全球人工智能伦理治理的发展特征与基本准则

内容提要：人工智能正通过日渐强大的自动决策能力越来越多地取代人类决策，相应的，人们对其引发的伦理道德问题越发担心。全球人工智能伦理治理尚处于起步阶段，但已就一些基本准则达成共识。本研究通过梳理全球 2015-2020 年间出台的 118 份关于人工智能伦理治理的政策文件，探索全球人工智能伦理治理的发展特征和发展趋势。

一、全球人工智能伦理治理尚处于起步阶段

人工智能以数据为驱动，通过自动决策技术影响和代替人类决策，逐渐对人类的伦理准则形成挑战。各国政府自 2015 年起纷纷将人工智能伦理治理纳入国家人工智能发展战略。美国将人工智能伦理治理列为《国家人工智能研究和发展战略规划》八大战略之一；欧盟委员会在《欧洲人工智能》中强调人工智能治理是人工智能生态体系的重要组成部分；英国在《人工智能发展的计划、能力与志向》中提到要制定国家层面的人工智能治理准则；日本提出要构建有效且安全的“AI-Ready 社会”。但总体来看，国际社会的人工智能伦理治理工作尚处于起步阶段。

在学术研究领域，斯坦福大学以人为本智能研究所开展的统计分析发现，2015 年以来，在提交给国际知名的人工智能会议的论文中，包含“伦理”相关关键词的论文标题数量显著增加，但总数仍然不多（见图 1）。从 2020 年开始，伦理议题才被大家更频繁地讨论。如神经信息处理系统（NeurIPS）会议（世界上最大的人工智能研讨会之一）在 2020 年首次要求研究人员提交文章的同时，也要提交“广泛影响性”声明。

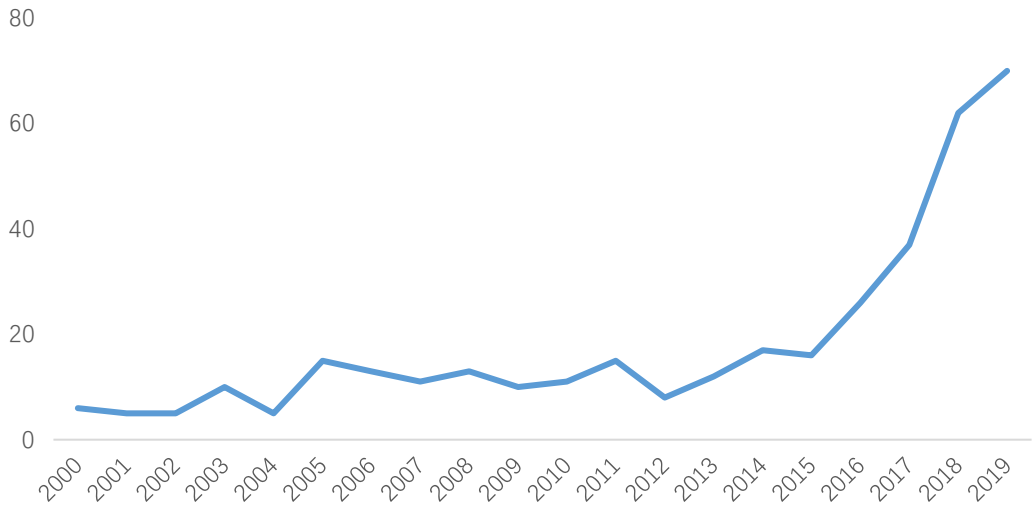

图 1 2000-2019 年国际上主要的人工智能会议上提及“伦理”的论文标题数量变化

资料来源：Prates et al.,2018 [1]，斯坦福以人为本智能研究所（2021）[2]

在实践领域，欧盟、日本、美国等各国政府，微软、SAP、谷歌、Salesforce、IBM 等大型企业，IEEE、ACM 等研究机构以及 OECD、G20 等国际组织近年来纷纷发布人工智能道德规范与人工智能伦理准则（见图 2）。全球 IT 研究与顾问咨询公司 Gartner 于 2020 年发布报告[3]，总结了国际社会五大人工智能伦理准则，即透明、公平、安全可靠、负责任、以人为本。本文对全球 2015-2020 年间出台的 118 份关于人工智能伦理治理的政策文件进行分析后发现，在以上五大准则方面大家已形成共识，除此之外，可持续发展也是国际社会人工智能伦理治理的基本准则之一（见

[1] Prates, M., Avelar, P., & Lamb, L. C. On quantifying and understanding the role of ethics in AI research: A historical account of flagship conferences and journals [J]. arXiv preprint arXiv:1809.08328.,2018.
[2] 斯坦福以人为本智能研究院. 人工智能指数 2021 年度报告[R]. 斯坦福以人为本智能研究院，2021.
[3] Buytendikk F., Sicular S., Brethenoux E., & Hare J. AI Ethics: Use 5 Common Guidelines as Your Starting Point. [R].Gartner.2020

图 3）。

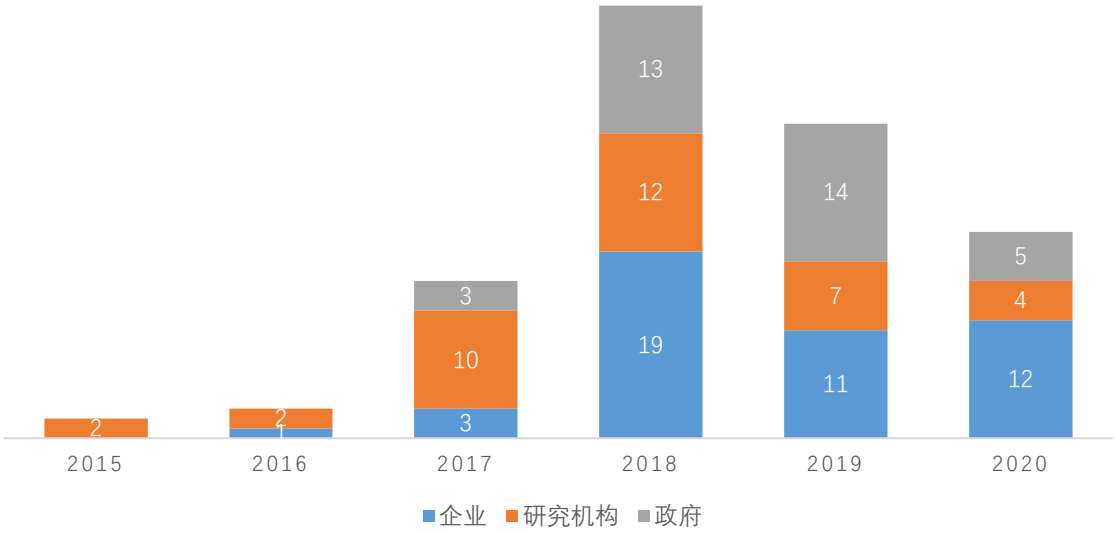

图 2 2015-2020 年按组织类型划分的人工智能伦理准则文件数量

资料来源：各国人工智能伦理治理准则文件总结

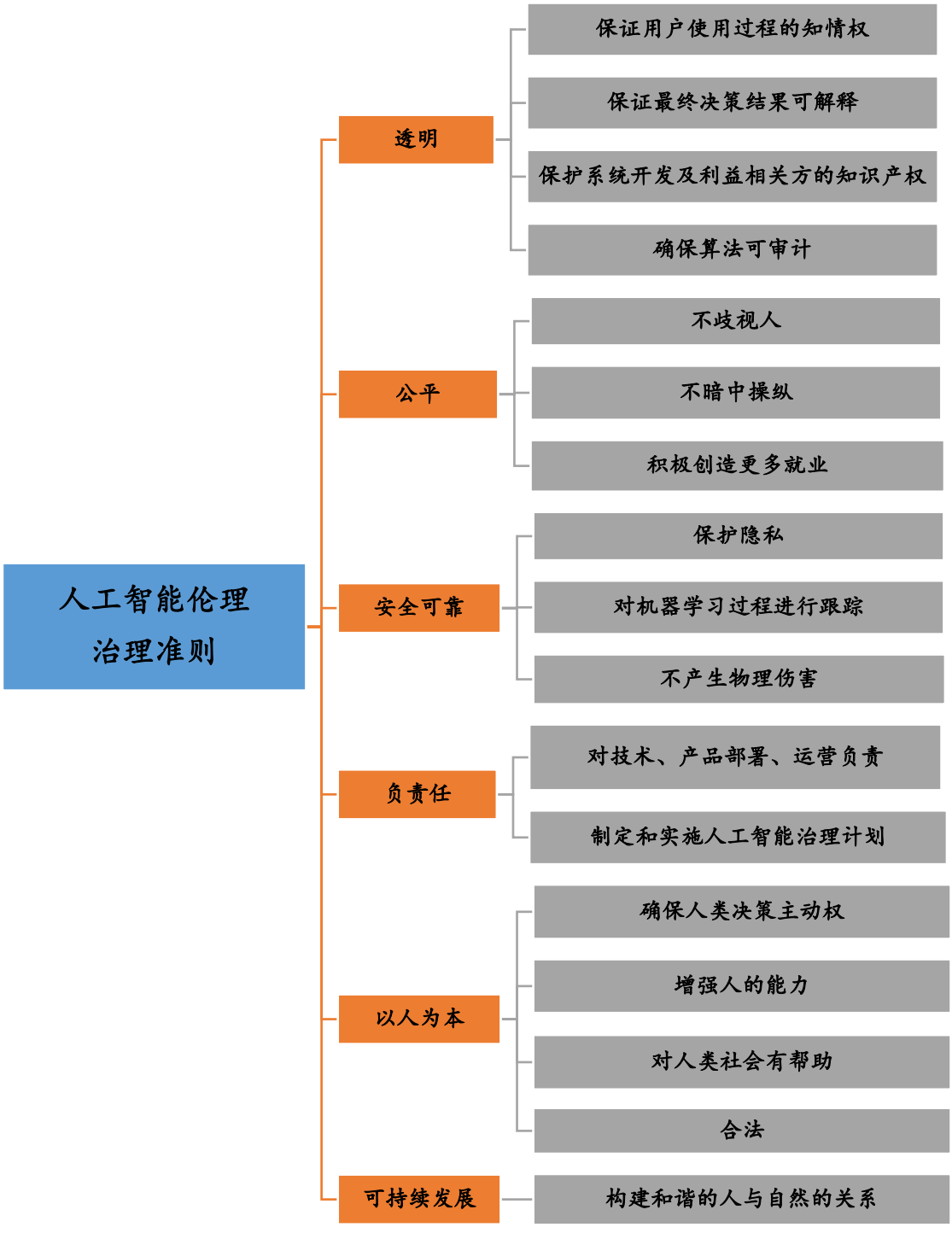

图 3 人工智能伦理治理准则

资料来源：各国人工智能伦理治理准则文件总结

二、全球人工智能伦理治理的六大基本准则

（一）透明（Transparency）

透明性是人工智能伦理治理的首要准则，几乎全部的规范文件中都会提及这一点。人工智能深度嵌入人类决策的过程中，但其复杂性与专业性会加重人与智能系统之间的信息不对称，降低人类对人工智能的安全感、信赖感以及认同感，影响人类的知情权与主体地位。所以透明性原则要求人工智能要保证人类了解自主决策系统的工作原理，从而使结果可预测，进而保障人类的知情权。具体来讲：

第一，要保证用户使用过程的知情权。即要确保用户在知情的情况下与人工智能系统进行交互并且了解人工智能在此过程中的功能。

第二，要保证最终决策结果可解释。虽然人工智能输出的结果往往不容易解释，但可通过对争议结果增加人工测试流程、对自动决策系统设置人工干预流程、加强反驳人工智能决策的假设、设置结果接受范围以及在自动决策系统中设定过滤条件等方法来确保决策结果的可解释性。

第三，透明是有限度的。需保护人工智能系统开发方及利益相关方的知识产权。谨防过度公开引起的诈骗、滥用等风险，健全透明度披露的风险管理办法。

第四，需要对模型、算法、数据以及决策结果进行记录，以备第三方机构审计。

（二）公平（Fairness）

在公平性问题上，人工智能是把“双刃剑”。人工智能的输出结果可实现全局最优，但对个体而言也可能并不“公平”。有时，人工智能会被刻意用来加剧不公平，如特定算法导致种族与性别歧视、大数据杀熟以及窃取个人信息等。此外，一些问题尚存在争议，例如在金融领域，基于用户画像判断发放贷款是否有失公平？由此可见，是否“公平”有时很难定义，需要根据实际情况进行判断。在共同富裕背景下，公平（相对于效率）在我国经济社会发展中的重要性日益提升，人工智能也应尽快将公平性作为约束条件引入目标函数中。目前，针对人工智能的“公平性”，已形成以下基本共识。

第一，不歧视人。MIT 研究发现人工智能可能放大人类的偏好差异，所以在运用技术的过程中要追求实质性公平，避免对特殊的人群或个体造成偏见与歧视，避免让弱势群体处于更不利的地位。

第二，不暗中操纵。人工智能系统的出现，根本上是为了让人类生活变得更美好，而不应该用其数据抓取与分析能力进行不合理推荐，也不应造就“知识茧房”。

第三，会造成结构性失业，但也应积极创造更多就业。从机械臂到机器人到自动化生产，人工智能替代人类劳动的情况层出不穷。2019 年初布鲁金斯研究中心的研究报告指出，约 3600 万美国人面临被人工智能替代就业的危险。各国政府近年来纷纷出

台相关规定，推动人工智能教育及人才培养。美国 2018 年发布《人工智能就业法案》，提出要营造终身学习和技能培训环境，应对人工智能对就业的挑战；英国 2018 年发布《英国发展人工智能的计划、意愿和能力》，提到要重视人工智能专业人才的培养，加强对公民的再培训。

（三）安全可靠（Security & Safety）

安全可靠是新技术普遍面临的伦理问题，核心是确保人工智能不对人类造成物理和心理上的伤害。其中：

第一，隐私保护是人类建立与人工智能之间信任感的基础。一方面要确保技术的可靠性与稳定性，监控人工智能机器学习过程，防止系统被恶意攻击使用户受到网络骚扰，导致信息泄露以及其他更严重的数字犯罪等；另一方面要合理搜集和使用数据，确保敏感数据加密且不用于其他目的。

第二，为保证系统不被恶意攻击，还需要对机器学习进行全程跟踪并不断测试，尤其是在发生异常情况时。

第三，当人工智能技术与机器人技术结合时，要确保可以无误差地预测风险，以确保机器与系统不会对人类造成物理伤害。

对安全可靠准则的监管已较成熟，尤其在数据治理方面。欧盟 2018 年通过《通用数据保护条例》，明确个人对自身数据的控制权。随后美国、日本、巴西、新加坡等国家也纷纷出台或修订个人数据保护有关法律。我国今年 11 月 1 日正式生效的《个人信息保护法》，对个人信息的收集和使用进行了严格的规定，要

求收集个人信息要“告知-同意”，且符合最小必要原则，并列举了用户的权益清单。

（四）负责任（Accountability）

人工智能基于数据和算法形成判断并做出选择，具有自动性但不具有自主性，具有行动能力但不具有行动意识、思维能力。人工智能本质上仍是人类为了达成某种目标而制造的工具，其是否具有主体地位仍存在争论。围绕人工智能的责任问题，实质仍是创造它的人的责任问题。当前大多数的治理准则中建议利益相关者制定并实施人工智能治理计划，加深对责任主体的研究。

（五）以人为本（Human-Centric）

以人为本是人与人工智能关系的根本原则。要求人工智能的主要目的是为人类服务，支持人类的目标，促进人类社会繁荣，增加人类社会的福祉。

第一，要确保人类决策的自主权。虽然人工智能可以代替人做决策，但是最终人工智能需要受到人类控制。

第二，人工智能的定位是补充人类尚不完美的行动能力，任何只注重效率而侵犯人类权利的技术都是违反伦理的，不宜提倡和鼓励。

第三，人工智能应该对人类社会有帮助，并对人类社会的重大问题做出积极贡献。

第四，人工智能的使用必须要符合人类社会的基本准则，即合法。需要保证开发和使用的合法性，遵循当地法律规范。

（六）可持续发展（Sustainability）

新冠疫情的爆发让社会更加关注人与自然的关系，人工智能的发展也要为此服务。可持续发展要求企业在人工智能的部署过程中考虑对全球生态变化以及生物多样性的影响；在人工智能的设计过程中以提高能源使用效率，减少生态足迹为目的。

三、我国人工智能伦理治理准则已与国际社会全面接轨

近年来，我国人工智能伦理治理体系不断完善。2017 年，国务院发布《新一代人工智能发展规划》，明确提出人工智能治理“三步走”的战略目标：2020 年部分领域的人工智能伦理规范和政策法规初步建立；2025 年初步建立人工智能法律法规、伦理规范和政策体系；2030 年建成更加完善的人工智能法律法规、伦理规范和政策体系。国家新一代人工智能治理专业委员会于 2019 年 6 月发布《新一代人工智能治理原则——发展负责任的人工智能》，提出人工智能治理的框架和行动指南，强调发展负责可信的人工智能八项治理原则：和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理；后又于 2021 年 9 月发布《新一代人工智能伦理规范》，深化治理原则，明确提出人工智能各类活动应遵循增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等六项基本伦理规范。整体来看我国的人工智能伦理准则已与国际社会全面接轨。

美团一直非常重视人工智能伦理治理工作。今后，美团一方

面会持续加大在科技创新方面的投入，另一方面也将积极与政府部门、行业组织、科研机构等社会各界一道，共同探讨和推进人工智能伦理治理工作，让人工智能技术更好服务经济社会发展。

美团研究院　刘婉莹 厉基巍 赫建营