全球人工智能安全治理

The Global Governance of Artificial Intelligence: Safety and Security Perspective

SAISD
上海市人工智能与社会发展研究会
Shanghai Association for AI and Social Development

360天枢智库
360 DIPPER RESEARCH

天津智慧城市数字安全研究院

RCGGC
网络空间国际治理研究中心
RESEARCH CENTER FOR GLOBAL GOVERNANCE OF CYBERSPACE

全球人工智能安全治理

The Global Governance of Artificial Intelligence: Safety and Security Perspective

邹权臣 鲁传颖 主编

360 天枢智库
上海国际问题研究院网络空间国际治理研究中心

联合发布

参与单位

天津智慧城市数字安全研究院
360 AI 安全实验室
大数据协同安全技术国家工程研究中心
中国工商银行业务研发中心
中信基金会金融实验室
中国电信研究院
上海市人工智能与社会发展研究会
上海赛博网络安全产业创新研究院

2022 年 6 月

前言

《全球人工智能安全治理》是 360 天枢智库与上海国际问题研究院网络空间国际治理中心合作发表的首份联合研究报告，是国内学术界和产业界在人工智能安全领域的联合研究成果。双方希望在人工智能技术快速发展过程中，探索出既能发挥技术效益，又能控制其安全风险的治理之道。

安全治理是决定人工智能未来发展的关键领域。双方研究团队从政策制定和产业实践的角度，分析了人工智能安全治理面临的挑战、治理模式和行业实践。我们看到，人工智能作为中、美、欧等国家或地区都在积极发展的关键新兴技术，其在发展过程中所产生的安全挑战也更为复杂多元。在政策层面，世界主要国家和地区已经将安全治理列为人工智能战略的优先事项。在行业领域，作为人工智能技术的开发者，企业是人工智能安全治理的重要参与者，通过参与标准制定，推出产品服务，努力探索人工智能安全治理的技术路径和解决方案。包括 360 在内的多家国内外企业已经通过自身实践走出了技术赋能、行业规制、平台监测的多种道路。

人工智能安全是数字时代的基础性安全领域。人工智能安全治理，当前没有成熟方案，只有最佳实践，需要广泛协作和开放创新。为应对人工智能与安全面临的挑战，360 发挥龙头企业优势，践行社会责任，承担了国家级开放创新平台的建设任务。例如，360AI 安全实验室自 2019 年承担建设的“安全大脑国家新一代人工智能开放创新平台”，致力于解决人工智能自身存在的安全风险、人工智能创新环境匮乏及在网络安全领域创新成果落地难等问题。同时，该平台还将赋能中小初创

安全企业、垂直行业、人工智能产业，提升国家人工智能安全整体防御能力，打造人工智能的“安全底座”。我相信，人工智能安全产业发展必将助力人工智能成为集技术革新与安全可控于一身的战略性技术。

值得一提的是，《全球人工智能安全治理》报告得到多家单位的参与和协助，汇集了国内产、学、研等多方力量的支持。例如，天津智慧城市数字安全研究院、360 AI 安全实验室、大数据协同安全技术为研究团队提供了一手的研究资料；中国工商银行业务研发中心、中信基金会金融实验室、中国电信研究院、上海市人工智能与社会发展研究会、上海赛博网络安全产业创新研究院为研究团队提供了难得的调研机会。在此，向上述单位表示诚挚的感谢，相信通过政策界、学术界和企业界的紧密结合，秉持“科技向善”理念，人工智能技术一定会持续造福人类。

当然，我们的研究成果也还存在不足之处，欢迎读者批评指正，也欢迎更多人加入我们的联合研究。

360 集团副总裁兼首席安全官　杜跃进博士

2022 年 6 月

目录

摘要

全球人工智能安全治理……1

一、安全治理成为人工智能战略的优先事项 /1

（一）美国人工智能安全治理……3

（二）欧盟人工智能安全治理……3

（三）中国人工智能安全治理……5

二、人工智能安全治理面临的挑战 /6

（一）人工智能自身安全面临的挑战……6

（二）人工智能衍生安全面临的挑战……8

（三）人工智能赋能网络安全的机遇与风险并存 ……10

案例一：人工智能对数字城市建设带来的挑战……11

三、人工智能的风险治理 /13

（一）人工智能风险治理思路……13

（二）人工智能风险治理模式……15

（三）人工智能域外治理经验……18

四、人工智能安全治理的行业实践 /22

（一）技术解决方案……22

（二）行业标准……24

（三）应用案例……25

五、总结 /28

摘要

安全治理是事关人工智能未来发展的关键。目前，各国政府在制定人工智能发展战略时都已经将安全治理作为不可或缺的组成部分。在全球范围内，针对人工智能的伦理规范、风险框架，以及治理理念和模式的探索成为学术界和政策界的重点工作。

一、人工智能面临的安全风险

目前来看，全球范围内普遍关注的人工智能安全问题有以下十类：网络安全问题、企业合规问题、可解释性问题、隐私安全问题、声誉和伦理问题、未来岗位问题、公平性问题、人身安全问题、国家安全问题、政治稳定问题等 。在当下的人工智能技术发展阶段中所出现的争议基本包含在以上十个方面的问题之内。这十类人工智能安全可从三个方面来理解：

第一，人工智能自身安全，也称人工智能系统安全，主要包括 AI 框架等基础设施安全、算法安全，以及训练数据集、模型文件的安全防护等。这其中，算法安全尤其值得关注。算法设计或实施有误可能产生与预期不符甚至伤害性结果。

其中有三个风险值得关注 : 第一是“透明度”问题，指拥有决策算法的公司或个人通过设置“算法黑箱”而导致的监督审查缺失，这不仅会带来安全风险，更有悖于人工智能的可解释性原则，甚至会造成更高层面上的政治合法性问题。第二是“偏见”问题，指有偏见的数据集和决策规则导致人工智能的训练结果存在偏误。

第三是“自动化”问题，指算法作为治理工具的普遍存在引起了人们对人类能动性和自主性影响的担忧 。

第二，人工智能衍生安全。人工智能技术的应用将对经济、政治、军事、社会等领域产生重大冲击。在军事方面，随着人工智能技术的成熟 , 它将会被越来越广泛地应用于武器系统 , 人类社会可能在进入人工智能时代之后迎来一个不同的军事安全环境。在政治方面，人工智能技术及其背后的大数据和算法能够潜移默化地影响人类行为，直接对国内政治行为产生干扰，甚至影响国际竞争的内容与形态。在经济方面，在人工智能技术的影响下，资本与技术在经济活动中的地位获得全面提升，而劳动力要素的价值则受到严重削弱，由此引发结构性失业风险、贫富分化和不平等现象。

第三，人工智能赋能对网络安全引发的机遇与风险。随着人工智能的发展，这种基于海量数据训练而提供实时监控的技术似乎为解决网络安全问题开辟了新的通道。另一方面，人工智能如果被用于网络安全进攻，也会给网络安全带来新的挑战。

二、人工智能安全治理模式

面对如此复杂的问题，人工智能安全不会一蹴而就。

首先，要建立科学有效的风险评估和识别模式。一是基于未来风险预防的影响评估（impact assessment），即通过多利益攸关方的咨询、参与和审议，针对可能具有重大安全风险的技术内容，在付诸实践前先进行影响评估，对其可能产生的危害进行预防性监管。二是元监管（meta-regulation），即政府等监管机构并不为企业设置严格的合规框架，而是由企业自行论证开发活动的合规性，并由政府机构进行调查和处罚。三是 AI 系统警戒（AI system vigilance），即通过系统性的警戒以及对不良事件的透明跟踪，尽早发现问题和故障并及时进行纠错。

其次，需要在治理模式上进行创新。一是注重参与性设计模式，或称参与性治理，其核心是将利益相关者（如预期的最终用户）纳入设计过程，与专业设计师和研究人员一起工作，并参与决策。二是敏捷治理模式，其特点是通过系统性整合，与各利益攸关方共同组成一体化的人工智能治理生态，并通过灵敏、及时、持续的“咨询 - 反馈”机制，促进治理政策的迭代升级，以弥补政府治理中的信息的滞后性，形成对人工智能风险的前瞻性评估与治理。在整个敏捷治理的思路中，包括三个要素——两条“咨询 - 反馈”路径和一个“动态评估”机制，一起致力于治理政策更新。

三、人工智能产业发展与安全治理创新

总体上，世界主要国家和地区在政策层面将人工智能产业列为战略性新兴产业，加大支持力度。企业既是人工智能技术的开发者，也是人工智能安全治理的重要参与者。它们积极参与标准制定，推出安全治理产品服务，这些努力既探索了人工智能安全治理的创新路径，也拓展了产业持续发展的空间。

未来，政府、科技公司、学术机构和用户等所有利益相关者要联合起来，秉承构建人类命运共同体的理念，建立积极的生态规则，加强多方互动合作，构建一个开放的人工智能生态系统，用“科技向善”理念引领人工智能技术造福人类。

全球人工智能安全治理

人工智能是引领未来的战略性技术，正在对经济发展、社会进步和人类生活产生深远影响，各国均在战略层面上予以高度关注。作为一种数字技术，人工智能自身存在数字安全威胁和隐患。而且，随着人工智能工程化、场景化、平台化落地不断加快，人工智能安全需求早已超越单纯技术范畴。因此，人工智能安全治理成为世界主要国家和地区人工智能战略的优先事项，各方希望在技术升级的过程中，找到既能发挥人工智能技术效益，又能控制其安全风险的治理之道。

一、安全治理成为人工智能战略的优先事项

安全治理是各国人工智能战略中的优先议题，各主要国家希望在推动人工智能发展时，能够避免其所带来的安全风险和挑战。根据经济合作与发展组织（OECD）统计，全球已经有 61 个国家的中央政府发布了 444 项人工智能发展规划[1]。随着人工智能发展战略部署的不断落地，主要大国从自身产业发展需要出发，逐步将政策布局重点转向技术伦理和法规的建设。美国、欧盟、中国等较早制定人工智能战略的国家和国家联合体，在安全治理上开展了各种探索。

[1] OECD,“OECD AI’s Live Repository of over 260 AI Strategies & Policies,” https://oecd.ai/en/dashboards.

表 1-1 主要国家和国家联合体人工智能战略

文件名称	核心内容	发布时间	发布机构
《人工智能伦理原则》	负责任、公平、可追溯、可靠、可控，呼吁国防部增加对人工智能研究、培训、 道德评估的投入。	2019 年 10 月	美国国防部
《人工智能应用规范指南》	公众信任人工智能、公众参与、科学诚信与信息质量、风险评估与管理、灵活性、收益和成本、公正和非歧视、公开透明、安全和保障、机构间协作。	2020 年 1 月	美国白宫
《可信赖的人工智能伦理指南》	遵守法律和伦理道德、尊重人类自由自治、受人监管、避免伤害、保证公平、 稳定可靠、保护隐私、透明可释、可审核评估、可问责、确保社会福祉。	2019 年 4 月	欧盟人工智能高级别专家组
《英国发展人工智能计划、意愿和能力》	确保人类共同利益、保证公平、容易理解、保护隐私、普及教育、避免伤害欺骗人类。	2018 年 4 月	英国上议院
《日本人工智能学会伦理准则》	贡献人类、遵守法律、尊重隐私、公平公正、保证安全、肩负社会责任。	2017 年 2 月	日本人工智能伦理委员会
《自动驾驶伦理准则》	保证交通参与者安全、驾驶系统需要官方批准监管、禁止将人群属性作为评价标准、禁止量化生命价值、责任共担。	2017 年 8 月	德国交通与数字基础设施部
《法国人工智能发展计划》	算法透明、责任承担、成立人工智能伦理委员会、组织伦理公共辩论。	2018 年 3 月	法国政府

表格来源：根据 OECD 人工智能国家政策与战略网站整理。

（一）美国人工智能安全治理

美国在人工智能安全治理上采取的手段是在人工智能技术部署、使用与监测的全过程中都进行验证与监管，建立与之配套的规范体系。

在部署阶段，强调提高人工智能的可解释性与透明度，减少由于技术障碍而导致的决策不准确问题，使不了解技术的工作人员可以知晓其工作方式与决策过程，从而提出改进建议。同时，建立可信任的输入数据库，降低人工智能的数据决策偏差。

在使用阶段，强调符合可验证性与可确认性，满足正式规范与用户的操作需求，让广泛复杂的人工智能系统以用户可见的方式进行操作，以用户可接受的形式输出，按用户的期望执行，形成透明、可信、可靠的人工智能交互方式。

在监管阶段，强调建立针对性的开发规范机制与评估方法[2]，以合适的方式对人工智能进行验证。此外，强调技术的持续更新，即通过自我监测、限制策略以及价值学习来实现人工智能的安全与优化，建立可审计、可恢复的人工智能系统。在此基础上应对可能存在的噪声污染与“对抗机器学习”，防止他国通过“污染”训练数据、修改算法等手段阻碍人工智能正确识别某一目标，从而达到危害人工智能系统的目的[3]。

（二）欧盟人工智能安全治理

相较于美国，欧盟更寄希望于运用监管框架与信任体系来对人工智能的安全进行规制，而其规制也更倾向于人权方向。欧盟认为欧洲的人工智能需要在七大方向做出关键改变，并针对七大方向提出了独立的发展措施。

---

[2] Networking & Information Technology Research and Development Subcommittee and The Machine Learning & Artificial Intelligence Subcommittee of The Nation Al Science & Technology Council, “Artificial Intelligence And Cybersecurity: Opportunities And Challenges,” 2020, p.1-4, https://www.nitrd.gov/pubs/ai-cs-tech-summary-2020.pdf.

[3] National Science and Technology Council, “The National Artificial Intelligence Research And Development Strategic Plan,” 2016, p.27-30, https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.

表 1-2 欧盟人工智能战略主要要求及措施

可信赖人工智能的关键要求	针对性措施
人类对人工智能的代理与监督	确保人工智能不会破坏人类的自主性
技术稳健性和安全性	整合安全设计机制
数据隐私和治理	在使用高质量人工智能系统的同时保证隐私和数据保护
决策透明度	要求人工智能系统可追溯
多样性、非歧视和公平	建立多样化的设计团队并建立机制，确保公民参与
社会和环境福祉	鼓励人工智能系统的可持续性和生态责任
人工智能问责制	建立机制，以确保人工智能系统对其结果负责，并可被问责

表格来源：根据欧盟《建立以人为本的人工智能信任体系》（Building Trust in Human-Centric Artificial Intelligence）内容整理而成。[4]

2020 年 2 月，欧盟发布《人工智能白皮书——欧洲追求卓越和信任的方法》，提出了基于风险的监管方法。人工智能应用程序在满足两个条件时将会被视为“高风险”——在应用于某一行业预期会发生重大风险事件和人工智能的使用方式存在重大风险。除此之外，当涉及员工权利与侵入性监视技术时，将始终被视为“高风险”行为，适用该等级的监管方法。具体而言，基于风险的监管方法主要关注训练数据、数据的记录保存方式、应用所要求提供的信息、人工智能的准确性以及人为监管力度，某些特定人工智能应用还会有相对应的例外要求。这一方法在充分保护人工智能信任

[4] European Commission, “Building Trust in Human-Centric Artificial Intelligence,” 2019, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168&qid=1650694295419.

体系的同时不至于过于严格，以免对企业造成不必要的负担[5]。

同时，欧盟认为人工智能的出现使得法律在执行力度、适用性、经营者责任分配、安全概念方面存在不足，进而对现有人工智能立法框架做出两方面的调整。一是目前的产品安全立法应该扩展到防止产品所产生的各种风险，要在保护用户的基础上鼓励创新。二是在监管框架方面，设置更为灵活的监管框架，使其能够应对技术的实时更新，并提供必要的法律确定性。

（三）中国人工智能安全治理

2017 年 7 月 8 日，中国国务院发布了《新一代人工智能发展规划》。2019 年，中国发布《新一代人工智能治理原则——发展负责任的人工智能》，明确了人工智能的治理框架和行动指南。中国的人工智能安全治理力图形成内含研发、管理和应用的全流程安全保障机制[6]，涵盖基础框架研制、基本安全原则、供应链管理实践指南、安全服务能力、应用领域的标准研制等方面。在网络安全领域构建了人工智能安全威胁分类体系，制定面向人工智能系统安全性的评估标准体系，将可解释性、隐私性等多方面安全要素纳入考虑之中，以应对可能存在的攻击与污染[7]。同时，还将人工智能的技术要点纳入了安全治理之中。针对算法、数据与模型都做出了明确的规定，要求开展匿名用户数据使用管理、人工智能数据安全、人工智能数据标注安全、人工智能算法模型可信赖等方向的规制。

中国也同样重视人工智能的社会伦理影响。2021 年 9 月发布的《新一代人工智

[5] European Commission, “WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust,” 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1650694043168.

[6] 国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部：《国家新一代人工智能标准体系建设指南》，2020 年 7 月。

[7] 中华人民共和国工业和信息化部：《网络安全产业高质量发展三年行动计划（2021-2023 年）（征求意见稿）》，2021 年 7 月，https://www.miit.gov.cn/cms_files/filemanager/1226211233/attach/20217/0e5071815ec641be9e2154566c09fe33.wps.

能伦理规范》强调，将伦理道德融入人工智能全生命周期，促进公平、公正、和谐、安全，避免偏见、歧视、隐私和信息泄露等问题，为从事人工智能相关活动的自然人、法人和其他相关机构等提供了伦理指引。[8]

二、人工智能安全治理面临的挑战

人工智能本身仍然是一种数字技术，“双刃剑”特征明显。国内外在竞相推动人工智能技术发展的同时，也在高度重视其可能蕴含的各种安全风险。归纳起来，人工智能安全涵盖三个方面，包括人工智能自身安全、人工智能衍生安全、人工智能赋能安全。

（一）人工智能自身安全面临的挑战

人工智能自身安全，也称人工智能系统安全，主要包括机器学习框架等基础设施安全、算法安全、以及训练数据集、模型文件的安全防护等。

人工智能基础设施安全风险主要表现为开发者的恶意或不完全建设、产品服务体系不成熟、应对能力不足、全球基础设施能力建设不平衡等问题。以实现人工智能算法时所依赖的机器学习框架为例，其安全风险具体表现在框架漏洞和第三方供应链安全问题上。谷歌开发的 TensorFlow 框架是目前被使用最多的机器学习框架之一，其下载量已达到数亿级别，据其官方代码仓库显示，目前已经有数百个安全问题被发现与修复[9]，相关的安全问题直接影响大量的智能服务开发者与智能服务用户。因特

[8] 中华人民共和国科学技术部：《新一代人工智能伦理规范》，2021 年 9 月 26 日，http://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html .

[9] https://github.com/tensorflow/tensorflow/blob/master/tensorflow/security/README.md

尔开发的 OpenCV（开源计算机视觉）框架支撑了谷歌、雅虎、微软等企业的人脸识别技术开发，但在 4.1.0 版本中却被发现有两个缓冲区溢出漏洞。[10] 由于这些漏洞附着于最底层的基础设施框架之上，因此对后续的研发和产业化环节的影响极大，必须进行快速的检测与修复。此外，还有 GPU、NPU 等硬件领域以及云平台方面的安全风险也都隶属于人工智能的基础设施安全。

在算法安全方面，算法设计或实施有误可能产生与预期不符甚至伤害性结果。其中有三个风险值得关注。第一是“透明度”问题，指拥有决策算法的公司或个人通过设置“算法黑箱”而导致的监督审查缺失，这不仅会带来安全风险，更有悖于人工智能的可解释性原则，甚至会造成更高层面上的政治合法性问题。第二是“偏见”问题，指有偏见的数据集和决策规则导致人工智能的训练结果存在偏误。比如说 MIT 研究员与微软科学家对微软、IBM 和旷视科技三家公司的人脸识别系统进行测试，发现其针对白人男性的错误率低于 1%，而针对黑人女性的错误率则高达 21%-35%。[11] 第三是“自动化”问题，指算法作为治理工具的普遍存在引起了人们对人类能动性和自主性影响的担忧 。[12]

人工智能需要大量数据集作为资源训练机器学习算法，给数据隐私带来了风险。人工智能的“数据饥渴”与人类的“隐私警惕”之间存在着不可避免的张力，人工智能技术是否会危害到数据和隐私安全也已成为社会公众普遍担心的问题。在微软的一项调查中，41% 的受访者表示不信任智能语音助理，并认为他们通过实时的被动声音收集损害隐私，约 52% 的人表示他们担心自己的个人信息不安全。[13]

[10] “OpenCV XML Persistence Parser Buffer Overflow Vulnerability,” Talos Vulnerability Report, 2020, https://talosintelligence.com/vulnerability_reports/TALOS-2019-0852.

[11] 中国信息通信研究院安全研究所，《人工智能安全白皮书 2018》，2018 年 9 月。

[12] Hildebrandt, Mireille, “The New Imbroglio – Living with Machine Algorithms,” 2016, p.55–60, https://doi.org/10.25969/mediarep/13395.

[13] Microsoft Advertising, “The 2019 Voice Report,” 2019, https://about.ads.microsoft.com/en-us/insights/2019-voice-report.

随着各国间数据交流深化，联邦学习、迁移学习等人工智能新技术应用，跨机构间人工智能研发协作进一步增多。特别是近年来对抗样本攻击、算法后门攻击、模型窃取攻击、模型反馈误导、数据逆向还原、成员推理攻击等新型安全攻击技术的快速涌现，个人隐私变得更易被挖掘和暴露。[14] 令 Facebook 市值蒸发 360 多亿美元的数据泄露事件主角——剑桥分析公司，就是通过关联分析方式获得了海量美国公民用户信息，借此实施各种政治宣传和非法牟利活动。2021 年 11 月 2 日，Facebook 公司宣布计划在当月关闭其已有 10 年历史的人脸识别系统，并删除超过 10 亿用户的面部扫描数据。[15]

（二）人工智能衍生安全面临的挑战

人工智能技术的应用将对经济、政治、军事、社会等领域产生重大冲击，从而导致技术发展的“科林格里奇困境”。[16] 如何在潜在的动荡风险尚未形成时做到未雨绸缪，使技术革命的影响不至于反噬人类本身，对人工智能未来发展至关重要。“人工智能应用正改变现有威胁，创造新的威胁类型，其可被用于社会渗透、数据窃取、破坏稳定等活动，加深了网络攻击和虚假信息活动构成的威胁。”[17]

在军事方面，“随着人工智能技术的成熟，它将会被越来越广泛地应用于军事领域，武器系统、军事策略、军事组织，甚至战争的意义可能会发生深刻改变，人类社

[14] 中国信通院安全研究所：《人工智能安全框架》，2020 年 12 月，http://www.caict.ac.cn/kxyj/qwfb/ztbg/202012/P020201209408499730071.pdf.

[15] Hill, Kashmir, and Ryan Mac, “Facebook, Citing Societal Concerns, Plans to Shut Down Facial Recognition System,” The New York Times, November 2, 2021, https://www.nytimes.com/2021/11/02/technology/facebook-facial-recognition.html.

[16] “科林格里奇困境”（Collingridge dilemma）意指一项技术的社会后果不能在技术生命的早期被预料到。然而当技术产生不良后果时，它往往已经成为了整个经济和社会结构中难以剥离的一部分，以至于难以对它进行控制。Collingridge, D, “The Social Control of Technology,” Milton Keynes, UK: Open University Press, 1980,pp.16-17; 转引自文成伟、汪姿君：《预知性技术伦理消解 AI 科林格里奇困境的路径分析》,《自然辩证法通讯》 第 43 卷，2021 年第 4 期，第 10 页。

[17] 秦浩：《美国政府人工智能战略目标、举措及经验分析》,《中国电子科学研究院学报》2021 年第 12 期，第 1243-1250 页。

会也有可能在进入人工智能时代之后迎来一个不同的军事安全环境。”[18] 作为人类科技史上最新的力量放大器，人工智能在军事领域已经展现出明显超越人类的能力与持续发展的潜力。面对这样的技术变革浪潮，所有具有相应技术基础的大国必然会千方百计地获取相关技术，一场以人工智能技术为核心的新的军备竞赛恐怕很难避免。此外，人工智能技术的介入，使大量无人作战武器参与作战成为可能，而完全自主性武器的广泛应用将带来巨大的军事伦理问题。

在政治方面，人工智能技术及其背后的大数据和算法能够潜移默化地影响人类行为，直接对国内政治行为产生干扰，甚至影响国际竞争的内容与形态。以人脸识别和深度造假技术为例，在美国 2020 年总统选举之前，一段众议院议长南希·佩洛西的演讲视频被深度造假技术所调整，这被视为对民主党领导人的抹黑。而更值得一提的是当发现这一视频是伪造后，美国各网络平台对此反应不一——YouTube 及时对视频进行了下架，但 Facebook 拒绝对此视频进行删除 。[19] 在这一事件中可以看到，对个人的生物信息识别涉及到隐私安全问题、在互联网社交媒体上传播捏造的信息又涉及网络安全问题和声誉问题、而针对政府领导人的行动又可能会危害到政治稳定和国家安全、在与互联网平台企业交涉时需要关注合规问题。

在经济方面，在人工智能技术的影响下，资本与技术在经济活动中的地位获得全面提升，而劳动力要素的价值则受到严重削弱，由此引发结构性失业风险、贫富分化和不平等现象。更进一步，人工智能技术带来的全球经济结构调整，将引导全球资本和人才进一步流向技术主导国家，由此留给发展中国家走上现代化道路的机遇期将变得更加有限。[20]

[18] 封帅，鲁传颖：《人工智能时代的国家安全：风险与治理》，《信息安全与通信保密》2018 年第 10 期，第 36 页。

[19] “Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread across Social Media,” Washington Post, May 24th 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/.

[20] 封帅，鲁传颖：《人工智能时代的国家安全：风险与治理》，《信息安全与通信保密》2018 年第 10 期，第 33 页。

当人工智能技术所推动的社会经济结构变革逐步深入时，资本和技术力量的垄断地位有可能结合在一起，在一定程度上逐渐分散了传统上由民族国家所掌控的金融、信息等重要的权力。[21] 例如，大型企业对于数据资源以及人工智能技术的控制能力正在造成其实际上的垄断状态。而这种垄断将深嵌于数字时代的方方面面，包括利用算法的黑箱为大众提供他们希望看到的内容，潜移默化地改变公共产品的提供方式。

（三）人工智能赋能网络安全的机遇与风险并存

随着人工智能的发展，这种基于海量数据训练而提供实时监控的技术似乎为解决网络安全问题开辟了新的通道。同时，人工智能如果被用于网络安全进攻，也会给网络安全带来新的挑战。

根据当前人工智能的技术特征和发展状况，人工智能被寄予厚望，期待在促进网络安全方面有以下五方面的功能：1、识别系统漏洞以提高系统的可信度；2、开展自动化网络行动和网络防御以平衡攻防态势；3、实现网络安全情报的实时收集和分析；4、自动化决策；5、人机交互。[22] 人工智能的这些功能不仅能够助力网络安全难题的解决，还能够进一步提升人们与网络空间的融合程度、挖掘网络的价值。

但人工智能技术的误用也会加剧网络风险，带来更深层次的安全威胁。其中一个重要的新问题就是生成对抗性神经网络，这种技术可以避开人工智能的监测系统，依据使用者的意图输出任意的结果。[23] 而且人工智能技术的发展和应用也依赖网络安全的保证。试想如果在未来，自动化武器的程序被黑客入侵并修改攻击对象，那么将造成灾难性的后果。因此，网络安全与人工智能的技术安全密不可分。

---

[21] 封帅，鲁传颖：《人工智能时代的国家安全：风险与治理》，《信息安全与通信保密》2018 年第 10 期，第 34 页。

[22] 参考“Artificial Intelligence and Cybersecurity: A Detailed Technical Workshop Report,” The Networking & Information Technology R&D Program (NITRD), 2020, https://www.nitrd.gov/pubs/AI-CS-Detailed-Technical-Workshop-Report-2020.pdf.

[23] S. Mathew Liao, Ethics of Artificial Intelligence, Oxford University Press , 2020, p.221.

案例一：人工智能对数字城市建设带来的挑战

从社会发展的角度看，目前人工智能技术的应用领域主要集中在政府城市治理和运营。根据 IDC 的数据，2020 年全球智慧城市投资规模达到 1,144 亿美元，中国智慧城市投资规模达到 241 亿美元，预计 2024 年超过 450 亿美元。[24]2020 年，中国人工智能市场主要客户来自政府城市治理和运营 (公安、交警、司法、城市运营、政务、交通管理、国土资源、监察、环保等)，应用占比达到 49%，互联网与金融行业紧随其后，占比分别为 18% 和 12%。[25]

自 2016 年以来，随着 5G、人工智能、大数据等技术深度融入智慧城市建设，智慧交通、智慧安防、智慧医疗、智慧园区等概念应运而生并迅速落地，人工智能与城市治理的产物——“城市大脑”正日益成为智慧城市建设的重中之重。然而，人工智能在充分助力城市建设和治理过程中，也不可避免地带来诸多挑战。

首先，人工智能可能导致过度监控。随着传感器、视觉监控等技术的广泛应用，对公共空间的监控可能遍及城市的各个角落。根据 IDC 预测，2020 年全球视频监控产生的数据约 18.1 PB，占同期物联网数据量的 83.1%，构成物联网数据的主体。[26] 人工智能在为城市管理者提供“上帝视

[24] IDC: 《中国智慧城市建设在疫情稳控形势下继续保持高质量发展》， 2020 年 12 月，. https://www.idc.com/getdoc.jsp?containerId=prCHC47212520.

[25] 艾瑞咨询: 《2020 年中国人工智能产业研究报告》, 2020 年, https://www.iresearch.com.cn/Detail/report?id=3707&isfree=0

[26] IDC: The premier global market intelligence company, “IDC: 安全新视界——打通视频监控系统端到端安全,” 2021, https://www.idc.com/getdoc.jsp?containerId=prCHC47327821.

角”便利的同时，也可能造成“全景监狱”的监管困境，让居民在获得前所未有的安全环境的同时，也产生“随时被监控”的不安全感，进而影响幸福指数和城市活力。

其次，人工会智能可能侵犯个人隐私。以深度学习为代表的人工智能广泛融入智慧城市和城市大脑建设，必然导致对大量数据尤其是个人数据的无止境需求。从个人身份信息到虚拟身份信息，从生活轨迹到消费记录，从生活习惯到生物识别，生活在智慧城市中的人，必将成为人工智能落地服务的目标指向，而人工智能技术在最大程度上满足个性化需求的同时，必将收集尽可能多的个人信息。在海量个人信息被收集、分析、再更新、再收集过程中，难免造成数据滥用、数据泄露、数据黑箱等侵犯个人隐私的后果。

第三，智慧城市发展的安全困境。随着深度感知技术的广泛应用，人工智能构成了智慧城市的神经系统，在深度学习和大数据分析技术支持下，进一步形成智慧城市的中枢处理系统，即“城市大脑”。但任何系统都有漏洞，越先进的系统漏洞可能会越多。未来，攻陷一座城市可能不再需要旷日持久的巷战，只需要占领并控制该城市的人工智能系统，便可以决定每一扇门的开关、每一条消息的发布，甚至是每个人的行动轨迹。

三、人工智能的风险治理

作为颠覆性的通用技术，人工智能的价值体现在政治、经济、社会、文化等各个方面，因而也蕴含着各个方面的风险。由此，对于安全风险的治理成为决定人工智能未来发展的关键领域。归纳起来，以下十个领域是全球范围内普遍关注的人工智能风险：网络安全问题、企业合规问题、可解释性问题、隐私安全问题、声誉和伦理问题、未来岗位问题、公平性问题、人身安全问题、国家安全问题、政治稳定问题等[27]。在当下的人工智能技术发展阶段中所出现的争议基本包含在以上十个方面的问题之内。

（一）人工智能风险治理思路

在新兴技术领域的诸多风险问题中，隐私安全、人身安全和企业合规等方面的问题可以经由专业的法律加以规范，但问题在于在短期内针对人工智能的专业化立法较为艰难。如图所示，从 2016 年至今，各国越来越多地在法律方面提到人工智能（图 3-1），但是专业化的立法进程依然推进缓慢（图 3-2）。

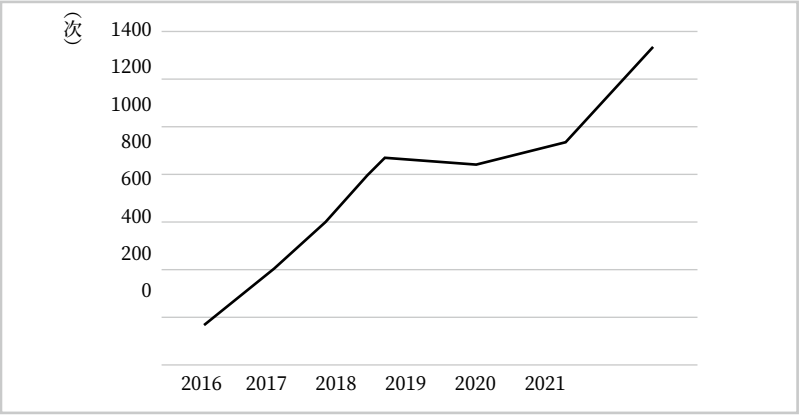


图 3-1 全球立法中提及“人工智能”的次数

数据来源：Stanford University “2022 AI INDEX REPORT”https://aiindex.stanford.edu/report/

[27] “Global Survey: The State of AI in 2020,” McKinsey, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020.

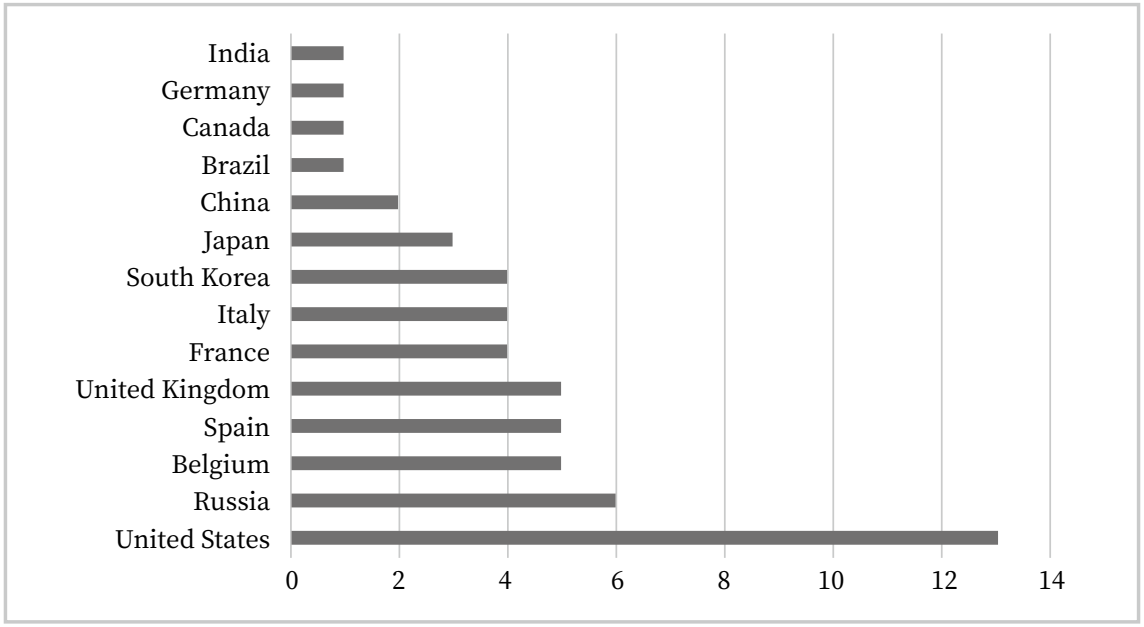


图 3-2 各国人工智能专业法案的数量

数据来源：Stanford University “2022 AI INDEX REPORT”https://aiindex.stanford.edu/report/

立法层面的缓慢进程说明当前的人工智能风险治理应该以实施监管和设置可操作性原则为主导，通过治理实践凝聚共识，并进一步助推人工智能的立法进程。由此衍生出以下风险治理思路：

1、基于未来风险预防的影响评估模式

影响评估（impact assessment）指通过多利益攸关方的咨询、参与和审议，针对可能具有重大安全风险的技术内容，在付诸实践前先进行影响评估，对其可能产生的危害进行预防性监管。影响评估主要包括隐私影响评估 (PIA)，数据保护影响评估 (DPIA)，社会影响评估（SIA）和伦理影响评估（EtIA）等。其中数据保护影响评估则被 GDPR 第 35 条引入，成为数据控制者的合规约束，也是 GDPR 问责机制中所关注的一个新环节。[28] 这种影响评估模式兼顾到了基础的数据层面以及高层次的

[28] GDPR.eu. “Data Protection Impact Assessment (DPIA),” August 9, 2018, https://gdpr.eu/data-protection-impact-assessment-template/.

社会伦理层面，希望通过实施影响评估在人工智能技术前沿使用者身上设置“责任感制动”，是一种风险的预防措施，但一定程度上依赖于企业等主体的自我意识。

2、基于自主性原则的元监管模式

元监管（meta-regulation）指政府等监管机构并不为企业设置严格的合规框架，而是由企业自行论证开发活动的合规性，并由政府机构进行调查和处罚。[29] 这种模式基本与人工智能技术活动同步开展，但政府角色退居幕后，最大程度地保证企业的自主性。但缺点也在于居于二线的政府的监管效率也会受到制约。

3、基于透明追踪的 AI 系统警戒模式

AI 系统警戒（AI system vigilance）指通过系统性的警戒以及对不良事件的透明跟踪，尽早发现问题和故障并及时进行纠错工作。[30] 这种警戒行为实际上发生在技术活动实施完毕之后，是从长期着眼的风险治理思路，有利于技术产品的长期稳定健康运行，但是监管成本也相应提高。

总体上，以上三种治理思路各有其独到的优势，也有其难以避免的缺陷。因为这些治理方式都仅能代表技术活动前、中、后阶段的特定风险管控，并没有贯通这一流程，形成成熟的治理模式。

（二）人工智能风险治理模式

近年来，各国开始探索综合性的人工智能治理模式。目前，以学界和产业界探索生成的“参与性治理模式”和 2018 年“世界经济论坛”提出的“敏捷治理模式”，逐

[29] 考 Simon, F. C., Meta-Regulation in Practice: Beyond Normative Views of Morality and Rationality, London: Routledge, 2017, p.2-17.

[30] 参考 Dubber, Markus Dirk, Frank Pasquale, and Sunit Das, The Oxford Handbook of Ethics of AI, Oxford University Press, 2020, p.88.

渐在各国的人工智能治理实践中得到有效运用。这些综合性的治理模式将有利于控制人工智能风险、协调各级主体之间的关系、为全球技术治理提供范例和思路。

1、参与性设计模式汇聚不同利益攸关方的贡献

参与性设计模式，或称参与性治理，其核心是将利益相关者（如预期的最终用户）纳入设计过程，与专业设计师和研究人员一起工作，并参与决策。而这需要先让用户充分了解人工智能技术的可能性和局限性，以便提出设计建议并做出设计决策，参与风险治理的整个过程。吸纳用户参与的治理模式直接针对了人工智能如算法偏差、决策黑箱等风险问题。在整套参与性治理的模式下，大致分为以下四个环节[31]：

一是明确参与设计的具体环节。在人工智能技术活动中，既需要政府来设置目标和底线，也需要不同的利益相关方来参与。明确具体的分工，有助于更好的发挥政府和非政府行为体的作用，让机制更加合理；

二是激发设计理念。由于参与性设计的主体多元化，许多用户并不具备专业知识，因此这一阶段通常通过一些参观或者研讨会向各设计方解释人工智能技术逻辑，而设计方则采用类似于前文提到的影响评估思路对该人工智能技术进行概念和功能上的设计；

三是选择和落实设计思路。参与性设计模式实际上为人工智能研发和运用提供了一个类似于“实地考察”的环节，但考察之后还需要专业人员进行方案选择和行动落实；

四是结果评估与审查。在落实之后，该人工智能技术活动则会进入到最后环节。一般来说，评估工作也遵从参与性设计的思路，由用户与专业人员共同评估该技术产品是否满足设计思路以及使用需求，是否蕴含着潜在风险。而审查工作则主要由政府

[31] 环节和特征部分参考 Zytko, Douglas, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. “Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains.” CHI Conference on Human Factors in Computing Systems Extended Abstracts, 1–4. New Orleans LA USA: ACM, 2022. https://doi.org/10.1145/3491101.3516506.

等审查机构进行，从另一个角度审查技术活动是否符合国家人工智能的发展方向和安全标准。

归纳一下，参与性设计模式贯穿了技术活动的整个流程，这种以用户为导向的治理思路也在一定程度上合理规避了前文提到的可解释性问题、隐私安全问题、声誉和伦理问题及公平性问题的风险，不失为未来人工智能风险治理的一种科学模式。

2、敏捷治理模式贡献应对技术不确定性的治理思路

相比于参与性治理的以用户为倚重的风险治理模式，敏捷治理则以政府为中心整合整个人工智能风险治理流程。[32]

敏捷治理模式的重要特点就是通过政府的系统性整合，与各利益攸关方共同组成一体化的人工智能治理生态，并通过灵敏、及时、持续的“咨询 - 反馈”机制，促进治理政策的迭代升级，以弥补政府治理中的信息的滞后性，形成对人工智能风险的前瞻性评估与治理。在整个敏捷治理的思路中，包括三个要素——两条“咨询 - 反馈”路径和一个“动态评估”机制，一起致力于治理政策更新[33]。

一是“咨询 – 反馈”路径。建立在政府所规划的人工智能“治理原则”与各利益攸关方掌握的人工智能技术“事实特征”之间。在敏捷治理模式下，政府并不需要制定明确严格的人工智能治理规范，而是需要设定一种原则框架，并通过这条“咨询 - 反馈”路径与掌握市场和技术信息的多利益攸关方进行配合，保证治理原则设置的科学性；

二是“咨询 – 反馈”路径。建立在政府所规划的人工智能“治理目标”与各利益攸关方掌握的“市场需求信息、价值观与公正性”之间。与治理原则类似，政府在

[32] 薛澜，赵静：《走向敏捷治理 : 新兴产业发展与监管模式探究》，《中国行政管理》2019 年第 8 期，第 28–34 页。

[33] 参考“Agile Governance Reimagining Policy-Making in the Fourth Industrial Revolution,” World Economic Forum, January 2018. https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf.

敏捷治理下所需要设计的第二个内容就是人工智能的治理目标，以规划整个治理活动的道路和方向。但是这条道路欲行稳致远，就需要满足各利益攸关方切身的市场需求和精神价值需求。

三是一套动态的评估机制。在这两条路径之外，政府还需要设置一套动态的评估机制，用于评估人工智能技术活动前是否符合原则、技术活动后是否满足目标并联通整个过程。此外，敏捷治理模式还通过企业设计试点技术与政府审查机构的动态评估相结合，塑造出一种政策的迭代机制，从而在兼顾战略目标导向的同时回应技术与环境的变化。这对于人工智能技术本身的“不确定性”风险应对提供了有力指引。

（三）人工智能域外治理经验

1、美国的人工智能风险管理框架

2022 年 3 月，美国国家标准与技术研究院 (The National Institute of Standards and Technology,NIST) 发布人工智能风险管理框架（AI RMF）草案[34]，该草案涉及 AI 系统的设计、开发、使用和评估方面的风险，从而促进值得信赖和负责任的人工智能（AI）技术和系统的发展和使用。

NIST 通过与私营和公共部门的合作，通过共识驱动、开放、透明和协作的开发过程，提出了人工智能风险管理框架。[35] 该框架有两个目标，一是促进开发创新，提高 AI 的可信度，包括准确性、可解释性、可靠性、隐私性、稳健性、安全性、弹性，以及减少意外和 / 或有害的偏见以及有害的用途等特征。二是在人工智能技术和系统的预设计、设计和开发、部署、使用以及测试和评估过程中考虑透明度、问责制和公

[34] “AI Risk Management Framework Concept Paper.” The National Institute of Standards and Technology (NIST), December 2021, https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf. .

[35] Federal Register. “Artificial Intelligence Risk Management Framework,” July 2021, https://www.federalregister.gov/documents/2021/07/29/2021-16176/artificial-intelligence-risk-management-framework.

平性等原则，从而提高人工智能技术和系统、产品和服务的可信度。

该框架采用三级特征分类法，在识别和管理人工智能系统相关风险的综合方法中应考虑三个要素——技术特征、社会技术特征和指导原则。其中，技术特征是指人工智能系统设计者和开发者直接控制的因素，可以使用评估标准来衡量，如准确性、可靠性和弹性。社会技术特征指人工智能系统在个人、群体和社会环境中的使用和感知方式，如“可解释性”、隐私、安全和管理歧视。在 AI RMF 分类法中，指导原则指的是更广泛的社会规范和价值观，表明了公平、问责和透明度等社会优先事项。

2、欧盟的人工智能监管框架

《人工智能法案》遵循了一种基于风险的方法，根据具体的风险水平将其划分为了不可接受的风险、高风险、有限风险和低风险四个层级：

一、不可接受的风险：这一风险层级要求明确禁止对人类安全、生计和权利构成明显威胁的有害人工智能做法，例如利用特定弱势群体、进行社会评分、以执法为由进行实时生物识别等。欧盟禁止在其所管辖的市场内投放或使用该类人工智能应用。

二、高风险：这一风险层级不仅取决于人工智能系统的功能，还取决于使用该系统的特定目的和方式。欧盟确定了两类主要的高风险人工智能系统：已经由第三方机构事先评估合格后作用于安全领域的人工智能系统；在生物识别、基础设施管理运作、教育和职业培训、就业与员工管理、获取公共服务与重要私人服务的权利、执法、移民与边境管制、司法行政与民主流程方面有所参与的人工智能应用。这些高风险的人工智能系统在符合某些强制性要求和事前合格评估的前提下，获准进入欧洲市场。

三、有限风险：这一层级包含在人工智能系统涉及与人类进行互动、用于检测情感或基于生物特征的分析、通过技术生成或操纵相关内容时，欧盟要求该类人工智能系统有义务告知人们存在此类情况。而当人工智能系统被用于生成与真实内容相似

的音视频内容时，也有义务披露该内容是在合法目的下自动生成的。

四、低风险或最小风险：这一层级包含其他只有低等级风险的人工智能系统，可以不受欧盟附加法规的限制。但是欧盟也在建立行为准则，以鼓励这类人工智能系统自愿遵从人工智能法案的高风险强制性要求，或是创建自己的强制性行为准则，由此保护环境、保证多样性、给予无障碍环境等。

欧盟《人工智能法案》使得人工智能在数据保护和利用上达成了有效平衡，与欧洲《通用数据保护条例》（GDPR）实现了有效衔接。人工智能的快速发展离不开数据的有效利用。《人工智能法案》的出台意味着人工智能可以在收集使用个人数据方面发挥积极作用，由此能够在更合理利用数据的同时更全面地保护数据。一方面，人工智能的使用将有助于促进个人数据的收集与使用更好地符合 GDPR 的要求，促进个人数据的保护。另一方面，在 GDPR 框架下规范使用个人数据也有利于人工智能技术发挥最大效用。两者相互扶持，共同推进人工智能向好向善发展。

3、联合国教科文组织的首份人工智能伦理全球协议

联合国教科文组织于 2021 年 11 月 25 日通过《人工智能伦理问题建议书》，“认为关于人工智能技术及其社会影响的规范框架应建立在共识和共同目标的基础上，以国际和国家法律框架、人权和基本自由、伦理、获取数据、信息和知识的需求、研究和创新自由、人类福祉、环境和生态系统福祉为依据，建议会员国动员包括工商企业在内的所有利益攸关方，使人工智能技术的开发和应用做到以健全的科学研究以及伦理分析和评估作为指导。”[36] 联合国教科文组织希望通过该建议书提供一个在价值观、原则和行动层面可以被各国普遍执行的研究框架，在私营部门与公民社会层面体现出指导意义，使人工智能全生命周期符合伦理规范，并以此建议书推动全球各利益攸关

[36] 《人工智能伦理问题建议书》，联合国教科文组织，2021 年 11 月 25 日，https://unesdoc.unesco.org/ark:/48223/pf0000380455_chi.

方的多元对话，由此达成惠及世界各国与社会各阶层的人工智能进步。该建议书提出了如下的价值观：

① 尊重、保护和促进人权和基本自由以及人的尊严

建议书中认为在人工智能全生命周期中，“任何人或人类社群在身体、经济、社会、政治、文化或精神等任何方面，都不应受到损害或被迫居于从属地位”[37]，人权与基本自由需要得到保证，且人工智能应当可以促进该类权力得到更完善的保护。

② 环境和生态系统蓬勃发展

建议书要求在保护环境、促进可持续发展的前提下开展人工智能活动，各国政府需要评估人工智能生命周期中对环境的影响，将其对气候变化和环境风险的影响因素降到最低，以此防止环境恶化与生态系统退化。而人工智能也应当有利于环境，促进环境保护运动的发展。

③ 确保多样性和包容性

建议书要求各国政府确保人工智能遵守国际法，尊重、保护和促进多样性和包容性。同时各国应积极开展国际合作，由此弥补人工智能技术鸿沟所带来的基础设施、教育、技能、法律的缺失，不应利用他国的技术不发达现状实施有损他国利益的行为。

④ 生活在和平、公正与互联的社会中

建议书认为人工智能行为者要使人工智能在和平与公正的社会中发挥参与和促进作用，由此惠及全民，并带来自由互联的未来。人工智能应当促进公平、包容的互联环境，使得社会形成有机、直接、出自本能的团结纽带。

建议书还提出了包括伦理影响评估、伦理治理和管理、数据政策、发展与国际合作、环境和生态系统、性别、文化、教育和研究、传播和信息、经济和劳动、健康和社会福祉在内的各政策领域要求，全方位规制人工智能的伦理原则。各会员国也被

[37] 《人工智能伦理问题建议书》，联合国教科文组织，2021 年 11 月 25 日，https://unesdoc.unesco.org/ark:/48223/pf0000380455_chi.

要求制定科学有效的方法，在可信透明的大框架下监测和评估与人工智能有关的政策、计划和机制。

四、人工智能安全治理的行业实践

政策层面，各方将人工智能产业列为战略性新兴产业，加大支持力度。企业既是人工智能技术的开发者，也是人工智能安全治理的重要参与者。它们积极参与标准制定，推出安全治理产品服务，将人工智能安全治理作为产业发展机遇。

（一）技术解决方案

从行业角度来说，人工智能应用领域涉及医疗、金融、零售、政府治理等，技术渗透度不断攀升。2018 年以来，谷歌、微软、IBM、腾讯、阿里、百度等国内外企业纷纷推出各自人工智能治理方案。

IBM 研发出一系列可信人工智能关键技术，例如，人工智能公平 360 工具箱（AIF360），可用于检测和缓解机器学习模型中的偏见；对抗性鲁棒性 360 工具箱（ART），可用于快速制作和分析机器学习模型的攻击和防御方法；人工智能可解释 360（AIX360），可用于支持机器学习模型和算法的可解释性。[38]

微软联合 MITRE、Bosch、IBM 等多家公司推出对抗机器学习的威胁矩阵。微软研制的 Counterfit 算法安全攻防工具，可以大规模地攻击多个 AI 模型，目前已成为微软人工智能业务安全测试的主要工具。[39]

[38] IBM Research Teams, “Trusted AI,” February 2021, https://research.ibm.com/teams/trusted-ai?_ga=2.264991951.1171655595.1652318757-354424505.1652318757.

[39] Microsoft Security Blog, “AI Security Risk Assessment Using Counterfit,” May 2021, https://www.microsoft.com/security/blog/2021/05/03/ai-security-risk-assessment-using-counterfit/.

腾讯于 2019 年基于黑盒测试和优化算法，在车道线系统攻击（即在路面部署干扰信息，导致车辆经过时对车道线做出错误判断，致使车辆驶入反向车道）方法基础上，设计了一套针对车道线系统的自动化攻击方法。[40]2020 年，腾讯发布业内首个 AI 安全攻击矩阵，并展示 AI 模型后门攻击研究成果。2021 年，腾讯推出声音防克隆研究成果，以及对抗 Deepfake 的新思路，即 MagDR。[41]

阿里于 2021 年发布自动化 AI 对抗攻击平台 CAA，联合清华、UIUC 举办 CVP2021 挑战者计划第六期；联合清华、RealAI 发布 AI 安全评估基准平台。

百度于 2018 年发布对抗攻击开源工具箱 AdvBox，2021 年提交 86 个 TensorFlow 漏洞，并举办首届自动驾驶 CTF。

华为于 2020 年发布 MindArmour，为 MindSpore 框架提供安全和隐私保护能力。

奇虎 360 于 2019 年承建“安全大脑国家新一代人工智能开放创新平台”，该平台于 2020 年 6 月获得科技部“科技创新 2030——新一代人工智能重大项目”资助，并已经与清华大学、中科院自动化所、中科院信工所、北京瑞莱智慧科技有限公司等开展共建合作。360 安全大脑是基于 360 十六年攻防实战经验积累，凭借全网安全大数据、人工智能技术、一线对抗和 APT 狩猎形成的安全知识及安全技术、安全专家等核心优势能力打造的网络空间“预警机和反导系统”。以该平台 360 安全大脑为中枢，构建了新一代网络安全能力体系，致力于帮助国家、政府、城市、行业和企业，整体提升应对高级威胁攻击的安全能力。[42]

---

[40] Jing, Pengfei, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu, “Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations,” 3237–54, 2021. https://www.usenix.org/conference/usenixsecurity21/presentation/jing.

[41] Chen, Zhikai, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. “MagDR: Mask-Guided Detection and Reconstruction for Defending Deepfakes.” ArXiv:2103.14211 [Cs], March 25, 2021. http://arxiv.org/abs/2103.14211.

[42] 《国家级 AI 创新力量集中亮相！ AIExpo2020‘新一代人工智能开放创新平台’高峰论坛成功举办》， 网易，2020 年 8 月 16 日， https://www.163.com/dy/article/FK5OQLU705118HA4.html.

其中，360 人工智能框架漏洞威胁感知系统能够从算法实现、漏洞类型、编译优化等多角度持续开展对机器学习框架的安全风险研究，并采用动、静结合的分析方法对 Python、C++、Go 等不同语言中的不同漏洞类型进行全面系统的检测，检测过程着重关注从训练到推理、从数据到模型、从云端到终端过程中存在的安全风险。[43] 该系统已经累计发现并帮助国内外 7 个厂商修复 7 款 AI 框架漏洞 180 余个。其中发现谷歌 Tensorflow 框架漏洞 135 个（CVE），包括严重漏洞 6 个。漏洞数在全球各大厂商排名第一，并首次发现危险等级为严重的漏洞。[44]

大数据协同安全技术国家研究中心开发的基于联邦学习的 APT 智能检测模型训练系统。该系统以 360 安全大数据为核心数据集，根据已披露的 APT 组织的攻击行为，结合 360ATT&CK 攻防知识图谱，对 APT 攻击进行建模。并且，采用纵向联邦和横向联邦相结合的方法，全面覆盖 APT 攻击模型特征，充分利用不同类型的样本库，在保证数据不出库和安全的条件下进行模型训练，提升 APT 攻击行为监测识别和挖掘分析能力，以形成对网络安全威胁特别是 APT 的全景视图。

（二）行业标准

在行业治理方面，国内外行业组织纷纷制定行业标准，并将其作为行业自律的重要组成部分。

表 4–1 人工智能行业标准范例

行业组织名称	标准内容与相关事件
国际标准化组织（ISO）	于 2017 年 10 月批准成立了 JTC 1/SC 42 人工智能分技术委员会，重点在术语、参考框架、算法模型和计算方法、安全及可信等方面开展标准化研究。

[43] 《360 两项成果入选“人工智能安全典型实践案例”》，360 政企安全， 2021 年 10 月 13 日， https://www.360.net/about/news/article61679316eec939004a2dfae0.

[44] TensorFlow Security Advisories, https://github.com/tensorflow/tensorflow/blob/master/tensorflow/security/README.md

国际电信联盟（ITU）	主要致力于解决智慧医疗、智能汽车、垃圾内容治理、生物特征识别等人工智能应用中的安全问题。
电气和电子工程师协会（IEEE）	开展了多项人工智能伦理道德研究，发布了多项人工智能伦理标准和研究报告，研制出 IEEE7000 系列标准，用于规范人工智能系统道德规范问题。
美国国家标准与技术研究院（NIST）	于 2019 年发布了关于政府如何制定人工智能技术标准和相关工具的指导意见。
全国信息技术标准化技术委员会（SAC/TC 28）	主要在人工智能术语词汇、人机交互、生物特征识别、大数据、云计算等领域开展了标准化工作。
中国信息通信研究院（CAICT）	在 ITU-T SG16 全会上，牵头创立了人工智能多媒体新课题，布局人工智能与多媒体融合研究与标准输出。
中国通信标准化协会（CCSA）	发布了《智能家居终端设备安全能力技术要求》《智能家居网络安全系统安全技术要求》等标准。
中国人工智能产业发展联盟（AIIA）	于 2019 年发布了《可信 AI 操作指引》，并公布了首批商用人工智能系统可信评估结果，涉及 11 家企业的 16 个人工智能系统，为用户选型提供参考。
中国人工智能开源软件发展联盟（AIOSS）	推出了有关机器翻译、智能助理等产品或服务评估标准，以及深度学习算法的可靠性评估标准。

资料来源：根据各行业组织官方文件整理得出。检索日期：2022 年 5 月 11 日。

（三）应用案例

1、Darktrace 人工智能网络安全 AI 公司与微软达成合作

2021 年 5 月，微软与 Darktrace 达成合作，Darktrace 的自学习 AI 帮助微软邮件、Microsoft 365、Azure 等服务用户应对网络威胁，并与 Microsoft Sentinel 实现整合，帮助企业在多云和多平台环境中构建安全能力。

Darktrace 是一家领先的人工智能网络安全 AI 公司，也是自动化相关技术的创始者。微软和 Darktrace 的技术集成将帮助微软通过分析恶意 IP 地址、域和 URL 来发现端点漏洞并提高 Web 保护能力。具体而言，采用自学习技术赋能邮件安全、Microsoft 365 和 Azure 组件，并与 SIEM 平台 Sentinel 高度集成。[45]

① AI 邮件安全。通过学习 Microsoft 365 中每个用户的正常“行为模式”，以识别组织中存在的异常行为。Darktrace 能够识别出新的电子邮件威胁，包括复杂的网络钓鱼、商业电子邮件泄露 (BEC) 和供应链攻击（或供应商电子邮件泄露）。

② Microsoft 365。自学习技术识别 Microsoft 365 产品组件中的网络安全威胁，包括凭据泄漏、管理员滥用、恶意内部人员和远程工作风险。当 Darktrace 在 Microsoft 365 中检测到网络事件时，对事件进行分类、解释和报告，以帮助组织对威胁快速响应。

③ Microsoft Azure 云安全。通过自学习技术，对 Azure 云环境中正常行为进行深入了解，将行为置于上下文中，识别出与正常“行为模式”的偏差，以检测威胁。自学习 AI 可以自动连接不同基础设施区域中异常行为之间的点，确保云安全不会与组织其他部门的监控相隔离。

④与事件管理与响应 (SIEM) 的整合。Darktrace 工作日志允许安全团队在 SIEM 中发送警报和网络事件。并可以按活动对这些进行分组，用户只需单击即可返回 Darktrace 威胁可视化工具，以便进行进一步调查。

2、瑞莱智慧 DeepReal 深度伪造内容检测平台

瑞莱智慧成立于 2018 年 7 月，孵化自清华大学人工智能研究院。其深度伪造内容检测平台 DeepReal, 依托第三代人工智能技术，通过辨识伪造内容和真实内容的

[45] 参考 Darktrace,“World–Leading AI for Cyber Security,” https://www.darktrace.com/en/.

表征差异性、挖掘不同生成途径的深度伪造内容一致性特征，能够快速、精准地对图像、视频、音频内容进行真伪鉴别，有效打击财产诈骗、色情黑产、虚假宣传、证据造假等违法行为。[46]

该平台以“实网高性能高精度检测能力”与“深度伪造治理生态共建”为特点，应用于网络内容合规检测、人脸验证安全、图像物证真实性检测、反侵权诈骗等场景。

该平台通过深度伪造检测算法对图像、视频和音频等内容进行检测，对内容进行监测鉴别和可解释性说明，并提供多维度检测报告。在技术上基于贝叶斯深度学习、多特征融合和多任务学习等方法进行设计研制；基于千万级数据集进行训练，具备领先的准确率和良好的鲁棒性；检测效率达 30 毫秒每帧。

3、360 集团的基于数字孪生的物联网安全攻防平台

2022 年 4 月，在工信部公示 2021 年物联网示范项目中，360“基于数字孪生的物联网安全攻防平台”入选关键技术攻关类示范项目。[47] 该平台建设旨在有效解决智慧城市物联网未来可能遇到的网络空间攻击威胁，提高城市整体防御水平和应急响应能力，护航智慧城市数字安全。

该平台有两大创新点，一是通过数字孪生技术与 AI 技术结合，建立智慧城市物联网的数字化孪生网络，从而支持更加灵活、更加便捷、更大规模的智慧城市物联网网络仿真，实现对区域智慧城市物联网设备群的仿真攻防能力智能验证。[48]

二是解决物联网设备自动化漏洞挖掘技术突破问题。该平台通过关键漏洞挖掘技术攻关，实现智慧城市场景下的相关设备固件静态扫描及动态模拟和智能化漏洞挖

[46] 《深度伪造内容检测平台 DeepReal》，瑞莱智慧 RealAI，https://www.real-ai.cn/products/9.html. Accessed May 12, 2022. https://www.real-ai.cn/products/9.html.

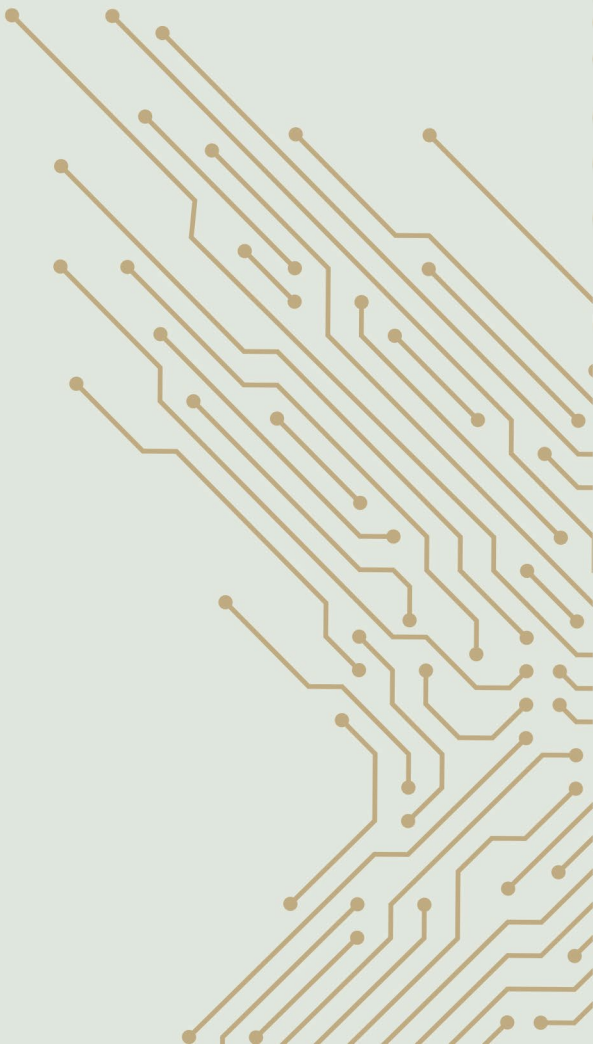
[47] 中华人民共和国工业和信息化部：《2021 年物联网示范项目公示》，2022 年 4 月 13 日， https://www.miit.gov.cn/zwgk/wjgs/art/2022/art_81b88cf50fd144f19267faeca3a35c2c.html.

[48] 《360 攻防平台入选工信部 2021 年物联网示范项目，护航智慧城市数字安全》，通信世界网，2022 年 4 月，http://www.cww.net.cn/article?id=561273.

掘能力，支持 ARM,MIPS,X86 等主流处理器架构，覆盖主流网络协议，以及在智能自动化漏洞挖掘方向进行持续攻关，模糊测试综合代码覆盖率在业界处于领先地位。

五、总结

人工智能作为中、美、欧等国家和地区都在积极发展的关键新兴技术，其在发展过程中所产生的安全挑战也更为复杂多元。为解决人工智能在自身安全、赋能安全、衍生安全、数字城市建设安全方面的挑战，人工智能风险治理模式应运而生。它以可操作性为主要指导原则，囊括影响评估、元监管与 AI 系统警戒三大治理思路，细分出以用户为考虑重点的参与性设计和以政府为主导力量的敏捷治理两条路径。在各国及各组织的治理实践之下，上述模式又形成了各自的人工智能官方监管框架。在此之下，包含 360 在内的多家人工智能龙头企业也以自身实践构建行业安全案例，走出了技术赋能、行业规制、平台监测的多种道路。相信各国都会加紧推进人工智能安全产业发展，继续助力人工智能成为集技术革新与安全可控于一身的战略性技术。



北京奇虎科技有限公司

联系方式：dipperresearch@360.cn
联系方式：北京市朝阳区酒仙桥路6号院2号楼
联系电话：010-58781000
传　　真：+86-10-56822000

上海国际问题研究院网络空间国际治理研究中心

联系方式：上海市徐汇区田林路195 弄15 号
联系电话：+86-21-54614900
传　　真：+86-21-64850100