CAICT 中国信通院　京东探索研究院 JD EXPLORE ACADEMY

No.202106

可信人工智能白皮书

中国信息通信研究院
京东探索研究院
2021年7月

前　言

当前，新一代人工智能技术迅猛发展，并向社会各个领域加速渗透，给人类生产生活带来了深刻变化。人工智能在带来巨大机遇的同时，也蕴含着风险和挑战。习近平总书记在 2018 年 10 月主持中央政治局第九次集体学习时强调，“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控”。增强人工智能使用信心，推动人工智能产业健康发展已经成为重要关切。

发展可信人工智能正在成为全球共识。2019 年 6 月，二十国集团（G20）提出“G20 人工智能原则”，强调要以人为本、发展可信人工智能，这一原则也得到了国际社会的普遍认同。欧盟和美国也都把增强用户信任、发展可信人工智能放在其人工智能伦理和治理的核心位置。未来，将抽象的人工智能原则转化为具体实践，落实到技术、产品和应用中去，是回应社会关切、解决突出矛盾、防范安全风险的必然选择，是关系到人工智能长远发展的重要议题，也是产业界急需加快推进的紧迫工作。

无论是回顾可信人工智能的背景和历程，还是展望新一代人工智能的未来，本白皮书认为人工智能的稳定性、可解释性、公平性等都是各方关注的核心问题。立足当下，本白皮书从如何落实全球人工智能治理共识的角度出发，聚焦于可信人工智能技术、产业和行业实践等层面，分析了实现可控可靠、透明可释、隐私保护、明

确责任及多元包容的可信人工智能路径，并对可信人工智能的未来发展提出了建议。

由于人工智能仍处于飞速发展阶段，我们对可信人工智能的认识还有待进一步深化，白皮书中存在的不足之处，欢迎大家批评指正。

目 录

图 目 录

表 目 录