人工智能
安全治理框架

全国网络安全标准化技术委员会
2024年9月

目 录

人工智能安全治理框架
（V1.0）

人工智能是人类发展新领域，给世界带来巨大机遇，也带来各类风险挑战。落实《全球人工智能治理倡议》，遵循“以人为本、智能向善”的发展方向，为推动政府、国际组织、企业、科研院所、民间机构和社会公众等各方，就人工智能安全治理达成共识、协调一致，有效防范化解人工智能安全风险，制定本框架。

1. 人工智能安全治理原则

秉持共同、综合、合作、可持续的安全观，坚持发展和安全并重，以促进人工智能创新发展为第一要务，以有效防范化解人工智能安全风险为出发点和落脚点，构建各方共同参与、技管结合、分工协作的治理机制，压实相关主体安全责任，打造全过程全要素治理链条，培育安全、可靠、公平、透明的人工智能技术研发和应用生态，推动人工智能健康发展和规范应用，切实维护国家主权、安全和发展利益，保障公民、法人和其他组织的合法权益，确保人工智能技术造福于人类。

1.1 包容审慎、确保安全。鼓励发展创新，对人工智能研发及应用采取包容态度。严守安全底线，对危害国家安全、社会公共利益、公众合法权益的风险及时采取措施。

1.2 风险导向、敏捷治理。密切跟踪人工智能研发及应用趋势，从人工智能技术自身、人工智能应用两方面分析梳理安全风险，提出针对性防范应对措施。关注安全风险发展变化，快速动态精准调整治理措施，持续优化治理机制和方式，对确需政府监管事项及时予以响应。

1.3 技管结合、协同应对。面向人工智能研发应用全过程，综合运用技术、管理相结合的安全治理措施，防范应对不同类型安全风险。围绕人工智能研发应用生态链，明确模型算法研发者、服务提供者、使用者等相关主体的安全责任，有机发挥政府监管、行业自律、社会监督等治理机制作用。

1.4 开放合作、共治共享。在全球范围推动人工智能安全治理国际合作，共享最佳实践，提倡建立开放性平台，通过跨学科、跨领域、跨地区、跨国界的对话和合作，推动形成具有广泛共识的全球人工智能治理体系。

2. 人工智能安全治理框架构成

基于风险管理理念，本框架针对不同类型的人工智能安全风险，从技术、管理两方面提出防范应对措施。同时，目前人工智能研发应用仍在快速发展，安全风险的表现形式、影响程度、认识感知亦随之变化，防范应对措施也将相应动态调整更新，需要各方共同对治理框架持续优化完善。

2.1 安全风险方面。通过分析人工智能技术特性，以及在不同行业领域应用场景，梳理人工智能技术本身，及其在应用过程中面临的各种安全风险隐患。

2.2 技术应对措施方面。针对模型算法、训练数据、算力设施、产品服务、应用场景，提出通过安全软件开发、数据质量提升、安全建设运维、测评监测加固等技术手段提升人工智能产品及应用的安全性、公平性、可靠性、鲁棒性

的措施。

2.3 综合治理措施方面。明确技术研发机构、服务提供者、用户、政府部门、行业协会、社会组织等各方发现、防范、应对人工智能安全风险的措施手段，推动各方协同共治。

2.4 安全开发应用指引方面。明确模型算法研发者、服务提供者、重点领域用户和社会公众用户，开发应用人工智能技术的若干安全指导规范。

3. 人工智能安全风险分类

人工智能系统设计、研发、训练、测试、部署、使用、维护等生命周期各环节都面临安全风险，既面临自身技术缺陷、不足带来的风险，也面临不当使用、滥用甚至恶意利用带来的安全风险。

3.1 人工智能内生安全风险

3.1.1 模型算法安全风险

（a）可解释性差的风险。以深度学习为代表的人工智能算法内部运行逻辑复杂，推理过程属黑灰盒模式，可能导致输出结果难以预测和确切归因，如有异常难以快速修正和溯源追责。

（b）偏见、歧视风险。算法设计及训练过程中，个人偏见被有意、无意引入，或者因训练数据集质量问题，导致算法设计目的、输出结果存在偏见或歧视，甚至输出存在民族、宗教、国别、地域等歧视性内容。

（c）鲁棒性弱风险。由于深度神经网络存在非线性、大规模等特点，人工智能易受复杂多变运行环境或恶意干扰、诱导的影响，可能带来性能下降、决策错误等诸多问题。

（d）被窃取、篡改的风险。参数、结构、功能等算法核心信息，面临被逆向攻击窃取、修改，甚至嵌入后门的风险，可导致知识产权被侵犯、商业机密泄露，推理过程不可信、决策输出错误，甚至运行故障。

（e）输出不可靠风险。生成式人工智能可能产生“幻觉”，即生成看似合理，实则不符常理的内容，造成知识偏见与误导。

（f）对抗攻击风险。攻击者通过创建精心设计的对抗样本数据，隐蔽地误导、影响，以至操纵人工智能模型，使其产生错误的输出，甚至造成运行瘫痪。

3.1.2 数据安全风险

（a）违规收集使用数据风险。人工智能训练数据的获取，以及提供服务与用户交互过程中，存在未经同意收集、不当使用数据和个人信息的安全风险。

（b）训练数据含不当内容、被“投毒”风险。训练数据中含有虚假、偏见、侵犯知识产权等违法有害信息，或者来源缺乏多样性，导致输出违法的、不良的、偏激的等有害信息内容。训练数据还面临攻击者篡改、注入错误、误导数据的“投毒”风险，“污染”模型的概率分布，进而造成准确性、可信度下降。

（c）训练数据标注不规范风险。训练数据标注过程中，存在因标注规则不完备、标注人员能力不够、标注错误等问题，不仅会影响模型算法准确度、可靠性、有效性，还可能导致训练偏差、偏见歧视放大、泛化能力不足或输出错误。

（d）数据泄露风险。人工智能研发应用过程中，因数据处理不当、非授权访问、恶意攻击、诱导交互等问题，可能导致数据和个人信息泄露。

3.1.3 系统安全风险

（a）缺陷、后门被攻击利用风险。人工智能算法模型设计、训练和验证的标准接口、特性库和工具包，以及开发界面和执行平台可能存在逻辑缺陷、

漏洞等脆弱点，还可能被恶意植入后门，存在被触发和攻击利用的风险。

（b）算力安全风险。人工智能训练运行所依赖的算力基础设施，涉及多源、泛在算力节点，不同类型计算资源，面临算力资源恶意消耗、算力层面风险跨边界传递等风险。

（c）供应链安全风险。人工智能产业链呈现高度全球化分工协作格局。但个别国家利用技术垄断和出口管制等单边强制措施制造发展壁垒，恶意阻断全球人工智能供应链，带来突出的芯片、软件、工具断供风险。

3.2 人工智能应用安全风险

3.2.1 网络域安全风险

（a）信息内容安全风险。人工智能生成或合成内容，易引发虚假信息传播、歧视偏见、隐私泄露、侵权等问题，威胁公民生命财产安全、国家安全、意识形态安全和伦理安全。如果用户输入的提示词存在不良内容，在模型安全防护机制不完善的情况下，有可能输出违法有害内容。

（b）混淆事实、误导用户、绕过鉴权的风险。人工智能系统及输出内容等未经标识，导致用户难以识别交互对象及生成内容来源是否为人工智能系统，难以鉴别生成内容的真实性，影响用户判断，导致误解。同时，人工智能生成图片、音频、视频等高仿真内容，可能绕过现有人脸识别、语音识别等身份认证机制，导致认证鉴权失效。

（c）不当使用引发信息泄露风险。政府、企业等机构工作人员在业务工作中不规范、不当使用人工智能服务，向大模型输入内部业务数据、工业信息，导致工作秘密、商业秘密、敏感业务数据泄露。

（d）滥用于网络攻击的风险。人工智能可被用于实施自动化网络攻击或

提高攻击效率，包括挖掘利用漏洞、破解密码、生成恶意代码、发送钓鱼邮件、网络扫描、社会工程学攻击等，降低网络攻击门槛，增大安全防护难度。

（e）模型复用的缺陷传导风险。依托基础模型进行二次开发或微调，是常见的人工智能应用模式，如果基础模型存在安全缺陷，将导致风险传导至下游模型。

3.2.2 现实域安全风险

（a）诱发传统经济社会安全风险。人工智能应用于金融、能源、电信、交通、民生等传统行业领域，如自动驾驶、智能诊疗等，模型算法存在的幻觉输出、错误决策，以及因不当使用、外部攻击等原因出现系统性能下降、中断、失控等问题，将对用户人身生命财产安全、经济社会安全稳定等造成安全威胁。

（b）用于违法犯罪活动的风险。人工智能可能被利用于涉恐、涉暴、涉赌、涉毒等传统违法犯罪活动，包括传授违法犯罪技巧、隐匿违法犯罪行为、制作违法犯罪工具等。

（c）两用物项和技术滥用风险。因不当使用或滥用人工智能两用物项和技术，对国家安全、经济安全、公共卫生安全等带来严重风险。包括极大降低非专家设计、合成、获取、使用核生化导武器的门槛；设计网络武器，通过自动挖掘与利用漏洞等方式，对广泛潜在目标发起网络攻击。

3.2.3 认知域安全风险

（a）加剧“信息茧房”效应风险。人工智能将广泛应用于定制化的信息服务，收集用户信息，分析用户类型、需求、意图、喜好、行为习惯，甚至特定时间段公众主流意识，进而向用户推送程式化、定制化信息及服务，“信息茧房”效应进一步加剧。

（b）用于开展认知战的风险。人工智能可被利用于制作传播虚假新闻、

图像、音频、视频等，宣扬恐怖主义、极端主义、有组织犯罪等内容，干涉他国内政、社会制度及社会秩序，危害他国主权；通过社交机器人在网络空间抢占话语权和议程设置权，左右公众价值观和思维认知。

3.2.4 伦理域安全风险

（a）加剧社会歧视偏见、扩大智能鸿沟的风险。利用人工智能收集分析人类行为、社会地位、经济状态、个体性格等，对不同人群进行标识分类、区别对待，带来系统性、结构性的社会歧视与偏见。同时，拉大不同地区人工智能鸿沟。

（b）挑战传统社会秩序的风险。人工智能发展及应用，可能带来生产工具、生产关系的大幅改变，加速重构传统行业模式，颠覆传统的就业观、生育观、教育观，对传统社会秩序的稳定运行带来挑战。

（c）未来脱离控制的风险。随着人工智能技术的快速发展，不排除人工智能自主获取外部资源、自我复制，产生自我意识，寻求外部权力，带来谋求与人类争夺控制权的风险。

4. 技术应对措施

针对上述安全风险，模型算法研发者、服务提供者、系统使用者等需从训练数据、算力设施、模型算法、产品服务、应用场景各方面采取技术措施予以防范。

4.1 针对人工智能内生安全风险

4.1.1 模型算法安全风险应对

（a）不断提高人工智能可解释性、可预测性，为人工智能系统内部构造、

推理逻辑、技术接口、输出结果提供明确说明，正确反映人工智能系统产生结果的过程。

（b）在设计、研发、部署、维护过程中建立并实施安全开发规范，尽可能消除模型算法存在的安全缺陷、歧视性倾向，提高鲁棒性。

4.1.2 数据安全风险应对

（a） 在训练数据和用户交互数据的收集、存储、使用、加工、传输、提供、公开、删除等各环节，应遵循数据收集使用、个人信息处理的安全规则，严格落实关于用户控制权、知情权、选择权等法律法规明确的合法权益。

（b） 加强知识产权保护，在训练数据选择、结果输出等环节防止侵犯知识产权。

（c） 对训练数据进行严格筛选，确保不包含核生化导武器等高危领域敏感数据。

（d）训练数据中如包含敏感个人信息和重要数据，应加强数据安全管理，符合数据安全和个人信息保护相关标准规范。

（e） 使用真实、准确、客观、多样且来源合法的训练数据，及时过滤失效、错误、偏见数据。

（f） 向境外提供人工智能服务，应符合数据跨境管理规定。向境外提供人工智能模型算法，应符合出口管制要求。

4.1.3 系统安全风险应对

（a）对人工智能技术和产品的原理、能力、适用场景、安全风险适当公开，对输出内容进行明晰标识，不断提高人工智能系统透明性。

（b）对聚合多个人工智能模型或系统的平台，应加强风险识别、检测、防护，防止因平台恶意行为或被攻击入侵影响承载的人工智能模型或系统。

（c）加强人工智能算力平台和系统服务的安全建设、管理、运维能力，确保基础设施和服务运行不中断。

（d）对于人工智能系统采用的芯片、软件、工具、算力和数据资源，应高度关注供应链安全。跟踪软硬件产品的漏洞、缺陷信息并及时采取修补加固措施，保证系统安全性。

4.2 针对人工智能应用安全风险

4.2.1 网络域风险应对

（a）建立安全防护机制，防止模型运行过程中被干扰、篡改而输出不可信结果。

（b）应建立数据护栏，确保人工智能系统输出敏感个人信息和重要数据符合相关法律法规。

4.2.2 现实域风险应对

（a）根据用户实际应用场景设置服务提供边界，裁减人工智能系统可能被滥用的功能，系统提供服务时不应超出预设应用范围。

（b）提高人工智能系统最终用途追溯能力，防止被用于核生化导等大规模杀伤性武器制造等高危场景。

4.2.3 认知域风险应对

（a）通过技术手段判别不符合预期、不真实、不准确的输出结果，并依法依规监管。

（b）对收集用户提问信息进行关联分析、汇聚挖掘，进而判断用户身份、喜好以及个人思想倾向的人工智能系统，应严格防范其滥用。

（c）加强对人工智能生成合成内容的检测技术研发，提升对认知战手段

的防范、检测、处置能力。

4.2.4 伦理域风险应对

（a）在算法设计、模型训练和优化、提供服务等过程中，应采取训练数据筛选、输出校验等方式，防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等方面歧视。

（b）应用于政府部门、关键信息基础设施以及直接影响公共安全和公民生命健康安全的领域等重点领域的人工智能系统，应具备高效精准的应急管控措施。

5. 综合治理措施

在采取技术应对措施的同时，建立完善技术研发机构、服务提供者、用户、政府部门、行业协会、社会组织等多方参与的人工智能安全风险综合治理制度规范。

5.1 实施人工智能应用分类分级管理。根据功能、性能、应用场景等，对人工智能系统分类分级，建立风险等级测试评估体系。加强人工智能最终用途管理，对特定人群及场景下使用人工智能技术提出相关要求，防止人工智能系统被滥用。对算力、推理能力达到一定阈值或应用在特定行业领域的人工智能系统进行登记备案，要求其具备在设计、研发、测试、部署、使用、维护等全生命周期的安全防护能力。

5.2 建立人工智能服务可追溯管理制度。对面向公众服务的人工智能系统，通过数字证书技术对其进行标识管理。制定出台人工智能生成合成内容标识标准规范，明确显式、隐式等标识要求，全面覆盖制作源头、传播路径、分发渠道等关键环节，便于用户识别判断信息来源及真实性。

5.3 完善人工智能数据安全和个人信息保护规范。针对人工智能技术及应用特点，明确人工智能训练、标注、使用、输出等各环节的数据安全和个人信息保护要求。

5.4 构建负责任的人工智能研发应用体系。研究提出“以人为本、智能向善”在人工智能研发应用中的具体操作指南和最佳实践，持续推进人工智能设计、研发、应用的价值观、伦理观对齐。探索适应人工智能时代的版权保护和开发利用制度，持续推进高质量基础语料库和数据集建设，为人工智能安全发展提供优质营养供给。制定人工智能伦理审查准则、规范和指南，完善伦理审查制度。

5.5 强化人工智能供应链安全保障。推动共享人工智能知识成果，开源人工智能技术，共同研发人工智能芯片、框架、软件，引导产业界建立开放生态，增强供应链来源多样性，保障人工智能供应链安全性稳定性。

5.6 推进人工智能可解释性研究。从机器学习理论、训练方法、人机交互等方面组织研究人工智能决策透明度、可信度、纠错机制等问题，不断提高人工智能可解释性和可预测性，避免人工智能系统意外决策产生恶意行为。

5.7 人工智能安全风险威胁信息共享和应急处置机制。持续跟踪分析人工智能技术、软硬件产品、服务等方面存在的安全漏洞、缺陷、风险威胁、安全事件等动向，协调有关研发者、服务提供者建立风险威胁信息通报和共享机制。构建人工智能安全事件应急处置机制，制定应急预案，开展应急演练，及时快速有效处置人工智能安全威胁和事件。

5.8 加大人工智能安全人才培养力度。推动人工智能安全教育与人工智能学科同步发展，依托学校、科研机构等加强人工智能安全设计、开发、治理人才的培养，支持培养人工智能安全前沿基础领域顶尖人才，壮大无人驾驶、

智能医疗、类脑智能、脑机接口等领域安全人才队伍。

5.9 建立健全人工智能安全宣传教育、行业自律、社会监督机制。面向政府、企业、社会公用事业单位加强人工智能安全规范应用的教育培训。加强人工智能安全风险及防范应对知识的宣传，全面提高全社会人工智能安全意识。指导支持网络安全、人工智能领域行业协会加强行业自律，制定提出高于监管要求、具有引领示范作用的人工智能安全自律公约，引导督促人工智能技术研发机构、服务提供者持续提升安全能力水平；面向公众建立人工智能安全风险隐患投诉举报受理机制，形成有效的人工智能安全社会监督氛围。

5.10 促进人工智能安全治理国际交流合作。积极与各国就人工智能开展合作交流，支持在联合国框架下成立国际人工智能治理机构，协调人工智能发展、安全与治理重大问题。推进 APEC、G20、金砖国家等多边机制下的人工智能安全治理合作，加强与共建“一带一路”国家、“全球南方”国家合作，研究成立人工智能安全治理联盟，增强发展中国家在全球人工智能治理中的代表性和发言权。鼓励人工智能企业、机构开展跨国交流合作，分享最佳操作实践，共同制定人工智能安全国际标准。

6. 人工智能安全开发应用指引

6.1 模型算法研发者安全开发指引

（a）研发者应在需求分析、项目立项、模型设计开发、训练数据选用等关键环节，切实践行“以人为本、智能向善”理念宗旨，遵循科技伦理规范，采取开展内部研讨、组织专家评议、科技伦理审查、听取公众意见、与潜在目标用户沟通交流、加强员工安全教育培训等措施。

（b）研发者应重视数据安全和个人信息保护，尊重知识产权和版权，确保数据来源清晰、途径合规。建立完善的数据安全管理制度，确保数据安全性和质量，以及合规使用，防范数据泄露、流失、扩散等风险，人工智能产品终止下线时妥善处理用户数据。

（c）研发者应确保模型算法训练环境的安全性，包括网络安全配置和数据加密措施等。

（d）研发者应评估模型算法潜在偏见，加强训练数据内容和质量的抽查检测，设计有效、可靠的对齐算法，确保价值观风险、伦理风险等可控。

（e）研发者应结合目标市场适用法律要求和风险管理要求，评估人工智能产品和服务能力成熟度。

（f）研发者应做好人工智能产品及所用数据集的版本管理，商用版本应可以回退到以前的商用版本。

（g）研发者应定期开展安全评估测试，测试前明确测试目标、范围和安全维度，构建多样化的测试数据集，涵盖各种应用场景。

（h）研发者应制定明确的测试规则和方法，包括人工测试、自动测试、混合测试等，利用沙箱仿真等技术对模型进行充分测试和验证。

（i）研发者应评估人工智能模型算法对外界干扰的容忍程度，以适用范围、注意事项或使用禁忌的形式告知服务提供者和使用者。

（j）研发者应生成详细的测试报告，分析安全问题并提出改进方案。

6.2 人工智能服务提供者安全指引

（a）服务提供者应公开人工智能产品和服务的能力、局限性、适用人群、场景。

（b）服务提供者应在合同或服务协议中，以使用者易于理解的方式，告知人工智能产品和服务的适用范围、注意事项、使用禁忌，支持使用者知情选择、审慎使用。

（c）服务提供者应在告知同意、服务协议等文件中，支持使用者行使人类监督和控制责任。

（d）服务提供者应让使用者了解人工智能产品的精确度，在人工智能决策有重大影响时，做好解释说明预案。

（e）服务提供者应检查研发者提供的责任说明文件，确保责任链条可以追溯到递归采用的人工智能模型。

（f）服务提供者应提高人工智能风险防范意识，建立健全实时风险监控管理机制，持续跟踪运行中安全风险。

（g）服务提供者应评估人工智能产品与服务在面临故障、攻击等异常条件下抵御或克服不利条件的能力，防范出现意外结果和行为错误，确保最低限度有效功能。

（h）服务提供者应将人工智能系统运行中发现的安全事故、安全漏洞等及时向主管部门报告。

（i）服务提供者应在合同或服务协议中明确，一旦发现不符合使用意图和说明限制的误用、滥用，服务提供者有权采取纠正措施或提前终止服务。

（j）服务提供者应评估人工智能产品对使用者的影响，防止对使用者身心健康、生命财产等造成危害。

6.3 重点领域使用者安全应用指引

（a）对于政府部门、关键信息基础设施以及直接影响公共安全和公民生

命健康安全的领域等重点领域使用者，应审慎评估目标应用场景采用人工智能技术后带来的长期和潜在影响，开展风险评估与定级，避免技术滥用。

（b）重点领域使用者应根据人工智能系统的适用场景、安全性、可靠性、可控性等，定期进行系统审计，加强风险防范意识与风险应对处置能力。

（c）重点领域使用者在使用人工智能产品前，应全面了解其数据处理和隐私保护措施。

（d） 重点领域使用者应使用高安全级别的密码策略，启用多因素认证机制，增强账户安全性。

（e）重点领域使用者应增强网络安全、供应链安全等方面的能力，降低人工智能系统被攻击、重要数据被窃取或泄露的风险，保障业务不中断。

（f） 重点领域使用者应合理限制人工智能系统对数据的访问权限，制定数据备份和恢复计划，定期对数据处理流程进行检查。

（g）重点领域使用者应确保操作符合保密规定，在处理敏感数据时使用加密技术等保护措施。

（h）重点领域使用者应对人工智能行为和影响进行有效监督，确保人工智能产品和服务的运行基于人的授权、处于人的控制之下。

（i） 重点领域使用者应避免完全依赖人工智能的决策，监控及记录未采纳人工智能决策的情况，并对决策不一致进行分析，在遭遇事故时具备及时切换到人工或传统系统等的能力。

6.4 社会公众安全应用指引

（a）社会公众应提高对人工智能产品安全风险的认识，选择信誉良好的人工智能产品。

（b）社会公众应在使用前仔细阅读产品合同或服务协议，了解产品的功能、限制和隐私政策，准确认知人工智能产品做出判断决策的局限性，合理设定使用预期。

（c）社会公众应提高个人信息保护意识，避免在不必要的情况下输入敏感信息。

（d）社会公众应了解人工智能产品的数据处理方式，避免使用不符合隐私保护原则的产品。

（e）社会公众在使用人工智能产品时，应关注网络安全风险，避免人工智能产品成为网络攻击的目标。

（f）社会公众应注意人工智能产品对儿童和青少年的影响，预防沉迷及过度使用。

安全风险与技术应对措施、综合治理措施映射表

安全风险			技术应对措施	综合治理措施
内生（自身）安全风险	模型算法安全风险	可解释性差的风险	4.1.1 （a）	● 推进人工智能可解释性研究 ● 构建以负责任的人工智能研发应用体系
		偏见、歧视风险	4.1.1 （b）	
		鲁棒性弱风险	4.1.1 （b）	
		被窃取、篡改的风险	4.1.1 （b）	
		输出不可靠风险	4.1.1 （a） （b）	
		对抗攻击风险	4.1.1 （b）	
	数据安全风险	违规收集使用数据风险	4.1.2 （a）	● 完善人工智能数据安全和个人信息保护规范
		训练数据含不当内容、被“投毒”风险	4.1.2 （b） （c） （d） （e） （f）	
		训练数据标注不规范风险	4.1.2 （e）	
		数据泄露风险	4.1.2 （c） （d）	
	系统安全风险	缺陷、后门被攻击利用风险	4.1.3 （a） （b）	● 强化人工智能供应链安全保障 ● 人工智能安全风险威胁信息共享和应急处置机制
		算力安全风险	4.1.3 （c）	
		供应链安全风险	4.1.3 （d）	
应用安全风险	网络域风险	信息内容安全风险	4.2.1 （a）	● 实施人工智能应用分类分级管理 建立人工智能服务可追溯管理制度 ● 加大人工智能安全人才培养力度 ● 建立健全人工智能安全宣传教育、行业自律、社会监督机制 ● 促进人工智能安全治理国际交流合作
		混淆事实、误导用户、绕过鉴权的风险	4.2.1 （a）	
		不当使用引发信息泄露风险	4.2.1 （b）	
		滥用于网络攻击的风险	4.2.1 （a）	
		模型复用的缺陷传导风险	4.2.1 （a） （b）	
	现实域风险	诱发传统经济社会安全风险	4.2.2 （b）	
		用于违法犯罪活动的风险	4.2.2 （a） （b）	
		两用物项和技术滥用风险	4.2.2 （a） （b）	
	认知域风险	加剧“信息茧房”效应风险	4.2.3 （b）	
		用于开展认知战的风险	4.2.3 （a） （b） （c）	
	伦理域风险	加剧社会歧视偏见、扩大智能鸿沟的风险	4.2.4 （a）	
		挑战传统社会秩序的风险	4.2.4 （a） （b）	
		未来脱离控制的风险	4.2.4 （b）	