 程立

（3）负责人观点：主张始终坚持以人为本的价值观，确保在 AI 应用中优先考虑人类福祉，推动技术可持续发展；探索企业自律、行业共治与监管相结合的治理模式，保障隐私与安全。

2. 人工智能治理与可持续发展研究中心（AAIG）

（1）工作：AAIG 专注于推动人工智能技术的伦理与可持续发展研究，参与制定并发布了多个行业标准，如 YD/T 3658-2020（互联网垃圾内容治理系统技术要求）和 ITU 国际标准（增强认证框架、广告反垃圾技术等），这些标准为 AI 的应用提供了技术指引，确保技术创新遵循国际规范和伦理要求。

（2）负责人：薛晖

（3）负责人观点：强调迈向更强大、通用的人工智能系统需要长期探索与技术积累；重视安全隐患和社会伦理，构建可用、可靠、可信、可控的智能系统，并推动科技服务全社会，尤其是弱势群体，践行“人人受益、责任担当、开放共享”的理念。

（七）字节跳动

字节跳动下设产品研发与工程部（PDI）和 Flow 部门，主张 AI 技术发展与伦理治理相结合，确保符合社会伦理和法律规范，特别关注隐私保护和数据安全，倡导跨部门协作与人才培养，推动 AI 治理框架的完善，确保技术合规与安全。

1. 产品研发与工程部（PDI）

（1）工作：

① 建立专门团队：下设 Flow 部门，专注于 AI 大模型应用的研发和治理，负责 AI 产品的开发和治理，确保技术应用符合社会伦理和法律规范。

② 制定 AI 治理框架：制定了一个用于指导 AI 技术的研发和应用的 AI 治理框架，确保技术符合社会伦理和法律要求，涵盖数据隐私保护、算法公平性、透明度和可解释性等方面，确保 AI 系统的决策过程公正可靠。

③ 开展 AI 伦理研究：关注 AI 技术在实际应用中的伦理和社会影响，推动技术发展同时确保其符合伦理要求。

④ 加强人才培养：通过内部培训和外部招聘，提升团队的 AI 治理意识和能力，确保技术应用符合社会伦理和法律规范。

（2）负责人：大模型团队负责人朱文佳

（3）负责人观点：强调 AI 技术发展必须与伦理治理相结合，确保符合社会伦理和法律规范；特别关注隐私保护和数据安全；倡导跨部门协作，推动团队协同工作；强调人才培养以支持 AI 技术的健康、可持续发展。

2. Flow 部门

（1）工作：

① 产品研发与伦理考量：Flow 部门下设 AI 教育、国际化、社区和豆包四大业务线，推出了 AI 应用开发平台“Coze”、AI 问答助手“豆包”等产品。在产品开发过程中，部门注重 AI 伦理和安全，确保技术应用符合社会伦理和法律规范。

② 跨部门协作机制：Flow 部门隶属于 PDI，与其他部门密切合作，建立跨部门的治理机制，确保 AI 技术的合规性和安全性。

③ 人才引进与培训：积极招聘产研等岗位人才，特别是从其他知名企业引进具有 AI 治理经验的专业人士。同时，部门通过内部培训提升团队的 AI 治理意识和能力，确保技术应用符合社会伦理和法律规范。

（2）负责人：

① 产品负责人：产品与战略副总裁朱骏

② 技术负责人：技术副总裁洪定坤

（3）负责人观点：

① 朱骏（产品负责人）：主张技术创新应以人为本，注重隐私保护和数据安全；提倡建立完善的 AI 治理框架，关注算法公平性、透明度和可解释性；倡导跨部门协作，共同应对 AI 治理挑战，确保技术合规与安全。

② 洪定坤（技术负责人）：强调技术创新与伦理并重；主张完善 AI 治理框架，通过跨部门协作和人才培养提升治理能力，确保技术应用符合社会伦理和法律规范。

（八）百度

百度下设科技伦理委员会和百度研究院，主张 AI 技术发展必须以人为本，确保安全可控，推动技术健康发展，注重算法透明度与可追溯性，保障数据合法性与安全性，并通过产学研及国际合作推动全球 AI 技术的规范化与健康发展。

1. 科技伦理委员会

（1）工作：百度科技伦理委员会的设立旨在加强 AI 技术的伦理治理，确保人工智能的安全、可靠、可控，委员会提出并倡导六项核心 AI 伦理举措。

（2）负责人：百度 CEO 助理李震宇

（3）负责人观点：强调降低 AI 算法“黑箱风险”，提升治理透明度，避免技术滥用；秉持“以人为本、智能向善”的原则，推动产学研及国际合作，共同制定和实施全球 AI 伦理标准。

2. 百度研究院

（1）工作：

① 发布科技趋势预测：定期发布科技趋势预测，重点关注 AI 技术的普惠性和价值导向，尤其关注中小企业和社会弱势群体需求，推动 AI 技术的包容性发展。

② 参与行业标准制定：积极参与中国人工智能产业发展联盟安全治理委员会的工作，参与制定行业规范《代码大模型安全风险防范能力要求及评估方法》，推动 AI 安全标准的建设和完善。

③ 推动可信 AI 评测：在多个 AI 产品领域（如 AI 开发平台、OCR、内容审核和智能客服等）通过了中国信息通信研究院的可信 AI 评测。

④ 发布 AI 应用案例集：发布《AI 技术产业化 蓬勃发展正当时——百度生态伙伴 AI 应用案例集》，梳理了 16 大行业的 AI 应用场景和 50 个实际落地案例，推动各行业智能化升级。

（2）负责人：百度 CTO 王海峰

（3）负责人观点：主张 AI 技术发展必须以人为本，确保安全可控，推动技术健康发展，保障数据合法性与安全性，增强算法透明度与可追溯性，并通过跨国合作推动全球 AI 技术的规范化与健康发展。

（九）华为

华为任命了 AI 治理首席专家段小琴，负责公司在 AI 产业和技术政策领域的洞察、规划和策略制定。该角色旨在探索新产业机会，识别政策风险，推动公司决策和布局，确保 AI 技术的发展符合伦理和法律要求。

段小琴认为，人工智能技术的发展应重点关注公平性、透明性和责任性，强调全球 AI 治理需要多方合作，共同应对挑战，以实现人工智能的长期可持续健康发展。

（十）商汤科技

商汤科技在人工智能治理领域高度重视伦理与合规，设立了人工智能伦理与治理委员会，制定了明确的工作职责，并由资深负责人领导，积极推动 AI 技术的负责任发展。

1. 人工智能伦理与治理委员会

（1）工作：制定伦理原则、宣传伦理理念，以及推动具体伦理措施的落实。其中，核心工作之一是对所有产品线进行伦理审核，从算法、数据和社会影响三个方面进行评估，防范和应对可能的伦理风险。

（2）负责人：商汤科技智能产业研究院院长田丰

（3）负责人观点：强调人工智能治理应以价值为引领，技术为基础，多方参与，推动负责任且可评估的 AI 发展，构建全生命周期的伦理治理闭环。

（十一）科大讯飞

科大讯飞设立了科技伦理委员会，积极推进 AI 治理相关工作。

1. 科技伦理委员会

（1）工作：科大讯飞成立了科技伦理委员会，旨在加强人工智能技术的伦理治理，确保 AI 技术的安全、可靠、可控。委员会提出并倡导六项核心 AI 伦理举措，包括以人为本、智能向善、算法透明、数据安全、隐私保护和公平公正。

（2）负责人：科大讯飞高级副总裁、研究院院长刘聪

（3）负责人观点：调以人为本，确保 AI 技术安全可控，推动算法透明化与可追溯性，保障数据合法性与安全性，倡导全球合作制定 AI 伦理标准，促进技术的健康发展。

（十二）智谱华章

智谱华章在人工智能治理方面积极参与国内外的行业合作，与 OpenAI、谷歌等公司共同签署了前沿人工智能安全承诺，推动 AI 伦理的标准化。

1. 人工智能智库网络

（1）工作：智谱 AI 与百度、快手、华为、蚂蚁、腾讯等公司共同发起了人工智能智库网络，旨在促进 AI 治理的研究与实践。

（2）负责人：智谱 AI 董事长刘德兵

（3）负责人观点：强调，人工智能的发展应注重伦理和安全，倡导技术向善，并推动全球 AI 伦理标准的制定与实施。

三、人工智能治理的工程实践与工具

（一）AI 安全与隐私保护的工程实践

随着人工智能技术的广泛应用，AI 安全和隐私保护成为了治理中的核心议题。在实际的工程实践中，通过差分隐私、对抗样本防护和安全多方计算等技术手段，可以有效保障 AI 系统中的数据安全，防止敏感信息泄露，并提升系统对恶意攻击的防御能力。

1. 差分隐私技术：保护个人数据隐私

差分隐私是一种通过加入噪声保护个人隐私的技术，使得即使数据被公开分析，个体的信息也无法被准确推断出来。TensorFlow Privacy 是由 Google 开发的一个开源库，专门用于在深度学习模型中实现差分隐私。它可以在模型训练过程中引入噪声，从而保护训练数据中的个人信息。另一款常用工具是 PySyft，由 OpenMined 社区开发，它不仅支持差分隐私，还支持联邦学习等隐私保护技术，广泛应用于数据隐私保护领域，尤其是在涉及医疗、金融

等行业的 AI 应用中。

2. 对抗样本防护：提高模型的安全性和鲁棒性

对抗性攻击对 AI 模型的安全构成威胁，攻击者通过细微的扰动欺骗模型做出错误判断。为应对这一挑战，Adversarial Robustness Toolbox (ART)是由 IBM 提供的一个开源库，它提供了多种对抗性防护算法，如对抗训练和梯度掩蔽，帮助开发者增强模型的对抗鲁棒性。CleverHans 是由 Google Brain 提供的另一个开源库，旨在生成对抗样本并测试模型的抗攻击能力，帮助研究人员和工程师优化模型，提升其在面对恶意攻击时的稳定性和安全性。

3. 安全多方计算：保证数据隐私与安全

安全多方计算 (SMPC) 允许多个参与方在不暴露各自数据的前提下，共同进行计算。PySyft 是由 OpenMined 社区开发的一款开源工具，支持 SMPC，并能够在隐私保护的同时进行去中心化的机器学习训练，尤其在医疗、金融等领域得到了广泛应用。此外，HElib 由 IBM 和其他研究机构联合开发，是一个用于同态加密的开源库，它支持在加密数据上进行计算，确保数据隐私的保护。

4. 隐私保护的其他技术

除了差分隐私和安全多方计算，同态加密和隐私计算也是常用的隐私保护技术。同态加密允许对加密数据直接执行计算，而无需解密数据，从而确保数据的隐私性。IBM 提供的 HElib 是一个同态加密的实现，它广泛应用于需要对加密数据进行复杂计算的场景。隐私计算通过结合差分隐私、同态加密和安全多方计算等技术，确保在数据分析和共享过程中，隐私得到保护。例如，国内腾讯云提供的隐私计算平台将隐私计算与大数据分析相结合，支持多方安全计算，尤其在金融、医疗等行业中具有广泛的应用。

5. AI 安全与隐私保护的合规性和伦理考虑

AI 治理不仅依赖技术手段，还必须遵守相关的法律法规和伦理要求。欧盟《通用数据保护条例 (GDPR)》对数据隐私和安全提出了严格要求，要求所有涉及欧盟公民数据的 AI 系统必须保障数据主体的隐私权，明确数据收集、处理和存储的合规性。中国《个人信息保护法》同样提出了严格的数据保护要求，强调对个人信息的隐私保护和用户数据的合法使用。此外，中国工信部发布的《人工智能伦理治理指导意见》提出了 AI 伦理的治理框架，要求开发者遵守公平性、透明性、责任性等伦理原则，确保 AI 技术在应用中不偏离伦理和法律轨道。

（二）AI 公平性与偏见消除的工程实践

随着人工智能技术的广泛应用，AI 系统在决策过程中可能会引入或放大现有社会偏见，导致不公平的结果。为此，工程实践中需要采取有效措施，识别、评估并缓解 AI 模型中的偏见，确保其决策过程的公平性。

1. 数据集的多样性与代表性：从源头减少偏见

确保训练数据的多样性和代表性是减少模型偏见的根本途径。模型偏见往往源于训练数据的偏差，数据集缺乏多样性或对某些群体的代表性不足，可能导致模型在这些群体上的表现不佳。因此，在数据收集和准备阶段，开发者应确保数据集涵盖不同性别、种族、年龄等群体，避免数据偏差。此外，采用数据增强技术，如过采样少数群体数据，也有助于提高模型的公平性。

2. 数据预处理：确保训练数据的公平性

数据预处理是消除 AI 模型偏见的重要步骤，通过清洗和调整数据，确保训练数据的公平性。训练数据中的偏差可能导致模型在特定群体上的不公平表现。因此开发者应在数据预处理阶段采取措施，如删除敏感属性、平衡数据集中的不同群体比例，以及使用重加权（Reweighing）等技术，确保模型在各群体间的公平性。例如，IBM 的 AI Fairness 360 (AIF360)提供了多种预处理算法，帮助开发者在模型训练前调整数据，减少潜在偏见。

3. 偏见检测与评估工具：识别模型中的不公平性

有效的偏见检测对实现 AI 公平性至关重要，利用专门的工具可以识别模型在不同群体间的性能差异。偏见检测工具通过分析模型输出，评估其在不同人群中的表现差异，从而识别潜在的不公平性。例如，AIF360 提供了多种公平性指标和偏见检测算法，帮助开发者评估模型的公平性。此外，Google 的 Fairness Indicators 也是一款用于评估模型公平性的工具，支持在 TensorFlow Extended (TFX)管道中集成，实时监测模型的公平性表现。

4. 偏见缓解算法：减少模型决策中的不公正

在检测到模型存在偏见后，采用偏见缓解算法是消除不公平性的关键步骤。偏见缓解算法通过调整数据、模型或预测结果，减少模型决策中的不公正。AIF360 提供多种偏见缓解方法，包括预处理、在处理和后处理阶段的算法，如重新加权（Reweighing）、对抗性去偏（Adversarial Debiasing）等。这些方法可以在模型训练前调整数据分布，或在训练过程中加入公平性约束，亦或在预测结果后进行校正，从而有效减少偏见。

5. 公平性指标的选择与应用：量化模型的公平性

选择适当的公平性指标是评估和改进模型公平性的基础。公平性指标用于量化模型在不同群体间的表现差异，常用的指标包括统计均等（Statistical Parity）、均等机会（Equal Opportunity）和预测平价（Predictive Parity）等。开发者应根据具体应用场景选择合适的指标，全面评估模型的公平性表现。例如，在信用评分模型中，均等机会指标可以评估不同种族群体获得贷款批准的机会是否相等。

6. 模型审计与持续监测：确保公平性的长期维护

定期对模型进行审计和监测，有助于及时发现并纠正偏见，确保 AI 系统的持续公平性。模型在部署后，可能由于环境变化或数据更新而引入新的偏见。因此，建立持续的模型审计和监测机制，定期评估模型的公平性表现，及时采取偏见缓解措施，是维护 AI 系统公平性的关键。例如，使用 Fairness Indicators 等工具，可以实时监测模型在不同群体间的性能差异，确保其决策过程的公正性。

（三）AI 模型的可解释性与透明度提升实践

随着 AI 治理要求的提高，模型的可解释性和透明度逐渐成为焦点。提高模型的可解释性有助于用户理解模型决策过程，增强信任度，并确保模型在实际应用中的可靠性。

1. 模型结构优化：采用易于解释的算法

选择和设计易于解释的模型结构是提升可解释性的基础。在模型开发阶段，采用如决策树、线性回归等本身具有高可解释性的算法，可以使模型的决策过程更透明。这些模型通过清晰的规则或线性关系，便于用户理解其预测依据。例如，决策树模型通过树状结构展示决策路径，直观地反映特征对结果的影响。

2. 局部可解释性工具：揭示单个预测的决策依据

利用局部可解释性工具，可以深入理解模型对单个输入的预测原因。工具如 LIME（Local Interpretable Model-agnostic Explanations）和 SHAP（SHapley Additive exPlanations）通过构建简化模型或计算特征贡献，解释复杂模型的单次预测结果。LIME 通过线性模型近似复杂模型的局部行为，SHAP 则基于博弈论计算特征对预测的贡献值，帮助用户理解模型在特定输入下的决策逻辑。

3. 可视化技术：直观展示模型决策过程

可视化技术使模型的内部机制和决策路径更直观，便于用户理解。通过热力图、决策路径图等可视化手段，展示模型在处理输入数据时的关注点和决策流程。例如，在图像识别任务中，热力图可以显示模型关注的图像区域；在分类任务中，决策路径图展示模型的决策过程。这些可视化工具帮助用户直观地理解模型的工作机制，增强对模型的信任。

4. 数据透明度：公开训练数据来源和处理方式

确保数据透明度是提升模型透明度的重要环节。公开模型的训练数据来源、收集方法和预处理步骤，使用户了解模型的基础数据情况。这有助于评估模型的适用范围和潜在偏差，确保模型在不同应用场景中的可靠性。例如，详细记录数据清洗、特征工程等过程，提供数据文档，增强数据透明度。

5. 算法透明度：公开模型算法和决策规则

公开模型的算法细节和决策规则，增强用户对模型的理解和信任。提供模型的算法描述、参数设置和训练过程，使用户了解模型的工作原理。这有助于评估模型的性能和适用性，确保模型在实际应用中的有效性。例如，发布模型的技术白皮书，详细说明模型的架构、训练方法和评估指标，提升算法透明度。

6. 人工审查与监督：定期评估模型决策过程

引入人工审查机制，定期评估模型的决策过程，确保其合理性和公正性。建立专家团队，对模型的预测结果和决策过程进行审查，及时发现并纠正潜在问题。这有助于维护模型的可靠性，防止模型在实际应用中出现偏差或错误。例如，在医疗诊断模型中，医生定期审查模型的诊断结果，确保其准确性和可靠性。

（四）AI 伦理与合规的工程实践

在人工智能的开发和应用过程中，遵守伦理规范和法律法规至关重要。通过制定伦理准则、建立合规框架、实施伦理审查和培训等工程实践，确保 AI 系统的开发和部署符合伦理标准和法律要求。

1. 制定伦理准则：指导 AI 开发的道德方向

制定明确的伦理准则为 AI 开发提供道德指导，确保技术应用符合社会价值观。组织应根据自身业务特点和社会责任，制定 AI 伦理准则，涵盖公平性、透明性、责任性等核心原则。例如，谷歌发布了《AI 原则》，强调 AI 应用应对社会有益，避免偏见，确保安全。这些准则为开发者提供了明确的道德指引，规范 AI 技术的研发和应用。

2. 建立合规框架：确保法律法规的遵守

建立完善的合规框架，确保 AI 系统的开发和部署符合相关法律法规。组织应建立合规管理体系，识别并遵守法律要求。例如，欧盟的《通用数据保护条例（GDPR）》对数据处理提出严格要求，组织需确保 AI 系统在数据收集、处理和存储过程中符合 GDPR 规定，保护用户隐私。

3. 实施伦理审查：评估 AI 项目的道德影响

通过伦理审查机制，评估 AI 项目的道德影响，预防潜在的伦理风险。在 AI 项目启动前，组织应设立伦理审查委员会，对项目进行评估，确保其符合伦理准则。例如，微软设立了 AI 伦理委员会，对 AI 项目进行审查，评估其对社会的潜在影响，确保技术应用的道德性。

4. 开展伦理培训：提升员工的道德意识

通过伦理培训，提升员工的道德意识，确保 AI 开发过程中的伦理合规。组织应定期开展 AI 伦理培训，帮助员工理解伦理准则和法律法规，培养道德判断能力。例如，IBM 为员工提供 AI 伦理培训课程，涵盖公平性、透明性、隐私保护等主题，确保员工在开发过程中遵守伦理规范。

5. 参与行业合作：推动 AI 伦理的行业标准化

通过参与行业合作，推动 AI 伦理的标准化，促进技术的负责任发展。组织应积极参与行业协会和标准化组织的活动，共同制定 AI 伦理标准。例如，IEEE 发布了《Ethically Aligned Design》指南，提供了 AI 伦理设计的框架，组织可参与其中，推动伦理标准的制定和实施。

（五）AI 治理的集成工具与平台

在人工智能的开发和应用过程中，集成化的治理工具和平台对于确保 AI 系统的安全性、

透明性和合规性至关重要。这些工具和平台提供了全面的解决方案，涵盖模型管理、风险评估、合规监控等方面，帮助组织有效地治理 AI 系统。

1. IBM watsonx.governance：全面的 AI 治理平台

IBM watsonx.governance 是一个与现有解决方案无缝集成的平台，帮助组织在 AI 生命周期中实现全面治理，可帮助组织针对生成式 AI 和机器学习模型，加快负责任且可解释的 AI 工作流程。该平台可以管理来自任何供应商的产品，包括 IBM watsonx.ai、Amazon Sagemaker、Bedrock、Google Vertex 和 Microsoft Azure。通过自动执行和扩展端到端 AI 治理，从选择正确的用例到开发、部署、监控和替换模型，watsonx.governance 帮助组织在整个 AI 生命周期中实现全面治理。

2. Fairly AI：专注于 AI 公平性的治理工具

Fairly AI 提供了端到端的 AI 治理解决方案，强调变更管理、偏见检测和合规性，确保 AI 系统的公平性。Fairly AI 平台提供了强大的功能，包括持续监控、政策执行和公平性测试，这些对于维护合乎道德的 AI 实践至关重要。该平台旨在支持法律、审计、风险、合规和数据科学团队，促进协作并减少跨团队摩擦。通过这些功能，Fairly AI 帮助组织在 AI 系统中实现公平性和合规性。

3. 商汤科技的 SenseTrust：可信 AI 基础设施

商汤科技的 SenseTrust 提供了覆盖数据、模型、应用治理环节的可信 AI 治理工具，推动生成式 AI 的可信发展。SenseTrust 包含一套完整的工具体系，可为商汤自身及行业提供伦理、安全二维一体的检测与加固解决方案。该平台提供的数据去毒工具能够检测数据来源中是否包含带有后门、扰乱的有毒数据，并提供去毒方案。此外，SenseTrust 还能够提供“AI 防火墙”，从源头过滤对抗样本，综合检出率达到 98%。通过这些功能，SenseTrust 帮助组织在生成式 AI 的发展中实现可信性和安全性。

4. 蒲公英人工智能治理开放平台：系统支持治理原则落地

蒲公英人工智能治理开放平台（OpenEGLab）致力于打造系统、实用的人工智能伦理与治理基础设施，探索“规则-技术-场景-评测”一体协同的人工智能治理体系。OpenEGLab 目标打造系统、实用的人工智能伦理与治理基础设施，探索“规则-技术-场景-评测”一体协同的人工智能治理体系。该平台发布了包括规则集、治理图谱、风险展示、评测框架和行业方案五大部分，旨在促进人工智能治理理念的落地，服务提升人工智能治理效率，共促人工智能可信、可持续发展。

人工智能作为一项颠覆性技术，正在深刻影响全球经济和社会的发展，其治理问题也日益成为国际社会关注的焦点。尽管不同国家在 AI 监管思路上存在差异，全球合作的必要性不言而喻。为了确保 AI 技术的健康发展和最大化其社会效益，全球各国和地区必须克服地缘政治的壁垒，推动更多的跨国合作与对话。同时，企业在合规与信任的基础上，强化伦理与安全治理，既是对技术创新的保障，也是对社会责任的履行。在此基础上，AI 治理的工程实践和技术工具的不断完善，将为应对复杂的治理挑战提供有力支持。展望未来，全球 AI 治理的框架尚需进一步强化，国际合作与多方协作将是推动这一进程的关键。只有通过共同努力，才能确保人工智能技术的应用真正造福全球社会，推动人类向更加智能、更加公平的未来迈进。