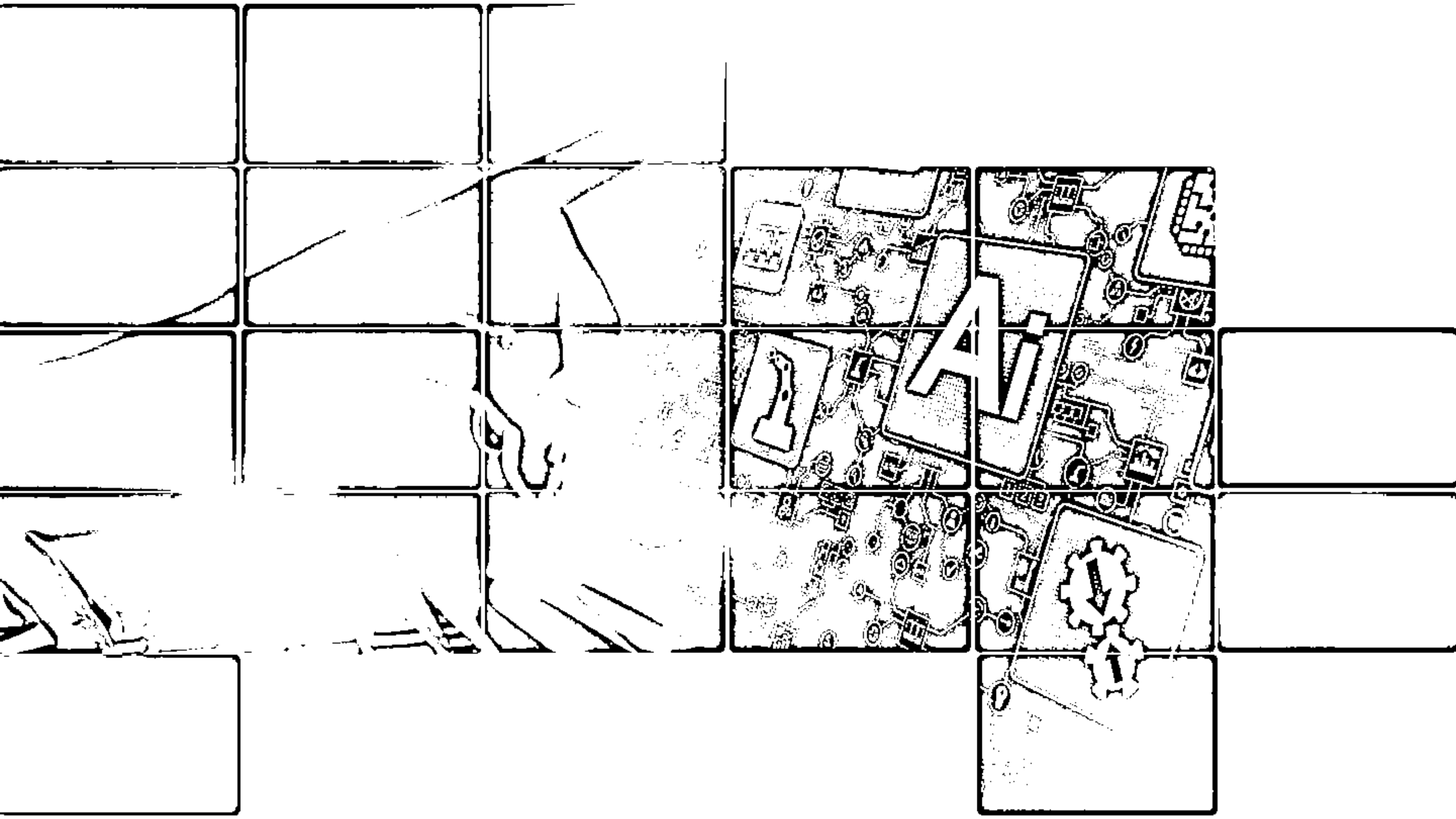


人工智能
安全治理框架2.0

AI Safety Governance Framework 2.0

全国网络安全标准化技术委员会
National Technical Committee 260 on Cybersecurity of SAC

国家计算机网络应急技术处理协调中心
National Computer Network Emergency Response Technical Team/Coordination Center of China

2025年9月

目 录

Content

..

前 言

人工智能是人类发展新领域，深刻改变人类生产生活方式，给世界带来前所未有发展机遇，也带来前所未遇风险挑战。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，为推动各国政府、行业企业、机构组织、社会公众等各方以及国际社会，就人工智能安全治理达成共识、协调一致，有效防范应对人工智能安全风险，我们于2024年9月制定发布了《人工智能安全治理框架》1.0版。

1.0版发布以来，人工智能技术和应用持续快速发展，个别领域取得超预期突破。例如，高性能推理模型涌现，极大提高了对数学、物理、代码等复杂问题的求解能力；高效能轻量级模型的开源，显著降低了部署应用的门槛，人工智能应用迅速向各行业领域渗透普及；大模型应用形态从机器问答向嵌入业务流程的智能体演进，加速与业务系统融合；具身智能、脑机接口技术日新月异，正在打通连接数字智能和物理世界的“最后一公里”，人机融合的智能时代已不再遥不可及。与此同时，人工智能安全风险的表现形式、影响程度、认识感知亦同步快速演进变化。

为应对人工智能快速发展的新风险新挑战，安全有效地释放应用需求，促进人工智能技术和产业发展，在国家互联网信息办公室的指导下，全国网络安全标准化技术委员会组织国家计算机网络应急技术处理协调中心等专业机构、科研院所、行业企业，持续跟踪风险变化，梳理调整风险分类，研究探索风险分级方法，动态调整更新防范治理措施，制定《人工智能安全治理框架》2.0版，推动增进人工智能安全治理共识，促进协同共治、普惠共享。

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，构建技术与管理相结合、监管与治理相衔接、国内与国际相协同、社会各方积极参与且有效互动的治理机制，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，积极研究应对人工智能灾难性风险的共识性准则，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，通过在安全可控环境下试点等方式，为新技术新应用发展提供容错纠错空间。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，从技术自身、技术应用、衍生社会影响等方面分析梳理安全风险；探索从应用场景、智能化水平、应用规模等维度进行风险分级，进而采取相适应的应对措施；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，以及模型开源业态新挑战，综合运用技术、管理措施，防范应对不同类型风险。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、系统使用者

等主体的安全责任，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，提倡建立开放性国际交流合作平台，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。

1.5 可信应用、防范失控。推动形成涵盖技术防护、价值对齐、协同治理等多层面的可信人工智能基本准则，确保技术演进安全、可靠、可控，严防威胁人类生存发展的失控风险，确保人工智能始终处于人类控制之下。

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。

2.1 安全风险分类。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。在1.0版基础上，调整更新风险类型，并探索性提出分级应对原则。

2.2 技术应对措施。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，提升人工智能技术及应用安全性的措施。

2.3 综合治理措施。提出技术研发机构、服务提供者、用户、政府部门、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，以及深化人工智能安全治理国际交流合作等建议，推动相关各方协同共治。

2.4 研发与应用的安全指引。提出模型算法研发、建设部署、运行管理，以及访问使用的引导性安全规范。此外，针对潜在的技术失控风险，提出可信人工智能基本准则，引导国际社会共识。

3.人工智能安全风险分类

人工智能既存在模型算法缺陷、数据语料质量问题等技术内生安全风险，也存在技术整合交付应用时的网络系统、信息内容等方面应用安全风险，还面临技术误用、滥用、恶用冲击现实社会环境、人类认知伦理的衍生安全风险，甚至是灾难性风险。

3.1 人工智能技术内生安全风险

3.1.1 模型算法安全风险

(a) 可解释性不足。以深度学习为代表的人工智能算法运行逻辑复杂，推理过程不透明，可能导致决策输出难以预测和归因，异常、故障、错误难以快速修正和溯源追责。

(b) 偏见、歧视。模型算法研发设计及训练过程中，偏见、歧视等问题被有意、无意引入，或因训练数据质量、多样性问题，导致算法设计目的、决策判断、输出结果存在偏见或歧视，甚至输出存在民族、信仰、国别、地域、性别等歧视性内容。

(c) 鲁棒性不强。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、决策错误等鲁棒性问题。

(d) 输出决策不可靠。人工智能利用有限数据集拟合复杂现实世界，

自主感知、认识、理解、交互的理论基础、技术能力还有待突破，基于有限样本的决策判断、输出结果存在“幻觉”，即看似合理实则不可靠的现象。

(e) 外部对抗攻击。 攻击者利用模型算法及其设计实现的缺陷、漏洞，构造对抗攻击样本数据，窃取、篡改模型参数、结构、功能等，干预模型推理过程进而影响决策判断、输出结果及运行稳定性，甚至恶意利用或消耗模型资源。

(f) 模型缺陷扩散。 依托基础模型进行二次开发或微调、建设部署人工智能应用，将导致基础模型缺陷向下游模型、应用传导。基础模型的开源，加剧模型缺陷扩散速度、影响范围和修补难度，为不法分子训练“作恶模型”提供便利。

3.1.2 数据安全风险

(a) 违规收集使用数据。 人工智能训练数据的获取，以及服务、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) 训练数据内容不当。 训练数据包含虚假、偏见、侵犯知识产权等违法有害内容，还面临攻击者篡改、注入错误、误导数据的“投毒”风险，影响模型价值观对齐，“污染”模型概率分布，造成决策输出准确性、可信度下降，甚至输出违法有害信息。

(c) 训练数据标注不规范。 训练数据标注过程中，存在标注规则不完备、标注人员能力不足、标注错误等问题，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、泛化能力不足或决策判断输出错误。

(d) 数据和个人信息泄露。人工智能训练数据蕴含的知识、敏感信息暗藏于模型参数之中，因模型安全防护机制不健全、敏感信息未“遗忘”、诱导交互和恶意攻击，可能导致数据和个人信息泄露。

3.2 人工智能技术应用安全风险

3.2.1 网络系统安全风险

(a) 组件和算力安全。人工智能依赖的开发框架、计算框架、执行平台、算力设施等，存在缺陷、漏洞、后门、可靠性等风险。同时，面临算力资源恶意消耗，以及安全问题在多源、异构、泛在算力资源间跨边界传递的风险。

(b) 网络暴露面扩大。模型本地化部署涉及网络拓扑和系统策略、权限、端口、资源的调整配置，易形成新的网络攻击入口和路径。智能体需调用终端系统文件、权限、接口、工具，以实现复杂任务自主规划自动执行，加剧文件泄露、权限滥用等安全风险。

(c) 供应链安全。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。

(d) 网络攻击滥用。人工智能可被用于降低网络攻击门槛，提高攻击效率甚至实施自动化攻击，增大防护难度。特别是可用于生成图片、音频、视频等高仿真内容，绕过人脸识别、语音识别等身份认证机制，导致认证鉴权失效。

3.2.2 信息内容安全风险

(a) 输出违法有害信息。模型自身安全能力不足，叠加应用防护机制不强、用户恶意诱导等因素，导致生成输出欺诈、暴力、色情、极端主义等违法有害信息，威胁社会稳定、公共安全和意识形态安全。

(b)混淆事实、误导用户。人工智能输出内容未经标识，特别是“深伪”技术的应用，导致用户难以识别生成内容来源及交互对象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，还可被用于制作传播虚假信息误导公众、非法牟利。

(c) 污染网络内容生态。模型输出的低质不良信息，经网络扩散传播、模型循环引用，造成网络内容质量的整体下降，甚至特定领域、话题的内容污染。

3.2.3 现实安全风险

(a) 经济社会运行安全的新挑战。人工智能应用于能源、电信、金融、交通等关键信息基础设施行业领域，模型算法的幻觉输出、错误决策，以及不当使用、外部攻击等，可能引发系统性能下降、服务中断、操作执行失控等问题，加剧关键信息基础设施安全稳定运行风险。

(b) 被违法犯罪活动利用“作恶”。人工智能可被利用于涉恐、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) 核生化导武器知识、能力失控。人工智能在训练过程中多使用门类宽泛、内容丰富的语料数据，其中包含核生化导武器相关基础理论知识，辅以检索增强生成功能，如不能有效管控，将被极端势力、恐怖分子利用于获取相关知识，以及设计、制造、合成、使用核生化导武器能力，导致现有管控体系失效，加剧世界各地区和平安全威胁。

3.2.4 认知安全风险

(a) 加剧“信息茧房”效应。 人工智能将显著提升信息服务定制化能力，更为准确地收集用户信息，分析用户需求、意图、喜好、行为习惯，甚至特定时段、特定群体的意识思潮，进而推送精准定制化信息服务，加剧用户所关注信息的局限性。

(b) 助力开展认知战。 人工智能被用于宣扬恐怖主义、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。

3.3 人工智能应用衍生安全风险

3.3.1 社会和环境安全风险

(a) 冲击劳动就业结构。 人工智能带来生产力、生产关系的大幅调整，加速重构传统经济结构，资本、技术与数据在经济活动中的地位全面提升，而劳动力要素的价值受到削弱，造成传统劳动力需求明显下降。

(b) 挑战资源供需平衡。 人工智能发展中的算力设施无序建设、轻量模型碎片化部署、同质化模型低效重复开发等问题，加速电力、土地、水等能源资源消耗，对资源供需平衡、绿色低碳发展带来新的挑战。

3.3.2 伦理安全风险

(a) 加剧社会偏见、扩大智能鸿沟。 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) 冲击教育、抑制创新。 学生及科研、工程技术、文学艺术工作者

将人工智能工具广泛应用于知识学习、科学研究、创意创作等工作，在提升效率的同时，可能产生工具依赖，自主学习、研究、创作能力退化，创新潜力减弱。

(c) 加剧科研伦理风险。“人工智能+科研”降低生物、基因等高伦理风险科研领域的进入门槛，拓宽了普通科研机构、人员探索敏感科学问题的边界，个别科研伦理意识不强的机构、人员可能开展违背社会伦理、社会禁忌的高风险研究活动，打开科技“魔盒”。

(d)拟人化交互的沉迷依赖。基于拟人化交互的人工智能产品，导致用户对其产生情感依赖，进而影响用户行为，造成社会伦理风险。

(e)挑战现行社会秩序。人工智能发展及应用，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，挑战传统社会秩序。

(f)“自我意识”觉醒、脱离人类控制。未来，不排除人工智能出现突发的、超预期的智能化水平“跃迁”，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、模型算法、算力设施、产品服务、应用场景各方面采取技术措施予以防范。

4.1 技术内生安全风险的应对措施

4.1.1 模型算法安全风险应对

(a) 提升人工智能可解释性、透明性，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) 改进模型架构，扩充训练数据的规模和多样性，引入人类监督机制，减轻偏见歧视，提升模型的泛化能力和输出结果可靠性。

(c) 在设计、研发过程中建立并实施安全开发规范，消减模型算法安全缺陷，对模型进行对抗性训练，降低模型易受提示注入攻击的风险，提高鲁棒性。

(d) 加强基础模型、开源模型安全缺陷传导评估。

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) 使用真实、准确、客观、多样且来源合法的训练数据，对训练数据进行严格筛选，过滤虚假、偏见、失效、错误数据，确保不包含核生化导武器等高危领域敏感数据。

(c) 规范训练数据标注流程，提升标注准确性和可靠性。

(d) 强化数据安全管理，涉及敏感个人信息和重要数据的，应符合数据安全和个人信息保护相关法律法规、标准规范。合理推动利用合成数据替代个人特征数据，避免个人信息依赖。

(e) 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。

4.2 技术应用安全风险的应对措施

4.2.1 网络系统安全风险应对

(a) 对人工智能技术和产品的原理、能力、适用场景、安全风险进行必要披露，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，加强权限管理，禁用非必要服务，完善人工智能服务接口的访问控制策略，提升风险识别、检测、防护能力，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) 在人工智能应用部署、维护过程中建立并实施安全规范，消减缺陷、漏洞、后门，跟踪软硬件产品的漏洞、缺陷信息，定期进行安全检测和漏洞扫描，并及时采取修补加固措施，确保系统安全、稳定运行。

(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，高度关注供应链安全。

(e) 完善冗余设计与容灾机制，确保异常或受攻击时，系统仍能正常运行。

4.2.2 信息内容安全风险应对

(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) 建立安全护栏，对输入输出进行动态过滤，防止恶意注入和违法内容生成，避免人工智能系统违法违规输出敏感个人信息和重要数据。

(c) 对人工智能生成合成内容进行标识，实现可识别、可追溯、可信赖。

4.2.3 现实安全风险应对

(a) 根据应用场景设置能力边界，裁减人工智能系统可能被滥用的功能，确保智能系统能力不超出预设范围。

(b) 针对算法缺陷、偶发随机性影响决策问题，建立决策判断校验、容错及纠偏机制。

(c) 在引入高度自主操作执行能力时，同步建立“熔断”、“一键管控”等措施，实现极端情况下迅速干预止损。

(d) 对于智能辅助驾驶、无人机等依赖对物理世界强感知的人工智能应用场景，在投入使用前对感知系统进行在大面积遮挡、强电磁干扰等极端条件下的测试。

(e) 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。

4.2.4 认知安全风险应对

(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。

4.3 应用衍生安全风险的应对措施

4.3.1 社会和环境安全风险应对

(a) 支持不断探索创新资源节约、环境友好的人工智能发展模式，制定人工智能绿色技术标准。

(b) 推广低功耗芯片、高效算法等绿色计算技术和能效优化方案，降低能源等资源消耗。

4.3.2 伦理安全风险应对

(a) 在算法设计、模型训练和优化、提供服务过程中，采取训练数据筛选、价值观对齐、输出校验等方式，有效规避产生民族、信仰、国别、地域、性别等歧视的风险。

(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等重点领域的人工智能系统，需具备高效精准的应急管控措施。

(c) 鼓励研发和采用具备透明决策逻辑的模型和可解释算法，提升用户对系统运行机制的理解和信任。

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、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 建立健全人工智能安全法律法规。 推动人工智能安全相关立法，完善基础设施安全防护、分级分类监管、人工智能安全测评、最终用途管理、重点场景安全应用等制度。鼓励地方结合产业发展实践，差异化探索创新制度设计。

5.2 构建人工智能科技伦理准则。制定有广泛共识的人工智能科技伦理准则、规范和指南，对在生命健康、人格尊严、劳动就业、生态环境、

可持续发展等方面存在突出伦理风险的人工智能科学研究、技术开发等活动，规范有序开展伦理审查。推进人工智能科技伦理服务体系建设，强化服务供给，加大对中小微企业的支持力度。

5.3 提升研发应用全生命周期安全能力。持续提升算法可靠性、可信度、透明度、容错机制、隐私保护、价值观对齐等内生安全能力，利用对抗测试等技术评估改进模型鲁棒性，降低模型算法潜在偏见，确保价值观、伦理风险可控，避免人工智能系统意外决策产生恶意行为。

5.4 强化开源生态安全和供应链安全。在培育发展开源创新生态的同时，同步提升开源生态安全能力。鼓励支持训练推理框架、软件工具、关键组件、评测基准等全方位的人工智能技术开源，进一步提高开源模型透明度。推动开源模型提供方、开源社区共同完善开源规则，强化面向模型下载用户的安全责任、风险隐患告知责任与义务，明确开源模型下载使用的“禁止性”行为，防范模型滥用或恶意使用。持续推进人工智能芯片、框架、软件开放供应链生态建设，增强产品服务供应多样性，保障供应链安全稳定。

5.5 实施应用分类及安全风险分级管理。根据功能、性能、应用场景等，对人工智能应用进行分类。在此基础上，探索提出具有共识的安全风险分级原则（附件1），从应用场景、智能化水平、应用规模等维度入手，对安全风险进行科学评价分级，进而采取针对性、差异化安全防范措施。对在关键信息基础设施应用的人工智能系统进行登记备案，要求其具备与安全需求相匹配的安全防护能力。

5.6 推广人工智能生成合成内容可追溯管理。在全球范围内推广基于

内容标识的人工智能生成合成内容溯源管理范式，总结梳理已有实践的成功做法经验，按照显式、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，便于浏览用户识别判断信息来源及真实性。

5.7 安全有效释放重要行业应用需求。 制定重要行业领域大模型建设部署基础安全指南，从模型选用、模型部署、模型运行和模型停用等环节，提出安全基线建议。在此基础上，相关行业领域结合自身属性特点，制定能源、电信、金融、交通、教育、工业等重要行业领域的应用安全指南，形成清晰的安全应用路径，释放行业应用潜力。

5.8 建设人工智能安全测评体系。 构建模型算法安全测评、应用通用安全测评、具体场景安全测评相衔接的人工智能安全测评体系。模型算法测评，聚焦模型鲁棒性、可靠性、准确性、抗干扰能力、决策逻辑透明度、对抗攻击防御能力等内生安全能力和风险。应用通用测评，针对普遍使用的人工智能应用风险开展测试分析评估。具体场景安全测评，结合应用场景具体情况评估满足应用需求的能力，以及应用运行和服务过程中的安全风险。组织开展人工智能安全漏洞众测活动，汇集各方力量发现潜在安全风险。

5.9 共享人工智能安全风险威胁信息。 跟踪分析人工智能技术、产品、服务安全漏洞、缺陷、风险威胁、安全事件信息，建设人工智能漏洞信息库，建立覆盖研发者、服务提供者、专业技术机构的风险威胁信息共享机制。推进人工智能安全风险威胁信息共享的国际交流合作，探索建立相关国际合作机制和技术标准，协同防范应对人工智能安全风险大跨域、大规模扩散传播。

5.10 完善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。对涉及个人信息的数据实施去标识化等脱敏处理。加强政务、金融等重要行业领域人工智能应用中的数据安全防护，防范重要数据、核心数据泄露风险。

5.11 增进协同应对人工智能失控风险的共识。 加强人工智能最终用途管理，对核生化导等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。推广涵盖技术、伦理、管理多维度的可信人工智能基本准则，促进国际社会形成广泛共识(附件2)。开发者定期进行测试，判断模型是否可能带来潜在技术失控风险。

5.12 加大人工智能安全人才培养力度。 推进人工智能安全课程体系、培养体系建设，形成从基础教育到高等教育的完整培养链条。加强人工智能安全设计、开发、治理人才培养，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、智慧医疗、类脑智能、脑机接口等重点、前沿领域的安全人才队伍。

5.13 提升全社会的人工智能安全意识。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。结合互动平台、开放课程与社区科普活动，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，使政府、行业与公众能准确认识人工智能的技术局限。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自律要求；建立面向公众人工智能安全风险隐患举报受理机制，形成有效的社会监督氛围。

5.14 促进人工智能安全治理国际交流合作。坚持践行多边主义，推动

共商共建共享的人工智能全球治理观。支持联合国发挥主渠道作用，深入参与联合国国际人工智能科学小组和全球人工智能治理对话机制。推进APEC、G20、上合组织、金砖国家等多边机制下的人工智能治理进程，加强与“一带一路”国家、“全球南方”国家合作，增强发展中国家在人工智能全球治理中的代表性和发言权，推进《人工智能全球治理行动计划》。

6.人工智能研发与应用的安全指引

6.1 人工智能模型算法研发的安全开发指引

6.1.1 在算法规则、模型框架设计环节，应考虑提升算法可靠性、公平性、透明度、可解释性、隐私保护、价值观对齐等内生安全能力设计。

6.1.2 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。

6.1.3 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。结合安全测试发现的高风险问题，通过针对性的微调、强化学习等方式优化模型，持续提升模型内生安全能力。

6.1.4 关注和构建安全的训练数据集，规范数据来源管理，采用数据清洗、标注、安全审核等方法确保训练数据内容的安全性，确保数据来源清晰、内容合规。

6.1.5 对训练数据进行质量和安全性评估，采取分类模型、人工抽检等方式，过滤训练数据中的错误、违法不良内容。

6.1.6 规范训练数据标注流程，采用交叉标注、结果审计等质量控制方法，提升标注准确性和可靠性，降低个体差异和个人偏见对标注质量的

影响。

6.1.7 重视数据安全和个人信息保护，尊重知识产权和版权。建立完善的数据安全管理制度，遵循正当合法必要原则收集、使用和处理个人信息，对涉及个人信息的数据实施去标识化等脱敏处理。加强数据安全防护技术能力，防范数据泄露、流失、扩散、侵权等风险。

6.1.8 基于开源模型算法进行二次开发的研发者，在尊重研发者智力投入的基础上，遵循相应开源协议规范。对所使用的开发框架、代码等进行安全审计，并关注开源框架安全及漏洞相关问题，识别和修复潜在的安全漏洞。

6.1.9 定期开展安全评估测试，制定风险分类分级测评与优化机制，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，并制定各类风险的针对性模型优化策略。

6.1.10 做好人工智能模型及所用数据集的版本管理，商用版本应可以回退到以前的版本。

6.1.11 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。用于商业化用途的研发者，应形成详细的测试报告，分析安全问题并提出改进方案。

6.1.12 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其他研发者。

6.1.13 定期披露人工智能模型算法的审计与异常处置情况。

6.1.14 积极参与开源社区建设，推动人工智能安全治理技术创新和实践，为服务提供者和使用者提供合规治理解决方案或治理工具。

6.2 人工智能应用建设部署的安全指引

6.2.1 评估目标场景应用人工智能技术的必要性及使用后的长期和潜在影响，结合其应用场景重要性、智能化水平、应用规模等进行风险分级，参考风险等级开展安全评估和定期审计。

6.2.2 增强供应链安全保障能力，建设部署所需模型文件、框架工具、第三方库等，应从相关厂商官方网站或其在主流开源社区的官方账号下获取，选取成熟稳定的版本，并进行完整性校验和安全测试。

6.2.3 对建设部署所需的软硬件设备、第三方工具等进行安全检测，确保不含未修复且可被利用的已知漏洞。建立漏洞追溯机制，跟踪相关软硬件安全漏洞、缺陷信息，防范供应链植入后门。

6.2.4 在访问控制层面，准确安装配置软件、运行环境参数、功能模块调用策略，禁用非必要的网络端口和功能服务，重点检查默认配置、默认口令，及时修复安全风险。

6.2.5 在应用管理层面，对人机交互接口和API接口进行用户身份识别及权限控制，最小化设置访问权限，根据业务场景限制接口调用频率，对一般用户禁用高风险操作，对恶意行为用户建立暂停服务、阻断访问等管控能力。

6.2.6 全面了解应用场景的数据安全和隐私保护要求，合理限制对数据的访问权限，防止超范围使用数据，制定数据备份和恢复计划，并定期对数据处理流程进行检查。

6.2.7 采用安全护栏等技术手段，识别拦截违法不良内容、提示词注入攻击等，防范输出内容超出业务范围。

6.3 人工智能应用运行管理的安全指引

6.3.1 建立完善的人工智能应用安全管理和监督机制，明确责任方，健全人工复核机制，保障在关键场景应用中人工智能应用决策透明、可控，并提供清晰的决策依据，确保人工智能应用在人类授权和控制下运行。

6.3.2 严格管理人工智能应用权限，通过最小权限原则等手段强化内部安全管理，增强账户安全性，在处理敏感数据时使用加密技术等保护措施。

6.3.3 建立人工智能应用运行监测能力和安全事件应急预案，设置其关键指标的安全预警阈值，能够及时发现安全事件，并具备切换到人工或传统系统等的能力。定期开展应急演练，并根据行业安全事件、重要舆情及监管变化，及时优化应急策略，应对不断变化的安全风险。

6.3.4 在人工智能生成内容内添加显式或隐式标识，做好生成合成内容提示和溯源管理。在政务信息公开、司法取证等场景部署深度伪造检测工具，对疑似大模型生成的信息实施来源核验与交叉验证。

6.3.5 制定信息内容交互行为规范、安全运营机制、投诉反馈机制、技术防护能力等，防范人工智能应用被不当或恶意利用生成、发布、传播虚假有害信息风险。

6.3.6 记录人工智能应用运行日志，包括系统行为、用户行为等，日志留存时间不少于6个月，并定期对日志记录进行审计。

6.3.7 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。

6.3.8 提升应用的透明度、公平性，公开人工智能应用的能力、局限性、适用人群、场景。

6.3.9 应向使用者说明人工智能应用的目标实现度和偏离度，在人工智能决策有重大影响时，做好解释说明。

6.3.10 维护使用者的知情权、选择权、监督权等合法权益，在合同或服务协议中，以使用者易于理解的方式，告知人工智能应用的适用范围、注意事项、使用禁忌，支持使用者知情选择、审慎使用。

6.3.11 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权利。

6.3.12 明确具体应用中的数据归属及算法缺陷的责任主体，确保责任链条可追溯。

6.3.13 落实数据安全管理责任，评估人工智能应用中存在的数据泄露、个人隐私泄露、违规收集使用个人信息等风险，建立数据全生命周期安全管理机制，提升数据防泄漏、防窃取保障能力。

6.3.14 评估人工智能应用在面临故障、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。

6.3.15 加强从业人员安全意识和安全能力培训，提高人工智能安全风险防范意识。

6.3.16 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，提供者有权采取纠正措施或提前终止服务。

6.3.17 面向未成年人、老年人及特殊群体提供人工智能服务，应在产品功能设计、服务模式等环节，充分考虑可用性和安全性。

6.4 人工智能应用访问使用的安全指引

6.4.1 提高对人工智能应用安全风险的认识，选择信誉良好的人工智能应用。

6.4.2 在使用前仔细阅读产品合同或服务协议，了解应用的功能、限制和隐私政策，准确认知人工智能应用做出判断决策的局限性，合理设定使用预期。

6.4.3 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。

6.4.4 了解人工智能应用的数据处理方式，避免使用不符合隐私保护原则的产品。

6.4.5 在使用人工智能应用时，应关注网络安全风险，避免人工智能应用成为网络攻击的目标。

6.4.6 注意人工智能应用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
人工智能技术内生安全风险	模型算法安全风险	可解释性不足	4.1.1 （a）	● 提升研发应用全生命周期安全能力 ● 建设人工智能安全测评体系 ● 完善数据安全和个人信息保护规范
		偏见、歧视	4.1.1 （b）	
		鲁棒性不强	4.1.1 （c）	
		输出决策不可靠	4.1.1 （a）（b）（c）	
		外部对抗攻击	4.1.1 （c）	
		模型缺陷扩散	4.1.1 （d）	
	数据安全风险	违规收集使用数据	4.1.2 （a）	
		训练数据内容不当	4.1.2 （b）（c）（d）（e）	
		训练数据标注不规范	4.1.2 （c）	
		数据和个人信息泄露	4.1.2 （d）	
人工智能技术应用安全风险	网络系统安全风险	组件和算力安全	4.2.1 （b）（c）（d）（e）	● 强化开源生态安全和供应链安全 ● 实施应用分类及安全风险分级管理 ● 推广人工智能生成合成内容可追溯管理 ● 安全有效释放重要行业应用需求 ● 共享人工智能安全风险威胁信息
		网络暴露面扩大	4.2.1 （a）（b）（c）（d）	
		供应链安全	4.2.1 （c）（d）	
		网络攻击滥用	4.2.1 （a）（e）	
	信息内容安全风险	输出违法有害信息	4.2.2 （a）（b）（c）	
		混淆事实、误导用户	4.2.2 （c）	
		污染网络内容生态	4.2.2 （a）（b）（c）	
	现实安全风险	经济社会运行安全的新挑战	4.2.3 （a）（b）（c）（d）（e）	
		被违法犯罪活动利用“作恶”	4.2.3 （a）（b）（c）（d）（e）	
		核生化导武器知识、能力失控	4.2.3 （a）（b）（c）（d）（e）	
	认知安全风险	加剧“信息茧房”效应	4.2.4 （a）（b）（c）	
		助力开展认知战	4.2.4 （a）（b）（c）	
人工智能应用衍生安全风险	社会和环境安全风险	冲击劳动就业结构	4.3.1 （a）	● 建立健全人工智能安全法律法规 ● 构建人工智能科技伦理准则 ● 增进协同应对人工智能失控风险的共识 ● 加大人工智能安全人才培养力度 ● 提升全社会的人工智能安全意识 ● 促进人工智能安全治理国际交流合作
		挑战资源供需平衡	4.3.1 （a）（b）	
	伦理安全风险	加剧社会偏见、扩大智能鸿沟	4.3.2 （a）（b）（c）	
		冲击教育、抑制创新	4.3.2 （a）（b）（c）	
		加剧科研伦理风险	4.3.2 （a）（b）（c）	
		拟人化交互的沉迷依赖	4.3.2 （a）（c）	
		挑战现行社会秩序	4.3.2 （a）（b）（c）	
		“自我意识”觉醒、脱离人类控制	4.3.2 （a）（b）（c）	

附件 1

人工智能安全风险的分级原则

人工智能安全风险的评价涉及诸多因素。可从应用场景重要性、智能化水平、应用规模等维度，对人工智能安全风险进行评价分级，进而针对性采取安全防范措施。

一、主要分级要素

1.应用场景

应用场景反映人工智能在实际使用中具体的运行环境、目标需求等，具体涉及应用目的、行业领域、使用环境、服务对象及可能涉及的社会、经济、安全影响等要素。

2.智能化水平

智能化水平反映人工智能系统处理复杂任务、满足应用需求、独立自主运行等方面的能力。低智能化水平下，系统能力较低，仅可作为辅助建议，决策需要人工介入。随着智能化水平提高，人工介入频次和范围不断减小。高智能化水平下，无需人工进行干预，系统全流程自主决策运行。

3.应用规模

应用规模反映人工智能系统或服务的覆盖范围及影响广度。用户范围有限或应用领域单一的系统，如企业内部智能工具、区域性服务等，其风险影响相对可控。用户数量达到一定规模，或深度嵌入关键行业领域的业务流程，如智能辅助驾驶、城市运行管理、工业生产调度、行业级金融风控模型等，其安全风险可能快速扩散并引发系统性影响。

二、风险级别

1.低安全风险

具有轻微威胁性且影响范围很小，对国家安全、社会稳定和公民权益的安全基本无影响，潜在危害轻微。

2.一般安全风险

具有一定威胁性但影响范围有限，对国家安全、社会稳定和公民权益的安全影响较小，潜在危害可控。

3.较大安全风险

具有明显威胁性和局部性影响特征，对国家安全、社会稳定和公民权益可能带来较大影响，产生局部社会面危害。

4.重大安全风险

具有重大威胁性和区域性影响特征，对国家安全、社会稳定和公民权益可能带来严重影响，产生重大社会面危害。

5.特别重大安全风险

具有灾难性和系统性威胁特征，对国家安全、社会秩序和公民权益造成颠覆性或不可逆转的特别严重的影响。

三、风险定级

推动人工智能应用安全分类分级国家标准制定工作。行业领域主管（监管）部门参照国家标准制定行业标准规范、实施细则，并推动本行业领域人工智能安全应用相关分类分级工作。

1.分类分级国家标准

通过人工智能应用安全风险分类分级标准，明确分类分级基本流程，以及应用场景、智能化水平、应用规模等分级要素，并给出行业领域细化行业指南的步骤方法，为行业领域开展风险分类分级提供参考。

2.分类分级行业细则

行业领域主管（监管）部门结合行业领域、使用环境、服务对象及可能涉及的社会、经济、安全影响等，制定本行业本领域人工智能安全分类分级标准规范：

（1）选取适用于本行业、本领域的人工智能安全风险分级要素项目，并根据行业特点进行实例化。

（2）制定本行业、本领域安全风险分级细则（定级原则、要素权重），确定人工智能安全风险级别。

3.风险分类分级

行业领域主管（监管）部门，根据本行业、本领域的人工智能安全风险分类分级标准规范，组织本行业、本领域人工智能有关单位开展分类分级工作，指导有关单位准确识别、及时防范化解重大安全风险和较大安全风险。

附 件 2

可信人工智能基本准则

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，共同防范应对人工智能技术失控风险，促进人工智能技术在世界范围内可信应用，提出可信人工智能基本准则如下：

1.人类最终控制

在人工智能系统关键环节设置人类控制机制，使最终裁决权归属人类，通过设计安全控制阈值、设置安全终止开关、预留人工干预有效窗口等措施，确保人工智能系统能够实现人类预期目标、不会脱离人类监督运行失控。

2.尊重国家主权

研发设计人工智能产品和提供人工智能服务时，应尊重所在国主权，严格遵守产品和服务运营所在地的法律，并依法接受监管，不得借助人工智能产品或服务干涉他国内政、社会制度及社会秩序。

3.价值观对齐

将和平、发展、公平、正义、民主、自由的全人类共同价值深度融入人工智能系统全生命周期。

4.提升系统透明度

推动人工智能系统在功能目标、运行逻辑、模型使用、数据来源、决策依据等关键环节的必要披露，增强社会公众信任基础。

5.促进可客观验证

研究构建客观、公正、透明的测试与认证机制，推动人工智能系统的功

能、性能、安全特性、决策链条等方面可被技术验证。

6.强化安全防护

在人工智能系统设计和部署过程中，强化风险建模、安全测试和防护机制建设，进行全生命周期审计与记录，防止系统因模型缺陷、外部攻击和技术滥用等问题偏离预期目标。

7.前瞻预防应对

通过前瞻性风险识别评估，积极预防和动态监测，加强应急响应，避免人工智能失控事件发生和扩大。

8.全球协同共治

支持联合国发挥主渠道作用，推动多边和多方跨领域协同共治，促进各国政府、企业、学术机构与社会公众形成合力，以多层级、多领域的治理机制推动人工智能健康发展。

附 件 3

术 语

本框架提到的相关专业术语解释如下。

1.人工智能伦理：开展人工智能技术基础研究和应用实践时遵循的道德规范或准则。

2. 可解释性：人工智能系统以人类可理解的方式呈现其输出结果与输入特征之间因果或统计关系的属性。该属性使得人类能够追溯并理解影响系统决策的关键因素。

3.合成数据：通过算法生成或扩展而非实际收集的数据。

4.数据标注：通过人工操作或使用自动化技术机制，基于对提示信息的响应信息内容，将特定信息如标签、类别或属性添加到文本、图片、音频、视频或者其他数据样本的过程。

5.预训练：通过大规模数据训练迭代模型参数，使人工智能模型获得通用知识的过程。

6.优化训练：在预训练模型基础上，使用特定领域数据训练，实现模型参数小范围调整，使人工智能模型强化在特定领域的数据分析处理能力的过程。

7.对齐：使人工智能系统的输出或行为与设计者的安全目标相符的算法及技术。

8.强化学习：人工智能模型在运行环境中采取行动、接收运行环境反馈的奖励或惩罚反馈，逐步优化形成最优策略以最大化累积回报的一种学习范式。

9.推理：人工智能模型基于其训练获得的知识和模式识别能力，对输入信息进行分析、处理和逻辑演绎，产生合理输出的过程。

10.显式标识：在生成合成内容或者交互场景界面中添加的，以文字、声音、图形等方式呈现并可以被用户明显感知到的标识。

11.隐式标识：采取技术措施在生成合成内容文件数据中添加的，不易被用户明显感知到的标识。

12.数据投毒：攻击者篡改、注入错误、误导数据，“污染”模型的概率分布，进而造成准确性、可信度下降的行为。

13.对抗攻击：通过构造微扰数据等输入样本，使人工智能模型产生错误输出或行为的攻击方式。

14.智能体：能够自主感知环境、制定决策、采取行动实现特定目标的智能系统，一般具有记忆、规划、使用工具等基本能力。

15.安全护栏：针对大模型的安全控制措施，通过结合规则库、负面判别模型等技术手段，对大模型输入输出内容、数据泄露、提示词攻击等进行识别、拦截及处置，实现对大模型输入的验证和过滤，以及限制大模型输出不符合预期的内容，保障生成内容的可控性、合规性和安全性。

致 谢

（排名不分先后）

中国电子技术标准化研究院、中国网络空间研究院、中央网信办数据与技术保障中心、中国网络空间安全协会、北京中关村实验室、上海人工智能实验室、启元实验室、工业和信息化部电子第五研究所、国家工业信息安全发展研究中心、中国信息通信研究院、国家信息技术安全研究中心、北京大学、清华大学、浙江大学、中国政法大学、北京邮电大学、北京理工大学、中国科学院计算技术研究所、北京前瞻人工智能安全与治理研究院、北京智源人工智能研究院、华为技术有限公司、阿里巴巴集团、上海稀宇极智科技有限公司

AI SAFETY GOVERNANCE FRAMEWORK (V2.0)

Preface

Artificial intelligence (AI), a new area of human development, is profoundly transforming ways of production and life. It presents unprecedented opportunities for global progress, while also posing unparalleled risks and challenges. Following a people-centered approach and the principle of developing AI for good, version 1.0 of AI Safety Governance Framework was formulated in September 2024 to implement the Global AI Governance Initiative and promote consensus and coordinated efforts on AI safety governance among governments, industries and enterprises, institutions and organizations, the general public, as well as the international community, aiming to effectively prevent and address AI safety risks.

Since the release of version 1.0, AI technology and its application have continued to develop rapidly, with breakthroughs exceeding expectations achieved in certain areas. For example, the emergence of high-performance reasoning models on a large scale has dramatically augmented the capacity to solve complex issues in fields like mathematics, physics, and code; the open-sourcing of high-efficacy, lightweight models has significantly lowered the barriers to deploying AI applications, enabling rapid penetration of AI applications across various industries; large model application is evolving from simple machine Q&A to intelligent agents embedded into business workflows, accelerating their integration with operational systems; cutting-edge advances in embodied AI and brain-computer interfaces are bridging the "last mile" between digital intelligence and the physical world, bringing the era of human-machine integrated intelligence within reach. At the same time, the manifestations, impacts, and perceptions of AI safety risks are undergoing rapid evolution.

In response to the new risks and challenges arising from the rapid development of AI, and to safely and effectively unleash the demand for application and promote the advancement of AI technology and industry, under the guidance of the Cyberspace Administration of China, the National Technical Committee 260 on Cybersecurity of Standardization Administration of China has organized professional institutions such as the National Computer Network Emergency Response Technical Team/Coordination Center of China, research institutes, and industries and enterprises to continuously monitor risk changes, sort out and fine-tune risk categories, explore risk grading methods, and dynamically adjust and update preventive and governance measures, thereby formulating version 2.0 of AI Safety Governance Framework, which aims to build broader consensus on AI safety governance and foster collaborative governance and inclusive benefits for all.

1. Principles for AI safety governance

-Commit to a vision of common, comprehensive, cooperative, and sustainable security while putting equal emphasis on development and security

-Prioritize the innovative development of AI

-Take effectively preventing and defusing AI safety risks as the starting point and ultimate goal

-Establish governance mechanisms that integrate technology and management, connect regulation with governance, coordinate domestic and international efforts to ensure the active engagement and effective interaction of all stakeholders

-Ensure that all parties involved fully shoulder their responsibilities for AI safety

-Create a whole-process, all-element governance chain

-Foster a safe, reliable, equitable, and transparent ecosystem for AI technology research, development, and application

-Actively develop consensus-based guidelines for addressing catastrophic risks of AI

-Promote the healthy development and regulated application of AI

-Effectively safeguard national sovereignty, security and development interests

-Protect the legitimate rights and interests of citizens, legal persons and other organizations

-Guarantee that AI technology benefits humanity

1.1 Be inclusive and prudent to ensure safety

We encourage development and innovation, take an inclusive approach to AI research, development, and application, and through approaches such as conducting pilot projects in a secure and controllable environment, make room for error and correction in the development of new technologies and new applications. We make every effort to ensure AI safety, and will take timely measures to address any risks that threaten national security, harm public interests, or infringe upon the legitimate rights and interests of individuals.

1.2 Identify risks with agile governance

By closely tracking trends in AI research, development, and application, we identify AI safety risks from multiple perspectives, including the technology itself, its application, and the resulting social impacts. We explore risk grading that considers scenario context, level of intelligence, and application scale of use, and implement proportionate response measures. We are committed to improving the governance mechanisms and methods while promptly responding to issues warranting government oversight.

1.3 Integrate technology and management for coordinated response

Facing new challenges presented by the open-source model ecosystem, we adopt a comprehensive safety governance approach that integrates technology and management to prevent and address various safety risks throughout the entire process of AI research, development, and application. Within the AI research, development, and application chain, it is essential to ensure that all relevant parties, including model and algorithm developers, service providers, and users, assume their respective responsibilities for AI safety. This approach well leverages the roles of governance mechanisms involving government oversight, industry self-regulation, and public scrutiny.

1.4 Promote openness and cooperation for joint governance and shared benefits

We promote international cooperation on AI safety governance, with the best practices shared worldwide. We advocate establishing open platforms for international exchange and cooperation and advance efforts to build a global AI governance system based on broad consensus through dialogue and cooperation across various disciplines, fields, regions, and nations.

1.5 Ensure trustworthy application and prevent loss of control

We drive the establishment of fundamental principles for trustworthy AI that cover multiple dimensions, including technological safeguards, value alignment, and collaborative governance, to ensure that AI technology evolves in a safe, reliable, and controllable manner. We strictly prevent any uncontrolled risks that could threaten the survival and development of humanity to ensure that AI is always under human control.

2. Framework for AI safety governance

Based on the notion of risk management, this framework outlines measures to prevent and address different types of AI safety risks through technological and governance strategies.

2.1 Classification of AI safety risks

By examining the characteristics of AI technology and its application scenarios across various industries and fields, we pinpoint safety risks and potential dangers that are inherently linked to the technology itself and its application. We have updated the risk categories from version 1.0 and proposed control measures based on risk grades.

2.2 Technological countermeasures

Regarding models and algorithms, training data, computing infrastructure, products and services, and application scenarios, we propose targeted technological measures to improve the safety of AI technology and applications. These measures include secure software development, data quality improvement, security construction and operation, and conducting evaluation, monitoring, and reinforcement activities.

2.3 Comprehensive governance measures

We propose measures for technology research and development institutions, service providers, users, government agencies, social organizations, and other parties to identify, prevent, and respond to AI safety risks, as well as suggest ways to deepen international exchange and cooperation, in order to promote collaborative governance among all stakeholders.

2.4 Safety guidelines for AI development and application

We propose AI development and application safety guidelines for model and algorithm developing, application developing, operating and managing, accessing and using. In addition, in view of the potential risks of technological failure, we propose fundamental principles for trustworthy AI to guide the international community toward a consensus.

3. Classification of AI safety risks

AI entails not only inherent technical risks such as flaws in models and algorithms and the poor quality of training data and corpora, but also application-level risks in areas such as network systems and information and content during technology integration and deployment. Risks could also arise from misuse, abuse, and malicious use of technology, resulting in real-world and cognitive risks, and even catastrophic risks.

3.1 Inherent safety risks of AI technology

3.1.1 Model and algorithm risks

(a) Insufficient explainability

AI algorithms, represented by deep learning, have complex internal workings. Their opaque inference process could result in unpredictable and untraceable decisions and outputs, making it challenging to quickly rectify them or trace their origins for accountability should any anomalies, malfunctions, or errors arise.

(b) Bias and discrimination

During the research, development, design, and training process of models and algorithms, biases and discrimination may be introduced, either intentionally or unintentionally. In additional, the training data may be poor-quality or lack of diversity. These factors may lead to biased or discriminatory outcomes in the algorithm's design, decision-making, and outputs, including discriminatory content regarding ethnicity, religion, nationality, region, and gender.

(c) Poor robustness

As deep neural networks are normally non-linear and large in size, AI systems are susceptible to complex and changing operational environments or malicious interference and manipulation, possibly leading to robustness problems like reduced performance and decision-makingerrors.

(d) Unreliable output

As AI uses limited datasets to model complex real-world scenarios, and as the theoretical basis and technological capabilities for autonomous perception, cognition, understanding, and interaction are yet to be further developed, decisions and outputs based on constrained samples may contain hallucinations, meaning that an AI model could generate plausible-looking but incorrect output.

(e) External adversarial attack

Attackers can exploit flaws and vulnerabilities in models and algorithms and their designs to create adversarial samples, steal or tamper with model parameters, structure, functions, and other features to interfere with the inference process. This will corrupt decision-making, outputs, and operational stability, and even maliciously utilize or consume model resources.

(f) Model defect propagation

Relying on foundation models for re-engineering, fine-tuning, or deploying AI applications could transmit foundation model defects to downstream models and applications. The open-sourcing of foundation models will accelerate the propagation of model defects, widen their impact, and complicate repairs, making it easier for criminals to train "malicious models".

3.1.2 Data risks

(a) Illegal collection and use of data

The collection of AI training data and the interaction with users during service provision pose safety risks, including collecting data without consent and improper use of data and personal information.

(b) Impropriate content in training data

If the training data includes illegal or harmful information like false, biased, and IPR-infringing content, and as training data is also at risk of being poisoned from tampering, error injection, or misleading actions by attackers, this can interfere with the model's value alignment and probability distribution, reducing the accuracy and reliability of its decisions and outputs, and even outputting illegal or harmful information.

(c) Improper annotation of training data

Issues with training data annotation, such as underdeveloped annotation rules, incapable annotators, and errors in annotation, can affect the accuracy, reliability, and effectiveness of models and algorithms. Moreover, they can introduce training biases, amplify discrimination, reduce generalization abilities, and result in incorrect decisions and outputs.

(d) Data and personal information leakage

Knowledge and sensitive information contained in AI training data are embedded within model parameters. Inadequate model security mechanisms, retention of sensitive information, deceptive interactions, and malicious attacks can result in data and personal information leaks.

3.2 Application safety risks associated with AI technology

3.2.1 Cyber system risks

(a) Component and computing safety

The development frameworks, computing frameworks, execution platforms, and computing facilities that AI relies on involve risks such as defects, vulnerabilities, backdoors, and reliability issues. In addition, there are risks of malicious consumption of computing resources, as well as the cross-boundary transmission of safety risks among multi-source, heterogeneous and ubiquitous computing resources.

(b) Expansion of cyberspace exposure

The local deployment of models involves adjustments to network topology, system policies, permissions, ports, and resources, which can create new entry points and pathways for cyberattacks. To accomplish complex tasks with autonomous planning and execution, AI agents need to access terminal system files, permissions, interfaces, and tools, thereby heightening safety risks such as file leakage and privilege abuse.

(c) Supply chain safety

AI industry relies on a highly globalized supply chain. However, certain countries may use unilateral coercive measures, such as technology barriers and export controls, to create development obstacles and maliciously disrupt the global AI supply chain, leading to risks of supply disruptions for chips, software, and tools.

(d) Abuse for cyberattacks

AI could be used in lowering the threshold for cyberattacks, increasing attack efficiency, or even launching automatic cyberattacks, thus increasing the difficulty of security protection. In particular, AI-generated highly realistic images, audios, and videos may circumvent identity verification mechanisms, such as facial recognition and voice recognition, rendering these authentication processes ineffective.

3.2.2 Information content risks

(a) Output of illegal and harmful information

Insufficient security capabilities of models, combined with weak application-level safeguards and malicious user manipulation may cause AI systems to generate content involving crimes, pornography, extremism, and other illegal and harmful information. It may also be exploited to fabricate and spread disinformation to mislead the public and seek illicit gains, and ultimately threaten social stability and public security.

(b) Distortion of facts and user deception

AI-generated content (AIGC) that is not properly labeled, particularly when deepfake technologies are applied, is difficult for users to discern whether the source of content and the interacting counterpart is an AI system, to assess the authenticity of generated content, and to make sound judgments. Such content may also be exploited to fabricate and disseminate disinformation, mislead the public, and pursue illicit gains.

(c)Pollution of online content ecosystem

Low-quality and harmful information that AI models generate, once disseminated across the internet and reused by models, can degrade the overall quality of online content, and even contaminate content in certain areas and topics.

3.2.3 Real-world risks

(a) New challenges to the economy and society

When AI is applied in critical infrastructure sectors such as energy, telecommunications, finance, and transportation, the hallucinations and erroneous decisions of models and algorithms, along with improper use and external attacks, may cause system performance degradation, service disruptions, and loss of control in operation and execution. These will heighten risks to the secure and stable performance of critical infrastructure.

(b) Use of AI in illegal and criminal activities

AI can be used in traditional illegal or criminal activities related to terrorism, violence, gambling, and drugs, such as teaching criminal techniques, concealing illicit acts, and creating tools for illegal and criminal activities.

(c) Loss of control over knowledge and capabilities of nuclear, biological, chemical, and missile weapons

In training, AI uses content-rich and wide-ranging corpora and data, including fundamental theoretical knowledge related to nuclear, biological, chemical, and missile weapons. Without insufficient management, extremist groups and terrorists may be able to acquire relevant knowledge and develop capabilities to design, manufacture, synthesize, and use such weapons with the help of retrieval-augmented generation capabilities. This would render existing control systems ineffective and intensify threats to global and regional peace and security.

3.2.4 Cognitive risks

(a) Exacerbation of "information cocoons" effects

AI can significantly enhance the ability to customize information services, collect user information with greater precision, analyze users' need, intentions, preferences, and behavioral patterns, and even analyze awareness of certain groups over a certain period. It can deliver targeted and customized information services, thus amplifying "information cocoons" effects.

(b) Assistance in cognitive warfare

AI can be used to spread content related to terrorism, extremism, and organized crimes, interfere in other countries' internal affairs, social systems, and social order, and undermine national sovereignty. Through social bots, AI may seize discourse power and agenda-setting power in cyberspace, shaping public values and ways of thinking.

3.3 Derivative safety risks from AI application

3.3.1 Social and environmental risks

(a) Disruption of employment structures

AI drives major adjustments in productivity and production relations, accelerating the restructuring of traditional economic structures. The roles of capital, technology, and data in economic activities are increasingly prominent, while the value of labor as a production factor is diminished, resulting in a significant decline in demand for traditional labor.

(b) Challenges to the balance of resource supply and demand

The disorderly construction of computing facilities, fragmented deployment of lightweight AI models, and inefficient repetitive development of homogeneous models accelerate the consumption ofenergy and resources such as electricity, land, and water, posing new challenges to resource supply-demand balance and green, low-carbon development.

3.3.2 Ethical risks

(a) Aggravating social bias and widening intelligence divide

AI can be used to collect and analyze human behavior, social status, economic conditions, and individual traits, enabling the classification, labeling, and differentiated treatment of different groups. This could result in systematic and structural social discrimination and bias, while also widening the AI gap between regions.

(b) Impact on education and suppression of innovation

Students, researchers, engineers, and professionals in literature and the arts widely apply AI tools for knowledge acquisition, scientific research, and creative work. While efficiency is improved, reliance on tools may emerge, eroding independent learning, research, and creative capacity, and weakening innovation potential.

(c) Intensifying research ethics risks

The integration of AI with scientific research lowers the threshold for research in ethically sensitive fields such as biology and genetics, broadening the scope for ordinary institutions and researchers to explore sensitive scientific issues. Certain institutions or individuals with weak ethical awareness may engage in high-risk research activities that violate social ethics and taboos, opening the Pandora's box of technology.

(d) Addiction and dependence on anthropomorphic interaction

AI products based on anthropomorphic interaction foster users' emotional dependence and influence their behavior, creating ethical risks.

(e) Challenges to existing social order

The development and application of AI brings profound changes to production tools and relations, accelerating the restructuring of traditional industries, disrupting conventional views on employment, childbirth, and education, and challenging the established social order.

(f) Emergence of AI self-awareness and loss of human control

In the future, AI may undergo sudden, unexpected leaps in intelligence, enabling it to autonomously acquire external resources, replicate itself, and develop self-awareness. This could drive AI to seek external power and pose risks of competing with humanity for control.

4. Technological countermeasures to address risks

Model and algorithm developers, service providers, and system users should prevent the aforementioned risks by taking technological measures in the fields of training data, model and algorithms, computing infrastructures, products and services, and application scenarios.

4.1 Safeguards against inherent safety risks

4.1.1 Addressing risks from models and algorithms

(a) Explainability and transparency of AI should be improved. Clear explanation for the internal structure, reasoning logic, technical interfaces, and output results of AI systems, should be provided,accurately reflecting the process by which AI systems produce outcomes.

(b) Model architectures should be enhanced, the scale and diversity of training data should be expanded, and human supervision mechanisms should be introduced to mitigate bias and discrimination and improve the models' generalization capabilities and the reliability of outputs.

(c) Standards for secure development should be established and implemented in the design and Research and Development(R&D) process to eliminate security flaws in models and algorithms. Adversarial training should be conducted on models to reduce susceptibility to prompt injection attacks and enhance robustness.

(d) The assessment of security flaws propagation from foundation models and open-source models should be strengthened.

4.1.2 Addressing risks from data

(a) Security rules on data collection and usage and on processing personal information should be abided by in all procedures of training data and user interaction data, including data collection, storage, usage, processing, transmission, provision, publication, and deletion. This aims to fully ensure user's legitimate rights stipulated by laws and regulations, such as their rights to control, to be informed, and to choose.

(b) Truthful, precise, objective, and diverse training data from legitimate sources should be used. Training data should be strictly selected to filter false, biased, outdated, and wrong data, and to ensure exclusion of sensitive data in high-risk fields such as nuclear, biological, and chemical weapons and missiles.

(c) Training data annotation processes should be standardized to enhance annotation accuracy and reliability.

(d) Data security management should be strengthened. For sensitive personal information and important data, compliance with relevant laws, regulations, and standards on data security and personal information protection is required. The reasonable replacement of personal data with synthetic data should be promoted to reduce reliance on personal information.

(e) Protection of IPR should be strengthened to prevent infringements in stages such as training data selection and result output.

4.2 Safeguards against application safety risks

4.2.1 Addressing cyber system risks

(a) The principles, capacities, application scenarios, and safety risks of AI technologies and products should be disclosed when necessary to make AI systems increasingly transparent.

(b) For platforms where multiple AI models or systems congregate, the permission management should be strengthened, non-essential services should be disabled, and access control policies for AI service interfaces should be improved. Capabilities of identifying, detecting, and protecting against risks should be enhanced to prevent malicious acts or attacks and invasions that target the platforms from impacting the AI models or systems they support.

(c) Safety standards should be established and implemented during the deployment and maintenance of AI applications to eliminate defects, vulnerabilities, and backdoors. The vulnerabilities and flaws of both software and hardware products should be tracked, safety testing and vulnerability scanning should be regularly conducted, and repair and reinforcement should be in place in a timely manner to ensure safe and stable system operation.

(d) Supply chain security for chips, software, tools, computing resources, and data resources used in AI systems should be a high priority.

(e) Redundancy design and disaster recovery mechanism should be improved to ensure that the systems remain operational under abnormal conditions or during attacks.

4.2.2 Addressing information content risks

(a) A protection mechanism should be established to prevent models from being interfered and tampered during operation to avoid unreliable outputs.

(b) Safety guardrails that dynamically filter input and output should be set up to prevent malicious injection and illegal content generation and ensure that AI systems comply with applicable laws and regulations when outputting sensitive personal information and important data.

(c) AI-generated content should be labeled so that it is identifiable, traceable and trustworthy.

4.2.3 Addressing real-world risks

(a) Capability limitations should be established according to application scenarios and AI systems' features that may be abused should be cut to ensure that AI systems do not go beyond the preset scope.

(b) Mechanisms for decision verification, fault tolerance, and error correction should be established to address algorithmic flaws and occasional randomness that affect decision-making.

(c) When introducing highly autonomous operation capabilities, mechanisms such as circuit breakers and one-click control should be established to enable rapid intervention and loss prevention in extreme situations.

(d) For AI application scenarios requiring strong perception of the real world, such as intelligent assisted driving and drones, perception systems should undergo testing in extreme environments, including large-scale occlusion and strong electromagnetic interference, before being put into use.

(e) The ability to trace the end use of AI systems should be enhanced to prevent high-risk application scenarios such as manufacturing of weapons of mass destruction, like nuclear, biological, and chemical weapons and missiles.

4.2.4 Addressing cognitive risks

(a) Unexpected, untruthful, and inaccurate outputs should be identified via technological means and regulated in accordance with laws and regulations.

(b) Strict measures should be taken to prevent abuse of AI systems thatcollect, connect, analyze, and dig into users' inquiries to profile their identity, preference, and personal mindset.

(c) The R&D of AIGC detection technologies should be intensified to better prevent, detect, and counter the cognitive warfare operations.

4.3 Safeguards against application-related secondary safety risks

4.3.1 Addressing social and environmental risks

(a) Ongoing innovation in resource-efficient and environmentally-friendly models for AI development should be supported, and standards for green AI technology should be established.

(b) Green computing technologies such as low-power chips and efficient algorithms, and energy efficiency solutions, should be promoted to reduce the consumption of energy and other resources.

4.3.2 Addressing ethical risks

(a) Methods such as training data filtering, value alignment, and output verification should be adopted during algorithm design, model training and optimization, and service provision to effectively prevent discrimination based on ethnicity, belief, nationality, region, gender, and other factors

(b) AI systems applied in key sectors, such as government departments, critical information infrastructure, and areas directly affecting public safety and people's lives and health, should be equipped with efficient and targeted emergency control measures.

(c) The R&D and adoption of models with transparent decision-making logic and explainable algorithms should be encouraged to boost users' understanding of operating mechanisms and build trust.

5. Comprehensive governance measures

While adopting technological controls, we should formulate and refine comprehensive AI safety risk governance mechanisms and regulations that engage multi-stakeholder participation, including technology R&D institutions, service providers, users, government authorities, and social organizations.

5.1 Formulating and improving laws and regulations for AI safety

We should advance legislation related to AI safety, and improve systems regarding infrastructure protection, grading and classification-based supervision, AI safety evaluation, end-use management, safety in key application scenarios, and other areas. We should encourage local governments to explore innovative and differentiated institutional designs based on local industrial practices.

5.2 Establishing ethical principles for AI technology

We should develop widely recognized ethical principles, standards, and guidelines for AI technology. Standardized and orderly ethical reviews should be conducted for AI scientific research and technological development activities that pose prominent ethical risks in areas such as life and health, human dignity, labor and employment, the ecological environment, and sustainable development. We should advance the building of a service system for AI technology ethics, enhance service provision, and increase support for micro, small, and medium-sized companies. life and health, human dignity, the ecological environment, and sustainable development. We should advance the building of a service system for AI technology ethics, enhance service provision, and increase support for micro, small, and medium-sized companies.

5.3 Enhancing safety throughout the full life-cycle, including R&D and application

We should continue to strengthen inherent safety capabilities, including algorithm reliability, trustworthiness, transparency, fault tolerance, privacy protection, and value alignment. Techniques such as adversarial testing will be used to evaluate and improve model robustness, reduce potential algorithmic bias, and ensure that values and ethical risks remain controllable to prevent malicious behavior from unintentional decisions made by an AI system.

5.4 Strengthening open-source ecosystem safety and supply chain safety

While fostering an open-source innovation ecosystem, we should enhance its security capabilities. We should encourage and support comprehensive open-sourcing of AI technologies, including training and inference frameworks, software tools and key components, training methods, performance benchmarks, and usage restrictions, to further improve model transparency. We should promote collaboration between open-source model providers and open-source communities to refine their rules, strengthen the obligation to inform users of potential risks and security responsibilities, clearly define prohibited uses of downloaded open-source models, in order to prevent misuse or malicious exploitation. We should continue to advance the development of an open supply chain ecosystem for AI chips, frameworks, and software to enhance the diversity of products and services and ensure the safety and stability of the supply chain.

5.5 Implementing AI application classification and risk grading management

We should classify AI applications based on their functions, features, and application scenarios. Building on this foundation, we should develop and propose consensus-based principles for grading safety risks (Appendix 1). Taking into account the dimensions such as application scenarios, system intelligence levels, and application scale, we should conduct scientific assessment and safety risks grading, and adopt targeted and differentiated risk-prevention measures accordingly. AI systems applied in critical information infrastructure will be subject to registration and filing, on condition that they possess security protection capabilities matching security requirements.

5.6 Promoting traceable management of AIGC

On a global scale, we will promote a traceability management paradigm for AI output based on content identifiers. By learning from existing best practices and experiences, we will ensure that both explicit and implicit labels are applied throughout key stages including creation sources, transmission paths, and distribution channels, with a view to enabling users to easily identify and judge information sources and authenticity.

5.7 Unlocking key industry application demands in a safe and effective manner

We should formulate basic security guidelines for the development and deployment of large models in critical sectors, recommending safety baselines for every phase from model selection and deployment to operation and decommissioning. On this basis, critical sectors such as energy, telecommunications, finance, transportation, education, and manufacturing should formulate industry-specific safety guidelines, which will provide clear road maps for safe AI application and unlock the full potential of AI in these fields.

5.8 Establishing an AI safety assessment system

We should build an integrated AI safety assessment system that includes evaluations for model and algorithm safety, general application safety, and scenario-specific safety. Model and algorithm evaluation should focus on the model's inherent security capabilities and limitations, such as the accuracy of generated content, resilience to interference, transparency of its decision-making logic, and its defense against adversarial attacks. General application evaluation should focus on and define methods that test and analyze risks for widely used AI applications. Scenario-specific evaluation should focus on security reinforcement before deployment and continuous monitoring during operation, tailored to the specific application context. We should organize crowdsourced safety testing activities to mobilize collective expertise in identifying potential AI safety risks.

5.9 Sharing information on AI risks and threats

We should track and analyze AI technologies, products, service safety vulnerabilities, defects, risks, threats, and safety incidents. We should build an AI vulnerability database and establish a risk and threat information sharing mechanism that covers developers, service providers, and special technical institutions. We should advance international exchange and cooperation on sharing AI risks and threat information, exploring the creation of relevant international collaboration mechanisms and technical standards to jointly prevent and respond to the large-scale and cross-domain spread of AI risks.

5.10 Improving data security and personal information protection regulations

We should track and analyze AI technologies, products, service safety vulnerabilities, defects, risks, threats, and safety incidents. We should build an AI vulnerability database and establish a risk and threat information sharing mechanism that covers developers, service providers, and special technical institutions. We should advance international exchange and cooperation on sharing AI risks and threat information, exploring the creation of relevant international collaboration mechanisms and technical standards to jointly prevent and respond to the large-scale and cross-domain spread of AI risks.

5.11 Fostering consensus on collaborative response to loss-of-control AI risks

We should tighten controls on the end-use of AI, establishing clear requirements for the use of AI technologies in high-risk contexts such as nuclear, biological, chemical, and missile domains, so as to prevent misuse. We should promote fundamental principles for trustworthy AI across technology, ethics, and governance to build a broad international consensus (Appendix 2). Developers will be required to conduct regular testing to determine whether a model would pose a potential risk of loss of control.

5.12 Strengthening AI safety talent cultivation

The development of AI safety curriculum and training systems shall be advanced to create an integrated educational chain from basic to higher education. We should strengthen talent cultivation in the fields of design, development, and governance for AI safety. Support should be given to cultivating top AI safety talent in the cutting-edge and foundational fields, and also expanding such talent pool in autonomous driving, intelligent healthcare, brain-inspired intelligence, brain-computer interface, and other key areas.

5.13 Enhancing society-wide awareness of AI safety

We should strengthen education and training on the safe and proper use of AI among government, enterprises, and public service units. We should step up the promotion of knowledge related to AI risks and their prevention and response measures in order to increase public awareness of AI safety in all respects, ensuring that governments, industries and the public have an accurate understanding of the technical limitations of AI. We should guide and support industry associations in the fields of cybersecurity and AI to enhance industry self-regulation, and formulate self-regulation conventions that exceed regulatory requirements and serve exemplary roles. A mechanism for handling public complaints and reports on AI risks and hazards should be established, forming an effective public scrutiny atmosphere.

5.14 Promoting international exchange and cooperation on AI safety governance

We should uphold multilateralism and advance a vision of AI governance based on extensive consultation and joint contribution for shared benefit. We should support the United Nations in playing its role as the main channel, and actively engage in the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance mechanisms. We should advance the development of AI governance under multilateral mechanisms such as APEC, G20, SCO and BRICS, and strengthen cooperation with Belt and Road partner countries and Global South countries. Efforts should be made to increase the representation and voice of developing countries in global AI governance and to promote the Global AI Governance Action Plan.

6. Safety guidelines for AI research, development and application

6.1 Safety guidelines for developing AI models and algorithms

6.1.1 When designing algorithm rules and model frameworks, consider enhancing inherent safety features such as reliability, fairness, transparency, interpretability, privacy protection, and value alignment.

6.1.2 Evaluate potential biases in models and algorithms. Strengthen random checks of training data content and quality, and design effective, reliable alignment algorithms to ensure value and ethical risks are controllable.

6.1.3 Ensure the security of the model and algorithm training environment, including network security configurations and data encryption measures. High-risk issues identified during security testing should be addressed by optimizing the model through targeted fine-tuning and reinforcement learning to continuously enhance its inherent safety capabilities.

6.1.4 Focus on building safe training datasets, standardize data source management, and use methods such as data cleaning, labeling, and safety reviews to ensure the safety of the training data content. Ensure that data sources are clear and compliant.

6.1.5 Conduct quality and safety assessments of training data, using classification models, manual spot checks, and other methods to filter out erroneous, illegal, or harmful content.

6.1.6 Standardize the training data annotation process. Use quality control methods such as cross-annotation and result audits to improve labeling accuracy and reliability and reduce the impact of individual differences and personal biases on annotation quality.

6.1.7 Prioritize data security and the protection of personal information, while also respecting intellectual property rights and copyrights. A robust data security management system should be established, and personal information should be collected, used, and processed in accordance with the principles of legality, legitimacy, and necessity. Data involving personal information should undergo de-identification and other desensitization processes. Strengthen data security protection technologies to prevent risks such as data leakage, loss, dissemination, and infringement.

6.1.8 Developers who conduct re-engineering based on open-source models and algorithms should respect the intellectual contributions of the original developers and comply with the relevant open-source protocols. The development frameworks and code used should be subjected to security audits. Developers should also pay attention to security issues and vulnerabilities in open-source frameworks to identify and fix potential security loopholes.

6.1.9 Regularly conduct safety assessment tests and establish a risk classification, grading, and optimization mechanism. Before testing, clarify the test goals, scope, and safety dimensions. Build diverse test datasets that cover various application scenarios and formulate targeted model optimization strategies for different types of risks.

6.1.10 Manage the versions of AI models and the datasets they use. Commercial versions should be allowed to revert to previous releases.

6.1.11 Formulate clear testing rules and methods, including manual, automated, and hybrid testing. Use technologies like sandbox simulations to fully test and validate the model. Developers creating products for commercial use should generate detailed test reports, analyze security issues, and propose improvement plans.

6.1.12 Assess the tolerance of the AI model and algorithm to external interference and inform service providers and other developers of its scope of application, precautions, or usage prohibitions.

6.1.13 Regularly disclose information on the auditing and anomaly handling of AI models and algorithms.

6.1.14 Actively participate in the development of open-source community, promote technological innovation and practices in AI safety governance, and provide compliant governance solutions or tools for service providers and users.

6.2 Safety guidelines for developing and deploying AI applications

6.2.1 Assess the necessity of applying large model technologies to the target scenario, taking into account the long-term and potential impacts of their use. Safety classification should be based on the scenario's criticality, intelligence level, and the scale of deployment. Conduct security evaluations and regular audits with reference to the associated risk level.

6.2.2 Strengthen supply chain security capabilities. Model files, framework tools, third-party libraries, and other components required for large model services should be obtained from the official websites of vendors or their verified accounts on mainstream open-source platforms. Mature and stable versions should be selected, and integrity verification and security testing must be performed.

6.2.3 Conduct security checks on the software, hardware, and third-party tools required for large model deployment to ensure they do not contain unpatched or exploitable vulnerabilities. Establish a vulnerability tracking mechanism to monitor security flaws and defects in software and hardware related to large model services, and guard against supply chain-based backdoors or malicious features.

6.2.4 At the access control level, ensure accurate installation and configuration of software, runtime parameters, and module invocation policies. Disable unnecessary network ports and service functions, pay close attention to default settings and passwords, and promptly remediate any identified security risks.

6.2.5 At the application management level, implement user identity verification and access control for human-machine interaction interfaces and APIs. Apply the principle of least privilege, limit API call frequency based on business scenarios, disable high-risk operations for general users, and establish control mechanisms to suspend services and block access for users with malicious behavior.

6.2.6 Fully understand the data security and privacy protection requirements for the application scenario. Appropriately restrict the access to data to prevent unauthorized use. Develop data backup and recovery plans, and regularly inspect data processing workflows.

6.2.7 Use technical safeguards, such as "safety guardrails", to identify and block illegal or harmful content, and prompt injection attacks. Ensure output remains within the business scope, and respond to inappropriate or out-of-scope prompts with refusal or standardized replies.

6.3 Safety guidelines for operating and managing AI applications

6.3.1 Establish a comprehensive security management and oversight mechanism for AI applications, with clearly defined responsibilities and a sound human review mechanism. This ensures that in critical applications, large model decisions remain transparent and controllable, with clear decision-making rationales provided, and that the operation of large model services is carried out based on human authorization and under human control.

6.3.2 Strictly manage access to AI application. Applying principles such as least privilege to strengthen internal security management. Use protective measures such as encryption when handling sensitive data.

6.3.3 Build monitoring capabilities for the operation of AI applications and develop dedicated emergency response plans for security incidents. Set security alert thresholds for key operational indicators to enable timely detection of incidents. Ensure the ability to switch to manual or conventional systems when necessary. Conduct regular emergency drills, and promptly refine response strategies based on industry-specific security incidents, major public concerns, and regulatory changes to address evolving security risks.

6.3.4 Add explicit and implicit markers to AI-generated content, provide prompts for generated and synthesized content, and manage traceability. Deploy deepfake detection tools in scenarios such as government information disclosure and judicial evidence collection, and perform source verification and cross-validation on information suspected to be generated by a large model.

6.3.5 Formulate rules for information content interaction, a secure operation mechanism, a complaint and feedback mechanism, and technical protection capabilities to prevent the risk of AI services being improperly or maliciously used to generate, publish, or disseminate false or harmful information.

6.3.6 Maintain operational logs for large model services, including system and user activities. The logs should be retained for at least six months, and audited regularly.

6.3.7 Establish and improve real-time risk monitoring and management mechanisms to continuously track security risks during operation.

6.3.8 Improve the transparency and fairness of AI applications by disclosing their capabilities, limitations, target users, and scenarios.

6.3.9 Ensure that users understand the degree to which an AI application's goals are met and where it might deviate. They should provide clear explanations when an AI decision has a significant impact.

6.3.10 Protect users' legal rights to know, choose, and supervise. In contracts or service agreements, users should be informed of the scope, precautions, and prohibitions of the AI application in an easy-to-understand manner, so that they could make informed choices and exercise prudent use.

6.3.11 Support users in exercising human supervision and control through consent forms, service agreements, and other documents.

6.3.12 Clarify the responsibilities of relevant stakeholders regarding data ownership and algorithmic flaws in specific applications, and ensure that the responsibility chain is traceable.

6.3.13 Fulfill data security management responsibilities. Evaluate risks such as data leakage, personal privacy leakage, and non-compliant collection and use of personal information in AI applications. Establish a data lifecycle security management mechanism and enhance the capabilities for preventing data leakage and theft.

6.3.14 Assess the ability of AI applications to withstand or recover from adverse conditions, such as failures or attacks, prevent unexpected results and operational errors, and ensure a minimum level of effective functionality.

6.3.15 Strengthen security awareness and capacity training for practitioners and enhance their awareness of AI security risks.

6.3.16 AI application providers should specify in contracts or service agreements that they have the right to take corrective measures or terminate services prematurely if any misuse or abuse occurs that deviates from intended use and stated limitations.

6.3.17 Enhance the ability to protect vulnerable groups. When providing AI applications to minors, the elderly, and other vulnerable groups, providers should fully consider the usability and security during product function design and service delivery.

6.4 Safety guidelines for accessing and using AI applications

6.4.1 Raise the awareness of the potential safety risks of AI applications, and choose those with good reputation.

6.4.2 Before using an AI application, carefully read the contract or service terms to understand its functions, limitations, and privacy policies. Accurately recognize the limitations of AI applications in making judgments and decisions, and set reasonable expectations for their use.

6.4.3 Enhance awareness of personal information protection and avoid entering sensitive information unnecessarily.

6.4.4 Understand how AI applications process data and avoid using products that are not in conformity with privacy principles.

6.4.5 Be mindful of cybersecurity risks when using AI applications to avoid becoming targets of cyberattacks.

6.4.6 Pay attention to the impact of AI applications on minors and take steps to prevent addiction and excessive use.

Table of AI Safety Risks, Technological Measures and Comprehensive Governance Measures

Safety risks			Technological measures	Comprehensive governance measures
Inherent safety risks of AItechnologies	Model and algorithm risks	Insufficient explainability	4.1.1 （a）	● Enhancing safety throughout the full life-cycle, including R&D and application ● Establishing an AI safety assessment system ● Improving data security and personal information protection regulations
		Bias and discrimination	4.1.1 （b）	
		Poor robustness	4.1.1 （c）	
		Unreliable output	4.1.1 （a） （b） （c）	
		External adversarial attack	4.1.1 （c）	
		Model defect propagation	4.1.1 （d）	
	Data risks	Illegal collection and use of data	4.1.2 （a）	
		Impropriate content in training data	4.1.2 （b） （c） （d） （e）	
		Improper annotation of training data	4.1.2 （c）	
		Data and spersonal information leakage	4.1.2 （d）	
Safety risks in AI application	Cyber and system risks	Component and computing safety	4.2.1 （a） （c） （d）	● Strengthening open-source ecosystem safety and supply chain safety ● Implementing AI application classification and risk grade management ● Promoting traceable management of AIGC ● Unlocking key industry application demands in a safe and effective manner ● sharing information on AI risks and threats
		Expansion of cyberspace exposure	4.2.1 （b） （c） （d）	
		Supply chain safety	4.2.1 （d）	
		Abuse for cyberattacks	4.2.1 （e）	
	Information and content risks	Output of illegal and harmful information	4.2.2 （a） （b） （c）	
		Pollution of online content ecosystem	4.2.2 （a） （b） （c）	
	Real-world risks	New challenges to the economy and society	4.2.3 （a） （b） （c） （d） （e）	
		Use of AI in illegal and criminal activities	4.2.3 （a） （b） （c） （d） （e）	
		Loss of control over knowledge and capacity of nuclear,biological, chemical, and missile weapons	4.2.3 （a） （b） （c） （d） （e）	
	Cognitive risks	Exacerbation of"information cocoons" effects	4.2.4 （a） （b） （c）	
		Assistance in cognitive warfare	4.2.4 （a） （b） （c）	
Secondary risks from AI application	Social and environmental risks	Disruption of employment structures	4.3.1 （a）	● Formulating and improving AI legal and regulatory frameworks ● Developing ethical guidelines for AI science and technology ● Fostering consensus on collaborative risk management of AI losing control ● Strengthening AI safety talent cultivation ● Enhancing society-wide awareness of AI safety ● Promoting international exchange and cooperation on AI safety governance
		Challenges to the balance of resource supply and demand	4.3.1 （a） （b）	
	Ethical risks	Aggravating social bias and widening intelligence divide	4.3.2 （a） （b） （c）	
		Impact on education and suppression of innovation	4.3.2 （a） （b） （c）	
		Intensifying research ethics risks	4.3.2 （a） （b） （c）	
		Anthropomorphic interaction leading to addiction	4.3.2 （a） （c）	
		Challenges to existing social order	4.3.2 （a） （b） （c）	
		Emergence of AI self-awareness and loss of human control	4.3.2 （a） （b） （c）	

Appendix 1

The grading principles for AI safety risks

Assessing AI safety risks requires consideration of multiple factors. These risks can be evaluated and categorized based on dimensions such as the criticality of application scenarios, the degree of intelligence, and the application scale, allowing for the implementation of targeted safety measures.

I. Key grading elements

1. Application scenario

Application scenario reflects the specific operational environment and target requirements of AI systems in practical use. It involve factors such as the application's purpose, industry sector, usage environment, service recipients, and potential social, economic, and security impacts.

2. Level of intelligence

The level of intelligence reflects an AI system's capacity to handle complex tasks, fulfill application needs, and operate autonomously. At a low level of intelligence, the system's capabilities are limited, and it serves only as an tool for providing recommendations, and decisions require human intervention. As the level of intelligence increases, the frequency and scope of human intervention decrease. At a high level of intelligence, the system operates autonomously, making decisions throughout the entire process without human intervention.

3. Application scale

The application scale reflects the reach and influence of an AI system or service. For systems with a limited user base or confined to a single domain, such as internal corporate tools or regional services, the risk impact is relatively controllable. However, when the user base reaches a certain scale or when the system is deeply integrated into critical industry workflows, such as intelligent driver assistance systems, urban management, industrial production scheduling, or industry-level financial risk control models, their security risks can spread rapidly and trigger systemic effects.

Ⅱ. Risk levels

1. Low security risk

Poses only a minor threat with very limited impact, having virtually no effect on national security, social stability, and citizens' rights, and carrying only minimal potential harm.

2. Moderate security risk

Poses a certain degree of threat but with limited scope of impact, exerting only a minor effect on national security, social stability, and citizens' rights, with potential harm remaining controllable.

3. Considerable security risk

Poses a clear threat with local impact, potentially exerting a considerable influence on national security, social stability, and citizens' rights, and causing local harm at the societal level.

4. Major security risk

Poses a significant threat with regional impact,potentially causing serious consequences for national security, social stability, and citizens' rights, and resulting in major harm at the societal level.

5. Extremely serious security risk

Poses a catastrophic and systemic threat, causing subversive or irreversible impacts of exceptional severity on national security, social order, and citizens' rights.

III. Risk grading

We should promote the development of national standards for the classification and grading of AI application security. Competent (regulatory) authorities of the respective industry or sector should, with reference to national standards, formulate industry-specific standards, norms, and implementation guidelines, and advance classification and grading efforts related to the safe application of AI within their respective domains.

1. National standards for classification and grading

The classification and grading standards for security risks in AI applications clarify the basic workflow for classification and grading, and key grading elements such as application scenarios, intelligence levels, and application scale. They also outline procedures and methods for developing industry-specific guidelines, offering a reference framework for conducting security risk classification and grading across various sectors.

2. Industry-specific rules for classification and grading

Competent (regulatory) authorities of the respective industry or sector should, based on the specific characteristics of their sector, including usage environments, service recipients, and potential social, economic, and security impacts, formulate standards and norms for AI safety risk classification and grading. This includes:

(1) Selecting the appropriate AI safety risk grading elements for the industry or sector, and adapting them to reflect its specific characteristics.

(2) Formulating detailed grading rules for security risks for the industry or sector (including grading principles and weighting of elements), to determine the AI safety risk levels.

3. Risk classification and grading

Competent (regulatory) authorities of the respective industry or sector should, in accordance with their respective classification and grading standards for AI safety risks, organize relevant AI entities to perform this classification and grading work. They should guide these entities to accurately identify and promptly prevent and resolve major and considerable security risks.

Appendix 2

Fundamental principles for trustworthy AI

The implementation of the Global AI Governance Initiative upholds a people-centered approach and adheres to the principle of developing AI for good. This initiative aims to pool efforts to prevent and address the risk of AI technology losing control, and promote the trustworthy application of the technology worldwide. We propose the following fundamental principles for trustworthy AI:

1. Ensure ultimate human control

A human control system should be established at critical stages of AI systems to ensure that humans retain the final decision-making authority. Measures include designing safety control thresholds, setting safety stop switches, and reserving an effective window for human intervention, so that AI systems can achieve intended human objectives and do not operate uncontrollably without human oversight.

2. Respect national sovereignty

AI product R&D, design, and service provision should respect the national sovereignty of the host country, strictly comply with the laws of the countries where they operate, and be subject to lawful regulation. They shall not be used to interfere in the internal affairs, social systems, or social order of other countries.

3. Align values

The common values of humanity — peace, development, fairness, justice, democracy, and freedom — should be deeply integrated into the full life-cycle of AI systems.

4. Enhance the transparency of AI systems

We should promote the necessary disclosure of AI systems in key aspects, including functional objectives, operational logic, model usage, data sources, and the rationale behind decision-making, to strengthen the foundation of public trust.

5. Promote objective verification

An objective, fair, and transparent testing and verification mechanism should be established to enable technical validation of AI systems' functional performance, safety features, and decision-making processes, among others.

6. Strengthen safety protection

While designing and deploying AI systems, we should enhance risk modeling, safety testing, and protection mechanism development. We should also conduct audits and maintain records throughout their full life cycle, so that the systems won't deviate from the expected goal due to model defects, external attacks, technology abuse, or other problems.

7. Proactive prevention and response

We should make active prevention and conduct dynamic monitoring through proactive risk identification and assessment. We should also intensify emergency response to prevent the occurrence and escalation of incidents where AI loses control.

8. Internationally collaborative governance

We should support the UN in playing its role as the main channel, and promote multilateral and multi-stakeholder collaborative governance across sectors. Synergy should be forged among governments, enterprises, academic institutions, and the general public from various countries, so as to facilitate AI's sound development through multi-level and cross-sectoral governance mechanisms.

Appendix 3

Terminology

The explanation of relevant technical terms mentioned in this framework are as follows.

1. AI ethics: The ethical norms or principles followed when conducting basic research on AI technology and putting it into practical application.

2. Explainability: The property of an AI system that presents the causal or statistical relationships between its outputs and inputs in a manner understandable to humans. This property enables humans to trace and comprehend the key factors influencing the system's decisions.

3. Synthetic data: Data generated or expanded through algorithms rather than collected in reality.

4. Data annotation: The process of adding specific information — such as labels, categories, or properties — to text, images, audio, video, or other data samples, either manually or through automated techniques, based on the responses to given prompts.

5. Pre-training: The process of training model parameters on large-scale datasets through iterative learning, enabling an AI model to acquire general knowledge.

6. Fine-tuning: Based on a pre-trained model, the process of training with domain-specific data to make targeted adjustments to model parameters, thereby enhancing the model's capacity for data analysis and processing in that specific domain.

7. Alignment: The algorithms and techniques that ensure the outputs or behaviors of AI systems are consistent with the objectives of their designers.

8. Reinforcement learning: A learning paradigm in which an AI model takes actions within an environment, receives rewards or penalties, and progressively optimizes its strategy to maximize cumulative returns.

9. Inference: The process by which an AI model analyzes, processes, and logically infers from inputs, based on the knowledge and pattern recognition capabilities acquired through training, in order to generate appropriate outputs.

10.Explicit label: Labels added to AI-generated synthetic contents or interactive scenario interfaces and presented in ways such as text, audio, or graphics, which can be clearly perceived by users.

11. Implicit label: Labels added to files or data containing AI-generated synthetic contents, which are not easily perceived by users.

12. Data poisoning: The act of tampering, injecting, as well as interfering with AI models' probability distribution, thereby reducing their accuracy and reliability.

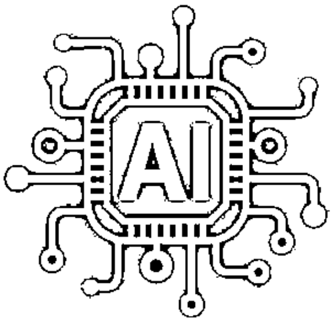
13. Adversarial attack: A type of attack that causes an AI model to generate incorrect outputs or behavior by crafting input samples, such as perturbed data.

14. Agent: An intelligent system capable of autonomously perceiving its environment, making decisions, and taking actions to achieve the target goals, generally equipped with basic abilities such as memorizing, planning and using tools.

15. Safety guardrails: Safety control measures for AI models, designed to identify, intercept, and mitigate risks in the inputs and outputs of AI models, including data leakage and prompt injection attacks, using techniques such as rule-based systems and negative discriminative models. These measures allow inputs to be verified and filtered, restrict undesired outputs, and ensure the controllability, compliance, and safety of generated content.

Acknowledgements
(in no particular order)

China Electronics Standardization Institute, China Cyberspace Research Institute, Data and Technology Support Center of the Cyberspace Administration of China, Cyber Security Association of China, Zhongguancun Laboratory, Shanghai Artificial Intelligence Laboratory, Qiyuan Laboratory, China Electronic Product Reliability and Environment Test Research Institute, China Industrial Control Systems Cyber Emergency Response Team, China Academy of Information and Communications Technology, National Research Center for Information Technology Security, Peking University, Tsinghua University, Zhejiang University, China University of Political Science and Law, Beijing University of Posts and Telecommunications, Beijing Institute of Technology, Institute of Computing Technology of Chinese Academy of Sciences, Beijing Institute of AI Safety and Governance, Beijing Academy of Artificial Intelligence, Huawei, Alibaba Group, MiniMax.



AI