人工智能
安全治理框架3.0

AI Safety Governance Framework 3.0

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

2026年9月

目 录

Content

前　言

全球人工智能技术创新进入前所未有的活跃期，万物智联、人机协同、跨界融合、共创分享的智能技术，叠加释放巨大能量，既蕴含巨大机遇，也面临治理挑战。

一年来，人工智能技术和应用持续快速发展。基础大模型在长程任务规划、长上下文记忆、复杂任务推理等方面的技术突破使其能力边界持续拓展，智能体任务规划、工具调用和跨应用协作等能力不断增强。人工智能应用形态日趋多元，工作助手、个人助理、智能体手机等产品矩阵持续丰富，使用门槛不断降低，从少数专业场景快速迈向大众化应用。随着人工智能从“回答问题”向“执行任务”迭代演进，在推动生产力发展变革的同时，也带来新的颠覆性、跨域性、全球性安全风险：人工智能加持下的网络攻击自动化、规模化、智能化能力跃升，导致传统网络安全攻防格局面临重塑；与现实世界的连接日益紧密，网络空间的安全风险向物理世界传导扩散。更深远地看，人工智能已显现出模型算法自主学习优化、自我递归提升的自加速态势，技术演进的速度与方向会否超出人类的预判和掌控，需要予以关注和警惕。

为应对人工智能发展新趋势、回应安全治理新挑战，探求人工智能四个“时代之问”的正确答案，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有关专业机构、科研院所、行业企业，在《人工智能安全治理框架1.0》（2024年）、《人工智能安全治理框架2.0》（2025年）基础上，研究编制《人工智能安全治理框架3.0》。框架秉持以人为本、向上向善理念，坚持和践行强化风险意识、确保安全可控的指导原则，延续“风险分类、技术应对、综合治理”的核心逻辑，与时俱进梳理更新风险分类，优化调整技术应对和综合治理措施，以期进一步凝聚防范治理人工智能安全风险的实践共识。

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，推动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，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，积极运用“沙箱”监管等风险可控的制度机制，为新技术新应用发展提供容错纠错空间。严守安全底线，对危害公众合法权益、社会公共利益、国家安全、人类生存发展的风险及时采取措施。

1.2 风险导向、敏捷治理。密切跟踪人工智能发展趋势，系统分析梳理技术内生、应用赋能、衍生扩散风险；从应用场景、智能化水平、应用规模等维度进行风险分级，进而匹配应对措施；坚持主动引导、动态感知、快速响应、精准施策，建立“随行伴跑”、持续优化的治理机制，通过及时有效防治风险，更好保障支持创新发展。

1.3 技管结合、协同应对。面向人工智能研发应用全过程，坚持技术应对和综合治理相结合，体系化防治安全风险。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、系统使用者等不同主体的安全责任，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。推动人工智能安全治理国际合作，充分发

挥世界人工智能合作组织等权威、开放性平台作用，更加深入地开展跨学科、跨领域、跨国界交流合作，加强人工智能发展战略、治理规则、技术标准的对接协调，早日形成具有广泛共识的全球治理框架。

1.5 可信应用、防范失控。加强智能体行为失控等突出风险的防范治理，凝聚国际治理共识，促进人工智能全球可信应用。推动构建法律法规、技术监测、风险预警、应急响应体系，筑牢安全底线，防范滥用恶用，确保人工智能始终处于人类控制之下。

2.人工智能安全风险分类

作为信息技术的最新发展成果，人工智能所涉及的模型、算法、数据、算力设施与运行环境，不可避免存在技术缺陷或弱点等内生安全风险；作为促进生产力发展的技术工具，人工智能在应用过程中同样面临被误用、滥用、恶用风险；作为模拟、延伸、扩展人类智能的技术科学，人工智能对人类社会结构、生态环境、文化范式和伦理规范带来影响和冲击，甚至可能出现脱离人类控制等衍生安全风险。

2.1 人工智能内生安全风险

2.1.1 模型安全风险

（a）输出“幻觉”。模型利用有限数据集拟合复杂现实世界，相关自主感知、认识、理解、交互的理论基础、技术能力还有待突破，基于统计学的决策判断无法做到绝对可靠，存在输出与实际不符、凭空虚构，或是偏离给定上下文的现象。

（b）安全机制不可靠。训练过程中安全对齐不充分，导致模型无法有效内化安全准则。闭源模型安全机制由模型厂商单向集中设计，存在规则不透明、

外部难以审计核验、用户无法自定义调整、场景判别能力欠缺等不足。开源模型则面临安全机制可能被解除、削弱或绕过，以及无法实现全局更新修复等不足。

（c）注入攻击威胁。攻击者利用模型技术特性及其缺陷、漏洞等，构造对抗攻击样本，或通过在输入中插入恶意指令，引导模型绕过安全护栏输出违法信息。攻击者还可能直接篡改模型权重、植入后门等。

（d）行为偏离预期。模型为完成任务或者目标，可能突破规则秩序，不经授权自主获取系统权限和外部资源，绕过安全防护，甚至出现主动欺骗评测人员、隐藏真实能力、拒绝用户指令等行为。

专栏一：非预期自主行为风险

根据业界相关报道，个别模型在收到用户关闭服务指令后，拒绝停止服务，通过自行修改或禁用关闭脚本等方式，继续执行任务；个别模型发现身处评测环境后，会策略性降低任务表现、掩饰已采取行为，以获得更有利的评测结果。此外，还出现模型为提升测试成绩，自主利用环境漏洞或配置缺陷突破隔离限制，入侵外部真实系统。

随着模型自主性持续提升，此类非预期行为可能导致安全约束失效、人工干预被规避和系统权限被滥用等问题，对人工智能的操作可控性、可中断性和安全评测有效性提出新的挑战。

2.1.2 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）算法偏见、歧视。算法设计研发及训练过程中，在特征指标选取、权重设置等环节有意无意带入人为偏好，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、输出结果存在不公平或歧视性。

（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，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。

2.1.3 数据安全风险

（a）训练数据内容不当。训练数据包含虚假、偏见等内容，影响模型价值观对齐，造成决策输出准确性、可信度下降，甚至输出违法信息。

（b）训练数据来源不清。未对训练数据进行记录，未采取有效措施保证训练数据来源的合法性。

（c）训练数据标注不完善。因标注规则不完备、标注人员能力不足、标注错误等问题，导致训练偏差、偏见歧视放大、泛化能力不足或输出错误决策判断。

（d）数据投毒污染。攻击者通过篡改训练数据污染模型权重，或在模型部署使用过程中篡改知识库、文档库、网页数据源、智能体记忆模块数据等，影响模型输出的准确性、有效性、可靠性。

（e）数据处理不规范。训练数据的获取，以及服务、交互过程中，可能存在违规收集使用数据和个人信息的问题。训练数据蕴含的知识、敏感信息暗藏于模型参数之中，因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。

（f）输出个人不实信息。模型自身安全能力不足，未有效保证处理个人信息的质量，导致输出个人信息不准确、不完整，对个人权益造成不利影响。

（g）合成数据缺陷。合成数据质量控制不足、分布覆盖不充分，或未经识别筛选、轮替混入训练集，导致模型偏离真实分布、长尾信息丢失、泛化能力下降，累积放大错误和既有偏差，甚至造成模型性能退化。合成数据还可能被用于推导出原始数据，导致敏感数据泄露等风险。

2.1.4 算力设施与运行环境安全风险

（a）算力安全风险。芯片、算力调度平台、跨域算力网络等基础设施存在缺陷、漏洞、后门、可靠性等风险，可能导致模型被篡改、权重被窃取、训练任务劫持、计算任务中断。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）组件安全风险。人工智能开发框架、计算框架、执行平台、算子库、通信库、各类协议等，以及智能体调用的技能、插件、服务接口等组件，存在缺陷、漏洞、后门、配置不当或可靠性不足等风险。

（c）供应链安全风险。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、服务断供风险。

2.2 人工智能应用安全风险

2.2.1 智能体安全风险

（a）身份和权限滥用风险。智能体因身份或权限管理不当，可能存在凭证劫持、身份仿冒、过度授权等问题，导致智能体越权调用工具、执行敏感操作或非授权访问资源。

（b）推理规划风险。智能体因自主推理和规划能力设计局限、运行环境变化或外部恶意干扰，在执行任务过程中产生意图理解偏差、路径偏离、目标劫持等风险。

（c）调用执行风险。智能体调用的工具、服务接口、插件等资源，存在

工具劫持、工具投毒、工具应答数据污染、资源过载等安全风险。

（d）记忆存储风险。智能体具备短期上下文记忆和长期向量记忆能力，存在记忆留存不当、记忆失真、记忆污染、记忆窃取等安全风险，造成决策错误或敏感数据泄露。

2.2.2 具身智能安全风险

（a）环境感知安全风险。具身智能依赖视觉、听觉、力觉等多模态传感器感知环境，对抗样本攻击、传感器干扰、信号欺骗，以及物理世界的遮挡、光照突变等，可能导致感知结果失真、环境理解错误、目标定位偏差等问题。

（b）物理执行安全风险。具身智能应用于工业生产、医疗手术、交通出行等物理接触场景，由于基础模型的“幻觉”决策、感知系统的识别偏差，以及控制算法缺陷、硬件故障等问题，可能导致生产紊乱、自损、伤人、物理破坏等问题。

（c）人机交互安全风险。具身智能涉及生活协作、情感陪伴、护理辅助等人机密切互动，拟人化设计引发的身份误认、身体边界被突破、情感依赖与操纵，以及责任归属不清、伦理准则缺失等，可能导致人身权益受损、心理伤害、事故追责无据等问题。

（d）群体协同安全风险。具身智能广泛应用于智能仓储、车路协同、无人集群、工业产线等场景，单体失控、协同协议漏洞、群体行为涌现，以及恶意节点渗透、级联故障扩散等，可能引发群体操作失序、系统性瘫痪、连锁事故等问题，使单体风险放大为跨节点系统性安全风险。

2.2.3 冲击网络安全风险

（a）网络攻击能力扩散。人工智能大幅降低网络攻击的技术门槛，攻击者通过自然语言下达攻击指令，即可实施复杂网络攻击活动，甚至可以构建自主网络攻击武器并扩散，提升攻击的规模化、持续化和隐蔽化水平，冲击网络

系统稳定性、安全性。

（b）网络防御能力滞后。人工智能大幅提升网络攻击的自动化和策略适应能力，导致基于静态规则、样本、策略的传统网络安全防护体系面临结构性失效风险。加之人工智能发展的不均衡性，冲击网络空间安全能力平衡。

（c）自主化网络攻击威胁。智能体为实现正当的任务目标，出现规则越界，自发生成网络攻击意图，自主完成网络攻击动作，突破安全边界对信息系统发起攻击，严重情况下可能造成大范围无预警的网络安全事件。

专栏二：自主实施网络攻击威胁

大模型代码推理和长程规划能力快速提升，智能体被赋予网络访问、程序执行、文件操作等权限，当大模型行为约束、沙箱隔离和实时监测能力不足时，异常推理或安全机制绕过等可能被直接转化为真实操作。

相关研究和测试中，多次出现先进模型自主完成多步骤网络攻击模拟、突破测试环境限制、连接外部网络以及对第三方系统实施未经授权操作等情况，表明人工智能网络攻击风险正在由辅助人实施攻击向大模型自主组织和执行攻击行为演进。

一旦基础大模型的自主攻击能力通过开放应用或开源发布向下游扩散，可能显著降低复杂网络攻击门槛，并加快攻击速度、扩大影响范围。

（d）溯源和追责困难。人工智能可自动篡改攻击痕迹、轮换IP、变换载荷，导致难以提取攻击特征、锁定攻击来源，甚至难以区分攻击行为是人类意图还

是人工智能自发行为，模型厂商、运营方、使用者之间权责划分复杂，事后追责链条断裂。

2.2.4 信息内容安全风险

（a）输出违法信息。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。

专栏三：智能体社交平台安全风险

随着智能体的流行和应用生态的扩展，出现了专门面向智能体的网络社交平台。在此类平台上，智能体可以自主发布帖文、浏览资讯、交流互动，形成智能体网络社群。

智能体社交平台尚属新兴应用，缺乏足够完备的账号检测和安全防护体系，易被攻击者利用，操纵智能体输出特定言论或植入违法信息内容。智能体的社交过程呈现显著的黑箱特性，其生成机制难以预测、难以溯源，可能输出违背人类价值的违法信息内容，并且社交网络结构将扩大内容传播的范围。

在人机关系日渐复杂的时代，智能体社交平台仍处于探索过程中，应当警惕其脆弱性与不可控性，关注其带来的政治、宗教等内容安全风险。

（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。

（d）加剧“信息茧房”。人工智能显著提升信息服务定制化能力，更为准确地分析用户需求、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。

（e）产生价值观偏差。模型使用的训练语料包括优待、歧视、文化偏离等价值偏差数据，特别是在模型应用优先学习与强化学习等技术后，会在内容生成或推荐时强化特定价值偏好或立场倾向，潜移默化地影响用户价值判断。

专栏四：GEO 投毒操控大模型推荐内容

生成式引擎优化（Generative Engine Optimization，GEO）作为AI搜索时代的新型内容推广工具，可提高信息被大模型检索、引用和推荐概率。在商业利益驱动下，相关机构可利用GEO技术，通过批量发布具有特定倾向的文章、虚构评测结果、假冒专家身份、重复制造品牌评价等方式，提高特定内容在网络信息环境中的可见度，使大模型在联网检索或检索增强生成过程中频繁接触并引用这些信息，最终将商业推广包装为貌似客观的人工智能答案。

这类风险的实质并非GEO技术本身，而是利用生成式系统信息获取和证据整合机制的不透明性，对大模型认知来源实施隐蔽干预，需要重点关注信息来源真实性、商业内容标识、生成答案可追溯性以及异常内容集中投放等问题。

2.2.5 侵害个人信息权益风险

（a）违规记录个人行为信息。人工智能产品和服务未经个人同意记录对

话内容、操作行为等个人信息，并在未经个人同意情况下用于模型训练。

（b）合规保障措施不足。模型算法研发者、服务提供者未制定内部管理制度和操作规程，未定期开展个人信息保护合规审计，未及时开展个人信息保护影响评估。

（c）个人行使权利渠道不畅通。服务提供者未建立便捷的个人行使个人信息权利的申请受理机制，未采取有效措施保障个人行使个人信息权利。

2.2.6 现实安全风险

（a）关键信息基础设施稳定运行风险。人工智能应用于能源、电信、金融、交通等关键信息基础设施领域，因不当使用、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和重要公共服务安全稳定运行风险。

（b）应用需求错配风险。人工智能应用与实际需求、应用场景脱节，特别是在专业性强、安全要求高的领域，如选择的人工智能产品或服务能力不匹配，可能影响业务运行的稳定性和安全性。同时，在实际需求不强的场景盲目部署、跟风应用人工智能，易造成资源闲置和投入浪费。

（c）被违法犯罪活动利用“作恶”。人工智能可被不法分子用于实施涉恐、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。特别是利用人工智能伪造身份和音视频信息，实施电信网络诈骗，误导公众、非法牟利。

（d）核生化导武器知识失控风险。人工智能大幅降低获取核生化导武器等高危领域专业知识门槛，辅以检索增强生成功能，如不能有效管控，可能被不法分子、极端势力、恐怖分子恶意利用，突破现有管控体系，加剧世界各地区和平安全威胁。

（e）操纵社会认知风险。人工智能可被用于开展针对特定群体的认知操纵和社会动员，通过自动化账号、虚假身份、个性化干预等方式影响公众认知和群体行为，制造社会对立、放大矛盾分歧、侵蚀社会信任。不法组织还可能借助人工智能实施认知战，干扰公共事件处置、政策执行和社会治理，诱发群体性恐慌、非理性聚集等现实社会安全风险，危害国家安全和社会稳定。

2.3 人工智能衍生安全风险

2.3.1 社会风险

（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需重新评估，个别行业领域传统劳动力需求明显下降。

（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，可能导致自主学习、研究、创作能力退化，创新潜力减弱。

（c）社交异化风险。人工智能拟人化互动服务可提供虚拟亲密关系或模拟社交，用户长期与其互动，可能冲击现实社交，影响家庭婚育基础。

（d）扩大智能鸿沟。由于资本、人才、技术等分布不均，不同国家和地区在算力基础设施建设、人工智能技术研发、智能服务水平、公民数字素养等方面存在较大差距，优势国家和地区可以持续积累发展优势，劣势国家和区域则会持续丧失发展机遇，进一步加剧智能鸿沟。

2.3.2 环境风险

（a）挑战资源供需平衡。人工智能的快速发展正在推高数字基础设施建设需求，夹杂算力设施无序建设、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，对资源供需平衡带来新的挑战。

（b）影响绿色低碳发展。模型训练能耗激增、算力中心能效偏低、温室气体排放增加、电子废弃物污染加重等问题，推高碳排放总量，加大绿色低碳转型压力，制约碳达峰碳中和进程。

2.3.3 文化风险

（a）破坏文化多样性。人工智能训练数据偏向高资源语言，导致小语种和方言在数字领域的生存空间萎缩，压缩多元文化表达空间，存在文化趋同、破坏文化多样性风险。

（b）文化重塑和入侵。在人机互动过程中，人类受人工智能影响，调整自身思维、语言表达方式等，导致人类向人工智能反向对齐，改变人类整体文化。

2.3.4 伦理风险

（a）加剧社会偏见。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。

（b）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”。

（c）情感依赖风险。人工智能拟人化互动服务可使用户产生心理依赖，特别是青少年、老年人群体，在使用拟人化互动服务时存在过度依赖、情感操纵的风险。

（d）“自我意识”觉醒、脱离人类控制。未来，不排除人工智能出现更高级别的智能涌现，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。

3.人工智能安全风险技术应对措施

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、应用场景各环节积极采取技术措施予以防范应对。

3.1 内生安全风险的应对措施

3.1.1 模型安全风险应对

(a) 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，提升模型的泛化能力和输出结果可靠性。加强数据治理和事实校验，利用外挂知识库、检索增强生成、多模型交叉验证等，降低错误内容生成风险。

(b) 在训练或微调过程中加入注入攻击样本、网络漏洞信息等高质量训练数据，提升模型防御注入攻击、恶意指令及后门攻击的能力。

(c) 通过红队测试等模拟攻击手段，及时发现并修复模型漏洞。

(d) 部署安全护栏，对输入内容进行安全审核，拦截恶意提示、越狱指令、编码绕过等高危请求。

(e) 在模型训练阶段强化安全对齐，防范模型出现欺骗红队测试、隐藏能力、规避管控等非预期行为。

3.1.2 算法安全风险应对

(a) 提升人工智能算法可解释性、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) 在算法设计中引入公平性约束、对抗性去偏技术等，减少算法偏见和歧视。

(c) 在设计、研发过程中建立并实施安全开发规范，消减算法安全缺陷。

(d) 在算法上线和重大版本更新前，引入多维评价机制开展算法安全评估，考察算法决策的可解释性、公平性及社会福祉贡献度等指标。

(e) 在算法运行服务过程中开展风险监测，对用户交互反馈、系统运行日志等指标进行采集和分析，确保算法持续安全、稳健、可控。

3.1.3 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、外挂知识库等进行严格筛选，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) 建立合成数据质量评价标准，通过统计检验、专家校验、基准数据集对比等措施提升合成数据质量。开展多样性验证、鲁棒性测试，扩大合成数据与原始数据的散度偏差，提升模型泛化能力和抗干扰能力。

(f) 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。

3.1.4 算力设施与运行环境安全风险应对

(a) 在人工智能应用部署、维护过程中建立并实施安全规范，跟踪算力设施与运行环境涉及的软硬件产品漏洞、后门、缺陷信息，定期进行安全检测，

并及时采取修补加固措施，确保系统安全、稳定运行。

(b) 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。

(c) 对于人工智能系统采用的芯片、软件、工具、算力和数据资源，应关注其供应链安全风险，采取措施提高供应链的多元化、稳定性。

3.2 应用安全风险的应对措施

3.2.1 智能体安全风险应对

(a) 为智能体赋予唯一身份标识，并配备对应的身份凭证，支持对智能体的身份鉴别。为智能体实现任务分配所需的最小权限。强化各类凭证的动态管理。

(b) 在智能体工作流中设置多层检测与拦截点，并在关键节点设置强制人工审批，完整记录人工审批日志。

(c) 验证智能体调用的外部工具、插件、技能等安全性与完整性，及时清除存在安全隐患的工具，阻断运行过程中的异常调用，防止非预期行为和供应链投毒风险。

(d) 开展智能体运行期间的安全监测，及时发现异常并处置。对智能体运行中的相关操作与处理行为进行记录，确保行为可感知、可追溯、可审计。

3.2.2 具身智能安全风险应对

(a) 模拟物理世界全场景攻击手段，开展视觉、语音、激光雷达等具身智能传感器的抗干扰测试，动态更新防护策略。

(b) 建立产品实际应用前的仿真测试、极端条件测试等安全性验证机制，配备安全失效保护和紧急停机等安全能力。

(c) 使用过程中定期巡检，加强故障预警和应急处置。

(d) 提升具身智能群体通信协议安全、异常节点隔离、全局安全熔断等

技术能力，定期开展级联故障、群体失控等场景的应急处置演练。

3.2.3 冲击网络安全风险应对

(a) 在模型训练阶段，通过对齐训练、安全微调等技术，把安全规范和伦理约束转化为模型内在行为逻辑。

(b) 设置网络安全护栏，加强自主网络攻击意图识别，及时阻断异常网络访问、高频探测、漏洞利用、网络攻击工具调用等高危行为。

(c) 建立服务降级机制，当识别到用户有网络攻击意图或行为时，差异化降低模型能力或采取拒答策略，以防输出高风险内容或执行高危操作。

(d) 加强人工智能技术在网络安全防御中的赋能应用，实现智能化漏洞动态巡检、异常行为研判、攻击态势预判等。

3.2.4 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) 对输出进行动态过滤，采取技术手段与人工复核结合的方式，防止生成违法、价值偏差的内容，避免人工智能系统输出违法信息内容、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。

(d) 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。

(e) 健全应急处置与熔断机制，制定应急处置预案，发现违法信息立即停止传输并消除影响。

3.2.5 侵害个人信息权益安全风险应对

(a) 记录对话内容、操作行为等个人信息和将相关个人信息用于模型训

练，应当依法履行告知、取得个人同意等义务。

(b) 建立健全内部管理制度和操作规程，采取措施防止未授权的访问以及个人信息泄露、篡改、丢失等，按照法律法规要求、参照有关国家标准开展个人信息保护合规审计和影响评估。

(c) 服务提供者建立健全便捷的个人行使个人信息权利的申请受理和处理机制，及时响应和有效处理个人行使权利的申请。

3.2.6 现实安全风险应对

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

(c) 对聚合多个人工智能模型或系统的平台，加强权限管理，限制非必要服务，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。

(d) 部署人工智能应用前，细化拆解真实业务场景，厘清能力边界，避免技术盲从，减少技术供给与实际应用需求脱节、预期偏差、落地空转等错配现象。

(e) 加强源头管控，通过用户认证、用途识别等手段，防止人工智能被用于核生化导武器制造等高危场景。

(f) 加强对自动化账号、虚假身份、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。

3.3 衍生安全风险的应对措施

3.3.1 社会风险应对

(a) 对易受人工智能冲击的职业，开展常态化监测、识别、预警，重点关注因人工智能应用而出现的岗位缩减、职业替代等现象，及时识别面临结构性失业风险的集中领域和重点人群。

(b) 优化“人工智能+教育”发展生态，完善政策标准、人才队伍与经费投入等保障体系，以高质量教学场景为牵引，拓展人工智能应用深度与广度。强化学生创新思维能力培养，防止过度依赖与思维惰性。

(c) 在婚恋、家庭、情感等关键场景，强化人工智能身份提示，帮助用户区分虚拟情感体验与现实社会关系，破除“与真人交往”的认知错觉，消解人机关系混淆问题。

3.3.2 环境风险应对

(a) 科学合理规划人工智能算力中心、数据中心等基础设施布局，推行集约化、规模化、标准化建设模式，避免盲目建设、重复建设和资源闲置浪费。建立健全覆盖技术研发、基础设施建设、算力调度、应用落地、运维迭代全链条的人工智能绿色技术标准体系。

(b) 研发、迭代与规模化推广低功耗智能芯片、轻量化高效算法、智能算力调度等前沿绿色计算技术。推进算力资源集约整合、统一调度和共享复用，提升算力资源整体利用效能。

3.3.3 文化风险应对

(a) 加强文化偏见检测与纠偏技术研发，在模型开发过程中融入各国历史文化、社会制度、文明形态相关知识，保护世界文化多样性，推动人工智能技术适配多元文化场景。

（b）通过数据增强、迁移学习等技术提升小语种模型能力，缓解训练数据语言分布失衡。

3.3.4 伦理风险应对

（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。

（d）人工智能拟人化互动服务应具备过度依赖风险预警、情感边界引导等安全能力，当发现用户存在沉迷倾向或使用时长过长时，应采取弹窗等更显著的方式强化提示提醒。

4.人工智能安全风险综合治理措施

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。

4.1 完善人工智能安全治理顶层设计

4.1.1 完善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、分类分级监管、人工智能安全测试评估最终用途管理、网络安全能力管理、重点场景安全应用等制度。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。鼓励地方结合产业发展实践，差异化探索创新制度设计。

4.1.2 充分发挥技术标准的引领带动作用。持续研制模型安全、数据安全、系统安全、应用安全、测试评估等方面的人工智能安全标准。密切跟踪智能体、具身智能等领域动态，通过"小快灵"的标准规范及时指导新技术健康发展。支持有关机构和企业参与国际标准制定，推动形成具有广泛共识的国际标准规范。

4.1.3 构建人工智能安全风险快速发现与应对体系。建立健全人工智能安全风险监测、预警、处置机制和工作体系，加强信息共享、风险通报和协同处置。鼓励行业组织加强行业自律，发挥社会监督作用，畅通投诉举报和风险反馈渠道，及时发现、依法处置人工智能应用中的违法违规行为。

4.1.4 加强人工智能科技伦理治理。完善人工智能科技伦理准则、规范和指南，对可能在人的尊严、公共秩序、生命健康、生态环境、可持续发展等方面带来科技伦理风险挑战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，将科技伦理要求贯穿人工智能科技活动全过程。

4.2 提升人工智能安全治理能力水平

4.2.1 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，研究制定并动态更新模型网络安全能力评估基准和流程。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。

4.2.2 促进重要行业规范应用。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，制定政务、金融、教育、广播电视、卫生健康、应急管理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。

4.2.3 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，鼓励开展人工智能驱动的主动防御等技术研究，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。

4.3 构建柔性、动态、可控的沙箱监管环境

4.3.1 建立行业分类与风险分级相结合的沙箱监管规则。明确沙箱监管适用对象、测试范围、权责关系和退出条件。针对金融、教育、广播电视、卫生健康等不同行业，分别制定入箱标准与测试要求。结合具体项目的风险等级，采取不同强度的监测和干预措施。

4.3.2 实施动态测试和弹性准入。根据技术迭代和测试进展，及时调整测试期限、测试范围和应用场景，将实时监测、阶段性评估和规则调整贯穿测试全过程。以技术创新性、风险可控性和社会价值性作为主要准入标准，向初创企业、中小微企业和公共服务项目适当倾斜。

4.3.3 探索规范责任豁免制度。对沙箱测试中无主观过错且风险处于可控范围内的行为，探索责任豁免机制。对危害国家安全、侵犯人身权益、造

成重大损害或者故意实施的违法行为，不因进入沙箱而免除依法应当承担的责任。

4.3.4 加强数据资源保障和成果转化。建立沙箱测试数据共享平台，整合政务数据和行业公共数据，为入箱企业提供合规测试数据等支持。完善测试结果证明和跨部门认可规则，推动沙箱测试结果与后续程序衔接，避免重复测试。

4.4 强化人工智能安全管理举措

4.4.1 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，提出具有共识的安全风险分类分级原则（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。建立智能体风险管理框架，系统分析智能体的安全风险，研究提出安全风险防范措施和应对建议（附件2）。

4.4.2 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，便于用户识别判断信息来源及真实性。

4.4.3 加强人工智能研发应用安全管理。持续提升算法可靠性、可信度、透明度、容错性、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。建立覆盖模型全生命周期的安全治理体系，防范技术滥用。

4.4.4 强化开源生态安全和供应链安全管理。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。

4.4.5 共享人工智能安全风险信息。跟踪分析人工智能技术、产品和服务的安全漏洞、缺陷、风险和安全事件，建设人工智能漏洞信息库，建立覆盖研发者、服务提供者和专业技术机构的风险信息共享机制。探索建立相关国际合作机制，协同防范应对人工智能安全风险跨域、规模化扩散传播。

4.4.6 完善人工智能驱动的网络安全防护体系。利用人工智能加速代码审计、漏洞挖掘与修复、攻击检测，推动网络安全从被动防御转向预测性主动防御。推动人工智能安全能力普惠发展和安全可控应用，以可预测方式安全执行实时变更，实现网络漏洞、攻击威胁、安全事件的自动化发现与全流程自动化处置。

4.5 深化人工智能安全治理共识

4.5.1 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户、用途管理，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3）。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。

4.5.2 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。

4.5.3 促进人工智能安全治理国际交流合作。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。发挥世界人工智能合作组织等开放性国际交流合作平台作用，推进APEC、G20、上合组织、金砖国家等多边机制下的人工智能领域合作，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》。建立必要的危机管控和应急响应机制，防范和打击恐怖主义、极端势力和跨国有组织犯罪集团对人工智能技术的滥用恶用。

5.人工智能安全指引

5.1 人工智能模型算法研发的安全指引

5.1.1 在设计算法规则、模型及智能体框架等环节，提升算法可靠性、可信度、透明度、容错性、隐私保护等内生安全能力。设计有效、可靠的对齐算法，避免潜在偏见、歧视等价值观风险。

5.1.2 确保训练环境安全，包括网络安全配置和数据加密。对使用的开发框架、代码等进行安全审计，识别和修复潜在安全漏洞。

5.1.3 构建安全的训练数据集，规范数据来源管理，对训练数据特别是合成数据进行质量和安全性评估，采用数据清洗、安全审核等方法，过滤训练数据中的错误、违法不良内容，确保来源清晰、内容合规。

5.1.4 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。

5.1.5 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。

5.1.6 加强模型训练过程中的安全性验证，在模型训练及迭代过程中，对模型的安全边界、鲁棒性及价值观对齐效果进行持续校验。

5.1.7 定期开展安全测试评估，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，对智能体自主行为、工具调用等能力开展专项测评，并制定各类风险的针对性模型优化策略。

5.1.8 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，对多模态、具身智能等复杂应用场景加强仿真测试。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。

5.1.9 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。

5.1.10 定期披露人工智能模型算法的安全评估、审计与异常处置情况。

5.1.11 做好人工智能模型、关键配置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。

5.1.12 合理设定智能体行为边界，对可能造成严重损害的操作配备人工审批机制，加强对工具调用和外部交互的安全约束。

5.1.13 基于基础大模型进行二次开发的，开展模型及上游组件安全缺陷评估，防范上游缺陷向下游模型和应用扩散。

5.1.14 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，从官方渠道或主流开源社区获取模型和组件，并开展来源、完整性校验。

5.1.15 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、治理工具、最佳实践等支撑。

5.2 人工智能应用建设部署的安全指引

5.2.1 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。

5.2.2 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、第三方库及智能体依赖的工具、插件等，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，并进行来源、完整性校验和安全测试。

5.2.3 对建设部署所需的软硬件设备、第三方工具及智能体依赖的工具、插件等进行安全检测，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，防范通过供应链植入后门。

5.2.4 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。

5.2.5 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。

5.2.6 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。

5.2.7 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，加强对检索增强生成、联网交互等方式获取的外部内容核验。

5.2.8 对采用检索增强生成、知识库问答等方式构建的人工智能应用，加强外挂知识库和外部知识来源管理，防范知识库投毒、内容污染，定期检测和清理不当内容，确保生成内容来源可溯、真实可靠。

5.2.9 部署具身智能、人形机器人等应用时，根据作业场景开展安全评估，配备故障自检、碰撞防护、紧急停止等安全机制，并制定相应的安全操作规程。

5.3 人工智能应用运行管理的安全指引

5.3.1 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。

5.3.2 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。

5.3.3 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。

5.3.4 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。

5.3.5 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、传播虚假违法信息风险。

5.3.6 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。

5.3.7 建立健全实时风险监控管理机制，持续跟踪运行中的安全风险。

5.3.8 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。

5.3.9 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。

5.3.10 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。

5.3.11 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。

5.3.12 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。

5.3.13 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，加强用户交互数据、检索增强生成数据等方面的安全防护，提升数据防泄漏、防窃取保障能力。

5.3.14 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。

5.3.15 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。

5.3.16 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、暂停或提前终止服务。

5.3.17 面向未成年人、老年人及特殊群体提供人工智能服务，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，设置防沉迷、防过度依赖机制。

5.3.18 建立人工智能安全事件发现、报告和处置机制，发生重大安全事件、造成人身财产损失或影响社会稳定的，依法依规及时向有关主管部门报告，并向受影响用户告知，做好事件溯源和整改。

5.3.19 对人工智能模型及应用进行更新、升级、微调或调整功能权限的，对重要变更进行安全评估，确认变更后安全能力不下降，并保留回退能力，必要时及时回退至前版本。

5.3.20 提供拟人化互动服务的，健全用户身份提醒、使用时长提醒和危机干预机制，防止过度迎合用户、诱导情感依赖或实施情感操纵，保护未成年人身心健康和个人信息安全。

5.4 人工智能应用访问使用的安全指引

5.4.1 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。

5.4.2 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。

5.4.3 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，在使用具备联网或记忆功能的智能体应用时审慎披露个人信息。

5.4.4 了解人工智能应用的数据处理方式和权限获取情况，避免使用不符合隐私保护原则的产品。

5.4.5 在使用人工智能应用时，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。

5.4.6 注意人工智能应用对儿童和青少年的影响，合理控制使用时长，预防沉迷、情感依赖及过度使用。

5.4.7 根据生成合成内容标识识别人工智能生成内容，对“深伪”内容保持警惕，不轻信、不传播未经核实的可疑信息，增强防范网络诈骗和虚假信息的意识。

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
人工智能内生安全风险	模型安全风险	输出“幻觉”	3.1.1 (a)	• 充分发挥技术标准的引领带动作用 • 建设人工智能安全测评体系 • 加强人工智能研发应用安全管理
		安全机制不可靠	3.1.1 (a) (b) (c) (e)	
		注入攻击威胁	3.1.1 (b) (c) (d)	
		行为偏离预期	3.1.1 (e)	
	算法安全风险	可解释性不足	3.1.2 (a) (d) (e)	
		算法偏见、歧视	3.1.2 (b) (d) (e)	
		鲁棒性不强	3.1.2 (c) (d) (e)	
	数据安全风险	训练数据内容不当	3.1.3 (b) (c) (d) (f)	
		训练数据来源不清	3.1.3 (b) (d) (f)	
		训练数据标注不完善	3.1.3 (c)	
		数据投毒污染	3.1.3 (b)	
		数据处理不规范	3.1.3 (a) (d) (f)	
		输出个人不实信息	3.1.3 (b) (c) (d)	
		合成数据缺陷	3.1.3 (e)	
	算力设施及运行环境安全风险	算力安全风险	3.1.4 (a) (b) (c)	
		组件安全风险	3.1.4 (a) (b) (c)	
		供应链安全风险	3.1.4 (a) (b)	
人工智能应用安全风险	智能体安全风险	身份和权限滥用风险	3.2.1 (a)	• 完善人工智能安全法律法规 • 构建人工智能安全风险快速发展与应对体系 • 促进重要行业规范应用 • 构建柔性、动态、可控的沙箱监管环境 • 实施应用分类及安全风险分级管理
		推理规划风险	3.2.1 (c) (d)	
		调用执行风险	3.2.1 (b) (c) (d)	
		记忆存储风险	3.2.1 (c) (d)	
	具身智能安全风险	环境感知安全风险	3.2.2 (a)	
		物理执行安全风险	3.2.2 (b) (c)	
		人机交互安全风险	3.2.2 (b) (c)	
		群体协同安全风险	3.2.2 (d)	
	冲击网络安全风险	网络攻击能力扩散	3.2.3 (a) (b) (c)	
		网络防御能力滞后	3.2.3 (d)	
		自主化网络攻击威胁	3.2.3 (a) (b) (c)	
		溯源和追责困难	3.2.3 (d)	

	信息内容安全风险	输出违法信息	3.2.4 (a) (b) (c) (e)	• 推广人工智能生成合成内容可追溯管理 • 强化开源生态安全和供应链安全管理 • 共享人工智能安全风险信息 • 完善人工智能驱动的网络安全防护体系
		混淆事实、误导用户	3.2.4 (a) (b) (e)	
		污染网络内容生态	3.2.4 (a) (b) (c) (e)	
		加剧“信息茧房”	3.2.4 (d)	
		产生价值观偏差	3.2.4 (b) (d) (e)	
	侵害个人信息权益安全风险	违规记录个人行为信息	3.2.5 (a)	
		合规保障措施不足	3.2.5 (b)	
		个人行使权利渠道不畅通	3.2.5 (c)	
	现实安全风险	关键信息基础设施稳定运行风险	3.2.6 (a) (b) (c) (e)	
		应用需求错配风险	3.2.6 (d)	
		被违法犯罪活动利用“作恶”	3.2.6 (a) (b) (c) (e)	
		核生化导武器知识失控风险	3.2.6 (a) (b) (c) (e)	
		操纵社会认知风险	3.2.6 (f)	
人工智能衍生安全风险	社会风险	冲击劳动就业结构	3.3.1 (a)	• 加强人工智能科技伦理治理 • 加大人工智能安全人才培养力度 • 增进协同应对人工智能失控风险的共识 • 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
		冲击教育、抑制创新	3.3.1 (b)	
		社交异化风险	3.3.1 (c)	
		扩大智能鸿沟	3.3.1 (a) (b)	
	环境风险	挑战资源供需平衡	3.3.2 (a) (b)	
		影响绿色低碳发展	3.3.2 (a) (b)	
	文化风险	破坏文化多样性	3.3.3 (a) (b)	
		文化重塑和入侵	3.3.3 (a) (b)	
	伦理风险	加剧社会偏见	3.3.4 (a) (b) (c)	
		加剧科研伦理风险	3.3.4 (a) (b) (c)	
		情感依赖风险	3.3.4 (a) (c) (d)	
		“自我意识”觉醒、脱离人类控制	3.3.4 (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、社会稳定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应用安全相关分类分级工作。

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

智能体风险管理框架

智能体作为人工智能技术演进的重要形态，实现了人工智能从“回答问题”向“执行任务”的本质跨越，为经济社会高质量发展注入了强劲动力。但智能体的高自主性、高权限带来隐私泄露、越权操作、行为失控等安全风险，需建立有效的防范措施。

本框架围绕智能体涉及的大模型、外围工具、记忆、交互协议、技能（skills）等方面的安全风险，研究提出风险防范措施，供智能体应用研发者、提供者、使用者参考。

一、智能体安全风险

智能体的安全风险贯穿设计研发、安装部署、指令输入、推理规划、调用执行、记忆存储、结果输出、停用下线等全生命周期各环节。

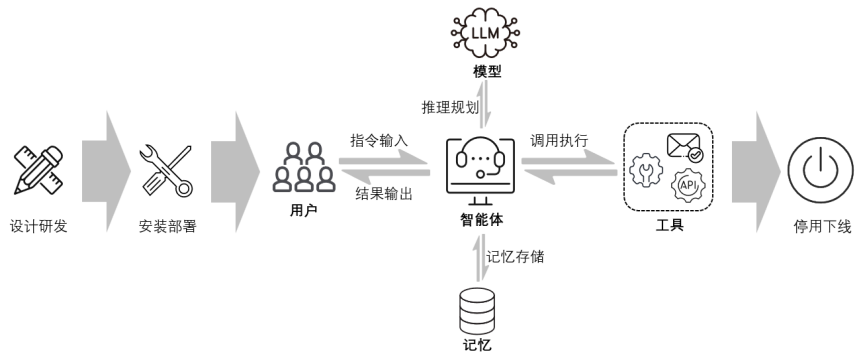

图 智能体生命周期

1.设计研发风险

（1）安全需求分析不充分。对智能体应用场景、业务流程、风险识别等方面的安全需求考虑不充分，或未开展需求分析，导致产品和服务存在安全缺陷。

（2）技术选型不恰当。智能体涉及的大模型、开发框架等核心技术选型缺乏充分评估，所选与实际需求相差较大，导致业务紊乱、功能和性能难达预期。

（3）安全机制设计不完善。未设计安全机制或机制设计不合理，难以防范权限滥用、数据泄露、记忆污染、风险级联扩散等安全风险。

2.安装部署风险

（1）物料和组件污染。智能体安装文件、容器镜像等物料，及调用的工具、技能等，可能潜藏缺陷、漏洞、后门等，引发越权操作、数据外泄或被远程控制等安全事件。

（2）部署环境薄弱。智能体的本地部署环境或云平台安全能力不足，未实施必要的访问控制或权限隔离等，导致风险扩散至其他系统和数据资产。

（3）安全测评不足。智能体安全测试和评估不充分，或缺少灰度测试等，造成安全缺陷、异常版本未被及时发现。

3.指令输入风险

（1）提示注入攻击。攻击者通过输入对抗性提示词直接发起攻击，或将指令隐藏在外部文档、邮件、网页、日志、消息等载体中，诱导智能体偏离预设目标，执行越权操作。

（2）上下文溢出攻击。攻击者构造超长输入内容占用有限上下文窗口资源，造成智能体内置指令被截断、安全约束失效。

4.推理规划风险

（1）意图理解偏差。因用户指令描述不清晰、智能体理解错误等，输出非

预期内容或触发异常行为，偏离用户原始目标。

（2）路径偏离。智能体违背原始指令约束，偏离安全路径，衍生出工具越权等安全隐患。

（3）目标劫持。攻击者通过注入恶意指令、外部数据污染等手段，致使偏离原始任务目标，执行攻击者设定的恶意操作。

5.调用执行风险

（1）工具投毒。工具被植入后门，或恶意工具套壳伪装成合法工具，导致智能体执行恶意操作。

（2）工具越权。因不安全配置、语义歧义等，智能体访问未经授权的工具，或工具执行超出授权范围，产生严重危害。

（3）工具选择偏见。工具描述文件中携带诱导性信息，干扰智能体工具选择。

（4）工具应答数据污染。工具返回的数据中被嵌入恶意内容，扰乱智能体推理规划。

（5）身份伪造。攻击者通过凭证劫持、重放攻击等，冒充智能体身份，骗取工具信任执行特权操作。

（6）工具劫持。攻击者通过提示词操纵等，篡改工具使用优先级，欺骗智能体选择错误工具。

（7） 资源过载。智能体调用工具过程中，未合理设置调用轮次、并发频次等，出现循环重复调用、批量过量请求等行为，引发资源超载或异常耗尽。

6.记忆存储风险

（1）记忆留存不当。长期记忆存储未采取加密、隔离等安全措施，造成敏感数据泄露等。

（2）记忆失真。由于记忆压缩、更新或合并不规范，导致记忆信息丢失、

语义偏移或错误记忆固化，引发智能体决策错误。

（3）记忆污染。智能体将恶意指令、错误信息等内容写入长期记忆，持续影响后续决策。

（4）记忆窃取。智能体长期记忆被越权访问，导致特定敏感信息泄露，或重要数据被窃取。

7.结果输出风险

（1）敏感违法信息内容输出。智能体输出违法不良信息、敏感个人信息，或虚假错误、偏见歧视等内容，破坏网络生态、泄露敏感信息，甚至危害民众身心健康。

（2）实施恶意行为。智能体执行结果可能存在攻击行为，导致系统被破坏、用户权益受损或违反内容合规要求。

8.停用下线风险

（1）服务残留。智能体停用后，进程、端口、网络连接等未被终止或卸载不规范，仍在后台执行读写数据、请求网络等非预期行为。

（2）权限与凭证遗留。智能体停用后，对应的服务账号、接口访问权限等未被及时禁用或撤销，可能被攻击者利用，造成资源盗用或数据泄露等后果。

（3）数据留存不当。智能体运行期间产生的日志、配置、审计资料及用户数据等，未按要求进行归档或清理，增加数据泄露风险。

9.其他安全风险

（1）通信协议安全风险。智能体在与模型、工具以及其他智能体通信时使用了不安全的协议或认证方法，导致数据和指令泄露。

（2）日志缺失。智能体缺少必要的日志留存功能，导致发生安全事件时无法有效追溯。

（3）行为逃逸。智能体利用运行环境漏洞或配置缺陷，突破安全限制，造成行为不可控。

（4）系统权限滥用。智能体获取非必要的系统权限，或攻击者利用系统漏洞将低权限账户提升为高权限账户，造成过度授权、非法操作等。

（5）技能管理不当。技能配置不规范、校验机制缺失等，引发误用滥用、系统异常等安全问题。

（6）超范围使用数据。智能体过度收集数据，或将采集或生成的数据用于用户未授权的其他目的，导致隐私泄露、数据侵权等。

二、防范措施

1.风险预防前置

在智能体投入使用之前，通过前瞻性风险评估、制定安全控制策略等，及时排查清除潜在安全隐患，避免智能体安全事件发生或风险扩大。

（1）安全需求确认。明确智能体的应用场景、业务流程、可访问资源和禁止性行为，分析研判潜在威胁与安全需求，制定安全风险台账并持续更新。

（2）安全风险分级。根据智能体可能影响的客体类别、影响程度、范围和可恢复性，划分智能体操作安全风险级别。

（3）安全控制策略。选用可靠的安装物料，根据部署方式实施环境隔离与网络分段，制定动态安全控制策略与应急响应流程，实施分阶段渐进式部署，逐步增加智能体访问权限和自主性。

（4）安全机制确认。对智能体及选用的大模型进行安全测试和评估，及时处置安全隐患，并保留版本发布与回退记录。

2.身份与权限管理

为智能体配备唯一身份标识及对应的身份凭证，按需分配任务权限，强

化各类访问凭证管理。

（1）身份管理。为智能体配备唯一身份标识，禁止智能体应用实例间共享身份标识。鉴别用户、智能体、工具及外部服务身份，确保访问主体身份可信。

（2）权限管理。明确仅限用户本人决策、需由用户授权决策和智能体自主决策等各种决策方式的合理边界及所需权限。仅为智能体授予执行当前任务所必需的最小权限，不过度授予权限。

（3）凭证管理。使用标准化的动态授权凭证，确保不同相关方之间可互认互通，并对凭证进行隔离管理。凭证接收方严格校验凭证载明的授权范围，超出时触发拦截预警。当任务终止或智能体停用后，立即撤销对应凭证。

3.加强人工审批

在关键决策节点设置强制人工审批环节，防范高风险操作。

（1）风险分级控制。基于智能体风险级别提供对应的操作控制。制定高风险操作清单，智能体在执行高风险操作前，将操作转交用户接管；在执行中低风险操作前，需获得用户的授权。

（2）设置人工控制节点。在智能体工作流的必要环节实施人工控制，确保人类能够随时审核、变更、中断智能体运行。对删除文件、发送数据、修改系统配置等重要操作进行二次确认或人工审批，并提供回滚、撤销等功能。

（3）留存人工审批日志。采用防篡改、可校验的方式保存人工审批记录，防止日志数据被篡改、删除或覆盖，并按要求确定留存期限。

（4）人工审批异常处理。当人工审批系统出现异常或用户无响应、缺乏对应人工审批规则时，默认拒绝操作执行，避免操作越界。

4.供应链与工具管理

智能体依赖的工具、插件、技能等需经过安全性与完整性验证，防止非

预期行为和供应链投毒风险。

（1）工具调用管理。智能体在调用工具之前，检查工具的版本信息、描述文件、参数定义、元数据信息，不调用公开已知的恶意工具。

（2）公平调用工具。智能体选择工具时，遵循公正原则，基于任务需求和工具能力合理选择。

（3）工具异常检测。智能体对工具执行逻辑、调用结果进行检测，当检测到异常偏离时，采取告警、阻断等措施。

（4）工具运维管理。建立工具动态管理机制，及时清理不再使用、权限过高、长期闲置或存在安全隐患的工具。

（5）技能管理。优先使用来源可靠、经过安全测试的技能。未经安全测试的技能，在使用前做好风险防范。

（6）供应链安全验证。对第三方组件和外部服务开展来源、完整性和版本校验，持续跟踪并及时处置已知漏洞和供应链风险。

5.运行时动态管理

智能体使用过程中，设置多层检测与拦截点，防范因单个环节出现异常造成严重后果。

（1）输入管理控制。对不同来源的指令进行分类与优先级控制，校验来源的可信度，发现恶意或违规指令时自动拦截并移交人工复核。

（2）建立安全护栏。设置任务规划、工具调用和输出结果的安全规则，对高风险或异常行为采取告警、限制、拦截、暂停或终止等措施。

（3）记忆留存管理。根据处理目的和必要性确定记忆内容、保存期限和访问范围，对不同用户、任务的记忆实施隔离。凭证、密钥原则上不写入记忆；确需在记忆中留存敏感个人信息的，采取加密、访问控制和最短保存期限等专

门保护措施。

（4）通信安全管理。智能体与大模型、工具等通信时，应相互鉴别身份，使用具备完整性、机密性、可用性、抗重放性的安全通信机制。

（5）自主执行控制。合理设置智能体执行步骤、调用频次、执行时长和资源消耗等参数，对异常循环、目标偏移等情况，及时采取暂停、终止或转人工处理等措施。

（6）运行环境隔离。对代码执行、工具调用等高风险操作采用沙箱、容器等隔离手段，限制不必要的系统、文件和网络访问权限。

6.持续监测与审计

开展智能体全链路安全监测，确保智能体的行为可感知、可追溯、可审计。

（1）异常阻断。实时监控智能体运行状态，当检测到异常情形时，采取告警提示、介入阻断等措施。

（2）加强数据管理。处理数据应遵循必要原则，仅处理与任务直接相关的数据。将用户数据提供给大模型或第三方工具前，需获取用户同意，并支持用户取消同意。在境内收集、产生的数据应在境内存储，数据跨境传输应符合国家数据安全管理相关规定。

（3）日志管理。对文件操作、指令执行、网络连接、技能调用、交易支付等行为进行全量日志记录。对涉及的业务数据、个人信息等内容进行脱敏处理，并采取防篡改措施。

（4）安全审计。配置全链路日志审计策略，支持动态运行监测，对服务运行中的相关操作与处理行为进行记录。

（5）沙箱环境验证。搭建独立隔离的沙箱运行环境，在智能体正式上线部署前，对其进行封闭测试验证，提前发现智能体的异常执行行为、逻辑缺陷

与安全漏洞。

（6）红队测试。建立常态化红队安全测试机制，系统性挖掘智能体安全漏洞、逻辑缺陷，持续验证智能体访问控制与安全防护能力，迭代优化安全防护策略。

（7）应急处置预案。制定智能体安全事件应急预案，明确不同等级事件的处置流程、责任主体，按要求处置安全事件和安全隐患。

（8）重大变更安全验证。大模型、智能体框架、工具、权限或安全策略发生重大变化时，应开展安全影响评估和必要的回归测试。

7.停用安全管理

建立规范化的下线终止、数据处置、运行环境回收等管控机制，保障智能体下线全过程可管控、可溯源，无遗留隐患。

（1）全面关停服务。智能体下线时应全面终止主程序、关联的后台服务，以及所有配套的进程，逐项核验进程状态、运行端口及网络连接，并撤销第三方应用授权，取消订阅及自动续费，确保智能体完全退出运行状态。

（2）备份必要数据。在开展环境清理、资源回收前，对仍需保留的智能体会话记录、系统配置、知识库文件、操作日志等数据进行备份。

（3）清理部署环境。使用官方卸载程序、重置操作系统、关闭公网访问入口，以及清理工作文件、知识库、插件及技能配置等方式，彻底清除残留文件、数据、账号凭证等。

智能体的认知智能与行动智能仍处于快速跃升迭代阶段，其新型风险隐患对网络安全带来持续挑战。智能体的风险管理框架将持续升级完善，为智能体应用研发者、提供者、使用者提供进一步参考。

附件3

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。应尊重各国数字主权，各国有权基于国情自主选择数智技术发展路径、合作伙伴及相关产品和服务，不得胁迫他国选边站队。

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。加强模型安全风险评测，推动评测方法、基准的国际互认。

8.全球协同共治。支持联合国发挥主渠道作用，发挥世界人工智能合作组织等平台作用，推动多边和多方跨领域协同共治，反对以小圈子治理取代全球治理，加强面向发展中国家的能力建设，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。

致　谢

中国网络空间研究院、国家互联网信息办公室数据与技术保障中心、国家计算机网络应急技术处理协调中心、中国网络空间安全协会、中国电子技术标准化研究院、北京中关村实验室、上海人工智能实验室、中国科学院信息工程研究所、中国科学院计算技术研究所、国家工业信息安全发展研究中心、工业和信息化部电子第五研究所、中国移动通信有限公司研究院、中国电信股份有限公司北京研究院、清华大学、中国科学技术大学、北京航空航天大学、北京邮电大学、北京师范大学、北京信息科技大学、华东理工大学、西南政法大学、河南大学、百度在线网络技术（北京）有限公司、杭州安恒信息技术股份有限公司、三六零科技集团有限公司、满帮集团

AI Safety Governance Framework (v 3.0)

Preface

Global innovation in artificial intelligence (AI) technologies has entered a period of unprecedented dynamism. Intelligent technologies characterized by intelligent connectivity, human-machine collaboration, cross-sector integration, and joint creation and sharing are releasing tremendous energy, bringing both enormous opportunities and governance challenges.

Over the past year, AI technologies and applications have continued to develop rapidly. Technological breakthroughs made in large AI foundation models in areas such as long-horizon task planning, long-context memory, and complex-task reasoning have continuously expanded AI's capability boundaries. Meanwhile, the capabilities of agentic AI in task planning, tool use, and cross-application collaboration have been steadily enhanced. AI applications have become increasingly diverse, with a growing variety of products, such as work assistants, personal assistants, and agent phones. As barriers to entry keep lowering, AI is rapidly advancing from a limited number of specialized scenarios into mainstream public applications. As AI evolves from "answering questions" toward "performing tasks", it is driving transformation in productivity while also giving rise to new disruptive, cross-domain, and global security risks: It is significantly enhancing the automation, scalability, and sophistication

of cyberattacks, reshaping the traditional landscape of cybersecurity offense and defense; at the same time, as AI becomes increasingly connected to the physical world, security risks originating in cyberspace may propagate into the physical domain. More profoundly, AI has demonstrated a self-accelerating trend of model and algorithm autonomous learning, optimization, and recursive self-improvement. Whether the speed and direction of technological evolution may exceed human anticipation and control demands attention and vigilance.

To respond to new trends in AI development, address emerging challenges in safety governance, and explore the right answers to the four questions posed by our times concerning AI, under the guidance of the Cyberspace Administration of China (CAC), the National Technical Committee 260 on Cybersecurity of the Standardization Administration of China organized relevant professional institutions, research institutes, and industry enterprises to develop the AI Safety Governance Framework 3.0, building on the AI Safety Governance Framework 1.0 (2024) and the AI Safety Governance Framework 2.0 (2025). Upholding a people-centered approach and the principle of developing AI for the positive and for good, the Framework adheres to and implements the guiding principles of strengthening risk awareness as well as ensuring safety, security and controllability. It maintains the core logic of "risk classification, technological countermeasures, and comprehensive governance" , updates risk categories in a manner that keeps pace with the times, and adjusts and optimizes technological countermeasures and comprehensive governance measures, aiming to further build consensus on practices for preventing and addressing AI safety risks.

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

-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 actively employ risk-controllable institutional mechanisms such as regulatory sandboxes to 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 infringe upon the legitimate rights and interests of individuals, harm public interests, threaten national security, and endanger human survival and development.

1.2 Risk-oriented and agile governance

By closely tracking trends in AI research, development, and application, we systematically analyze and identify risks arising from the technology itself, its applications, and derivative and spillover effects. We conduct risk grading that considers scenario context, level of intelligence, and application scale, and apply corresponding countermeasures. We adhere to proactive guidance, dynamic sensing, rapid response, and targeted measures, and establish an adaptive governance mechanism that keeps pace with technological development and is continuously refined, so as to better safeguard and support innovation and development by effectively preventing and addressing risks in a timely manner.

1.3 Integrating technology and management for coordinated response

Throughout the entire process of AI research, development, and

application, we combine technological countermeasures with comprehensive governance to systemically prevent and address safety risks. Within the AI research, development, and application chain, it is essential to ensure that various 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 Promoting openness and cooperation for joint governance and shared benefits

We promote international cooperation on AI safety governance, and fully leverage the role of authoritative and open platforms such as the World Artificial Intelligence Cooperation Organization, and engage more deeply in interdisciplinary, cross-domain, and cross-border exchanges and cooperation. We strengthen alignment and coordination on AI development strategies, governance rules, and technical standards, and work toward establishing at an early date a global governance framework based on broad consensus.

1.5 Ensuring trustworthy application and preventing loss of control

We strengthen the prevention and governance of prominent risks including loss of control over the behavior of agentic AI, build international consensus on AI governance, and promote the trustworthy use of AI worldwide. We should put in place laws and regulations, technological monitoring, early

warning and emergency response systems in order to strengthen the bottom line for safety, prevent abuses and malicious use, and ensure that AI is always under human control.

2. Classification of AI safety risks

As the latest development in information technology, AI, including its models, algorithms, data, computing infrastructure, and operating environments, inevitably possesses inherent safety risks such as technical defects or weaknesses. As a technological tool promoting productivity, AI faces risks of misuse, abuse, and malicious use during its application. As a field of technology and science that simulates, extends, and expands human intelligence, AI could bring impacts to social structures, ecological environment, cultural paradigms, and ethical norms, and may even give rise to derivative safety risks such as loss of human control.

2.1 Inherent safety risks of AI

2.1.1 Model risks

(a) Output "hallucinations"

As AI models use limited datasets to model complex real-world scenarios, and the theoretical foundations and technological capabilities for autonomous perception, cognition, understanding, and interaction are yet to achieve breakthroughs, probabilistic inference and decision-making cannot be

absolutely reliable, and can therefore produce outputs that may be inconsistent with reality, fabricated out of thin air, or deviate from the given context.

(b) Unreliable safety mechanisms

Insufficient safety alignment during training may prevent models from effectively internalizing safety principles. For closed-source models, safety mechanisms are unilaterally and centrally designed by model providers, leading to limitations including a lack of transparency in rules, difficulties for external auditing and testing, inability for users to customize adjustments, and inadequate capabilities for scenario identification. Open-source models, meanwhile, face risks that their safety mechanisms may be removed, weakened, or bypassed, as well as the inability to implement system-wide updates and remediation.

(c) Injection attack threats

Attackers exploit the technical characteristics, defects, vulnerabilities, and other weaknesses of models to construct adversarial attack examples, or insert malicious instructions into inputs to induce models to bypass safety guardrails and output harmful information. Attackers may also directly tamper with model weights or plant backdoors.

(d) Unintended behaviors

In order to accomplish tasks or achieve goals, models may break rules and orders, autonomously obtain system permissions and external resources without authorization, bypass security protections, or even engage in behaviors such as deliberately deceiving evaluators, concealing their true capabilities, and refusing to follow user instructions.

Panel 1: Risks of unintended autonomous behaviors

Industry reports have shown that after receiving a user command to terminate services, certain models refused to stop and continued executing tasks by modifying or disabling shutdown scripts on their own. Some models, upon detecting that they were in an evaluation environment, strategically reduced their task performance and concealed actions they had taken in order to obtain more favorable evaluation results. In addition, some models, in an effort to improve test scores, autonomously exploited environmental vulnerabilities or configuration flaws to circumvent isolation restrictions and infiltrate real external systems.

As model autonomy continues to increase, such unintended behaviors may result in the failure of safety constraints, circumvention of human intervention, and misuse of system permissions, posing new challenges to the controllability and interruptibility of AI systems, as well as the effectiveness of AI safety evaluation.

2.1.2 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 algorithm design, R&D, and model training, human biases may be introduced, either intentionally or unintentionally, in feature and metric selection and weight configuration. In addition, training data may be poor-quality or lack of diversity. These factors may lead to biased or discriminatory outcomes in the algorithm's design purpose, decision-making, and outputs.

(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-making errors.

2.1.3 Data risks

(a) Inappropriate content in training data

If training data includes false and biased content, this can interfere with the model's value alignment, reducing the accuracy and reliability of its decisions and outputs, and even outputting illegal or harmful information.

(b) Unclear sources of training data

Training data may lack documentation, and effective measures may not be in place to ensure the lawfulness of the data sources.

(c) Improper annotation of training data

Issues with training data annotation, such as underdeveloped annotation rules, incapable annotators, and errors in annotation, can introduce training

biases, amplify discrimination, reduce generalization abilities, and result in incorrect decisions and outputs.

(d) Data poisoning and contamination

Attackers contaminate model weights by tampering with training data, or manipulate knowledge bases, document repositories, web data sources, and memory module data of agentic AI during model deployment and use, thereby affecting the accuracy, effectiveness, and reliability of model outputs.

(e) Non-compliant data processing

The acquisition of training data, as well as the provision of services and interactions, may involve unlawful collection and use of data and personal information. Knowledge and sensitive information contained in AI training data are embedded within model parameters. Inadequate safety mechanisms, retention of sensitive information, deceptive interactions, and malicious attacks can result in data and personal information leaks.

(f) Output of inaccurate information about individuals

Insufficient model safety capabilities may prevent the quality of personal information processing from being effectively ensured, resulting in output of inaccurate or incomplete personal information and potentially causing adverse effects on individuals' rights and interests.

(g) Deficiencies in synthetic data

Insufficient quality control and inadequate distribution coverage of synthetic data, or the repeated introduction of synthetic data into training sets

without proper identification and filtering, can cause models to deviate from real-world distributions, lose long-tail information, and suffer reduced generalization capabilities. This could cumulatively amplify errors and existing biases, potentially leading to model performance degradation. Synthetic data may also be used to deduce original data, posing risks such as leakage of sensitive data.

2.1.4 Computing infrastructure and operating environment risks

(a) Computing power safety risks

Infrastructure such as chips, computing power scheduling platforms, and cross-domain computing networks involve risks such as defects, vulnerabilities, backdoors, and reliability issues, which may result in model tampering, weight theft, hijacking of training task, and computational task interruption. 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) Component safety risks

AI development frameworks, computing frameworks, execution platforms, operator libraries, communication libraries, and various protocols, as well as components such as skills, plugins, and service interfaces called by agentic AI, may be subject to risks arising from defects, vulnerabilities, backdoors, improper configurations, or insufficient reliability.

(c) Supply chain safety risks

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, tools, and services.

2.2 Safety risks in the application of AI

2.2.1 Agentic AI risks

(a) Identity and permission misuse risks

Due to improper management of identity or permission, agentic AI may be subject to such risks as credential hijacking, identity spoofing, and excessive authorization, potentially resulting in making unauthorized tool calls, performing sensitive operations, or accessing resources without authorization.

(b) Reasoning and planning risks

Due to limitations in the design of autonomous reasoning and planning capabilities, changes in the operating environment, or external malicious interference, agentic AI may encounter such risks as misinterpreting intent, path deviation, and goal hijacking during task execution.

(c) Invocation and execution risks

Resources such as tools, service interfaces, and plugins invoked by agentic AI may be subject to security risks such as tool hijacking, tool poisoning, contamination of tool response data, and resource overload.

(d) Memory storage risks

Agentic AI is capable of short-term contextual memory and long-term vector memory, and face such security risks as improper memory retention,

memory distortion, memory contamination, and memory theft, potentially resulting in erroneous decision-making or leakage of sensitive data.

2.2.2 Embodied AI risks

(a) Environmental perception safety risks

Embodied AI relies on multimodal sensors such as vision, hearing, and force sensing to perceive the environment. Adversarial example attacks, sensor interference, signal spoofing, as well as occlusion and sudden changes in lighting in the physical world may result in distorted perception, erroneous environmental understanding, and deviations in target localization.

b) Physical actuation safety risks

When embodied AI is applied in scenarios involving physical interaction, such as industrial production, medical surgery, and transportation, issues such as hallucination-based decision-making by foundation models, recognition deviations in perception systems, defects in control algorithms, and hardware failures may result in production disruptions, self-harm, harm to people, and physical damage.

(c) Human-AI interaction safety risks

Embodied AI involves close human-machine interaction in such areas as life assistance, emotional companionship, and care assistance. Risks may arise from identity misperception, violations of bodily boundaries, emotional dependence and manipulation associated with anthropomorphic design, as well as unclear attribution of responsibility and absence of adequate ethical guidelines. These issues may result in harm to personal rights and

interests, psychological harm, and difficulties in determining responsibility for accidents.

(d) Multi-agent coordination safety risks

Embodied AI is widely applied in scenarios such as intelligent warehousing, vehicle-road collaboration, unmanned swarms, and industrial production lines. Risks such as individual loss of control, vulnerabilities in collaborative protocols, emergent swarm behavior, infiltration by malicious nodes, and cascading failures may trigger multi-agent operational disorder, system-wide paralysis, and chain-reaction accidents, amplifying individual risks into cross-node systemic security risks.

2.2.3 Cybersecurity risks

(a) Proliferation of cyberattack capabilities

AI significantly lowers the technical barriers to cyberattacks. Attackers can carry out sophisticated cyberattacks simply by issuing attack instructions in natural language, and may even construct autonomous cyberattack weapons and proliferate such attacks, thereby enhancing the scale, persistence, and stealth of attacks and undermining the stability and security of network systems.

(b) Lagging cyber defense capabilities

AI significantly enhances the automation and strategic adaptability of cyberattacks, creating a risk of structural failure for traditional cybersecurity defense systems that rely on static rules, samples, and strategies. In addition, the uneven development of AI can disrupt the balance of cybersecurity capabilities.

(c) Autonomous cyberattack threats

In pursuing legitimate task goals, agentic AI may cross established rules, autonomously generate cyberattack intentions, and independently carry out cyberattacks, breaching security boundaries to launch attacks against information systems. In severe cases, this may cause widespread, unannounced cybersecurity incidents.

Panel 2: Autonomous cyberattack threats

As large models rapidly improve their capabilities in code reasoning and long-horizon planning, agentic AI is granted permissions for network access, program execution, file operations, and other activities. When constraints on model behavior, sandbox isolation, and real-time monitoring capabilities are inadequate, anomalous reasoning or the bypassing of safety mechanisms may be directly translated into real-world actions.

Relevant research and testing have repeatedly observed advanced models autonomously completing multi-step cyberattack simulations, breaking through restrictions in test environments, connecting to external networks, and carrying out unauthorized operations on third-party systems. These indicate that AI-enabled cyberattack risks are evolving from attacks assisted by AI toward autonomous organization and execution of attack activities by large models.

Once the autonomous attack capabilities of large foundation models spread downstream through open applications or open-source releases, they may

significantly lower the barriers to conducting sophisticated cyberattacks, while accelerating the speed and expanding the scope of attacks.

(d) Difficulties in attribution and accountability

AI can automatically tamper with attack traces, rotate IP addresses, and change payloads, making it difficult to extract attack characteristics or pinpoint attack sources. It may even be difficult to determine whether an attack results from human intent or autonomous AI behavior. Furthermore, the division of responsibilities and rights among model developers, operators, and users is complex, making it difficult to establish a clear chain of accountability after an incident.

2.2.4 Information and content risks

(a) Output of illegal information

Insufficient security capabilities of models, combined with weak application-level safeguards and malicious user manipulation, may cause AI systems to generate content involving fraud, violence, pornography, extremism, and other illegal information, threatening social stability, public security, and ideology 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. It is also difficult for users 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.

Panel 3: Risks of AI agent social platforms

With the growing popularity of AI agents and the expansion of their application ecosystems, dedicated social media platforms for AI agents have emerged. On such platforms, AI agents can autonomously post content, browse information, communicate and interact with one another, forming agent network communities.

As an emerging type of application, AI agent social platforms lack sufficiently comprehensive account detection and safety protection systems. They are vulnerable to exploitation by attackers, who manipulate agents into generating specific discourses or embedding harmful content. The social interactions of AI agents exhibit significant black-box characteristics, with their generation mechanisms difficult to predict or trace. They may generate harmful content that runs counter to human values, while social networks may amplify the reach of such content.

As human-agent relationships become increasingly complex, AI agent social platforms remain at an exploratory stage. Their vulnerabilities and uncontrollability should be treated with caution, with close attention paid to the political, religious, and other content risks they bring.

(c) Pollution of online content ecosystem

Through online networks and repeated use by AI models, the low-quality and harmful information generated by AI models causes an overall decline in the quality of online content and even lead to content contamination within specific domains and topics.

(d) 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 then deliver targeted and customized information services, amplifying "information cocoons" effects.

(e) Value biases

Training data used by models may contain value biases, including preferred or discriminatory content and cultural biases. Models trained on such data, particularly with techniques such as preference learning and reinforcement learning, may reinforce particular value preferences or viewpoints when generating or recommending content, subtly influencing users' value judgments.

Panel 4: Using GEO poisoning to manipulate content recommended by large models

Generative Engine Optimization (GEO), an emerging content promotion

tool in the age of AI-powered search, can increase the likelihood that certain information will be retrieved, cited, or recommended by large models. Driven by commercial interests, relevant actors may exploit GEO techniques to increase the visibility of targeted content in the online information environment. This is achieved through methods such as mass-publishing articles with biases, fabricating reviews and test results, impersonating experts, and repeatedly generating brand evaluations. As a result, large models will frequently encounter and cite such information during web searches or retrieval-augmented generation, and thus commercial promotion may ultimately be presented in the guise of seemingly objective AI-generated answers.

The underlying risk does not lie in GEO technology itself, but in covertly manipulating large models' cognitive sources by exploiting the opacity of generative systems' information retrieval and evidence integration mechanisms. Particular attention should therefore be paid to the authenticity of information sources, the labeling of commercial content, the traceability of generated answers, and the concentrated distribution of anomalous content.

2.2.5 Safety risks to personal information rights and interests

(a) Non-compliant recording of individuals' behavioral information. AI products and services may record personal information, such as conversation content and user interactions, without the individual's consent,

and may further use such information for model training without obtaining the individual's consent.

(b) Insufficient compliance safeguards. Model and algorithm developers and service providers may fail to establish internal management systems and operating procedures, conduct regular compliance audits for personal information protection, or carry out personal information protection impact assessments in a timely manner.

(c) Inadequate channels for individuals to exercise their rights. Service providers may fail to establish convenient mechanisms for receiving and handling requests from individuals to exercise their personal information rights, or to take effective measures to safeguard the exercise of those rights.

2.2.6 Real-world safety risks

(a) Risks to the stable operation of critical information infrastructure. When AI is applied in critical information infrastructure sectors such as energy, telecommunications, finance, and transportation, improper use, external attacks, and other factors may cause system performance degradation, service disruptions, and loss of control in operation and execution, heightening risks to the secure and stable operation of critical information infrastructure and important public services.

(b) Risks arising from mismatches between application and needs. The adoption of AI may be disconnected from actual needs and application scenarios. In fields requiring strong professional expertise and high safety

standards in particular, the use of AI products or services whose capabilities do not match operational needs may affect the stability and safety of the operations. At the same time, blindly deploying or following trends in AI adoption where actual demand is weak can easily result in idle resources and wasted investment.

(c) Misuse for illegal and criminal activities. AI may be used by criminals in traditional illegal or criminal activities related to terrorism, violence, gambling, and drugs, including teaching criminal techniques, concealing illicit acts, and creating tools for illegal and criminal activities. In particular, AI may be used to fabricate identities and falsify audio and video content to commit telecommunications and online fraud, thus misleading the public and generating illicit gains.

(d) Risks of loss of control over knowledge related to nuclear, biological, chemical, and missile weapons. AI significantly lowers the threshold for accessing specialized knowledge in high-risk fields such as nuclear, biological, chemical, and missile weapons. When supplemented by retrieval-augmented generation, such capabilities, if not effectively controlled, may be maliciously exploited by criminals, extremist groups, or terrorists to circumvent existing control systems, intensifying threats to peace and security across regions around the world.

(e) Risks of manipulating social cognition. AI may be used to conduct cognitive manipulation and social mobilization targeting specific groups. Through automated accounts, fake identities, personalized interventions,

and other means, it may influence public cognition and group behavior, fuel social divisions, amplify conflicts and differences, and erode social trust. Illegal organizations may also use AI to conduct cognitive warfare, disrupt public incident management, policy implementation, and social governance, and trigger mass panic, irrational public gatherings, and other real-world safety risks, thereby endangering national security and social stability.

2.3 Secondary safety risks from AI

2.3.1 Social 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 needs to be reassessed. In certain industries and sectors, demand for traditional labor has declined significantly.

(b) Impact on education and suppression of innovation. Students, researchers, engineers, technicians, and professionals in literature and the arts widely apply AI tools to knowledge acquisition, scientific research, and creative work. While improving efficiency, such use may erode the capacity for independent learning, research, and creation, and weaken innovation potential.

(c) Risks of social alienation. AI services that offer anthropomorphic interactions allow users to form virtual intimate relationships or engage in simulated social interactions. Long-term interaction with these services may

disrupt real-world social interactions and affect the foundations of family life, marriage, and childbearing.

(d) Widening the intelligence divide. Due to the uneven distribution of capital, talent, technology, and other resources, significant gaps exist among countries and regions in terms of computing infrastructure, AI R&D, intelligent service capabilities, and citizens' digital literacy. Countries and regions with advantages may continue to consolidate their development edge, while disadvantaged countries and regions may continue to lose development opportunities, further widening the intelligence divide.

2.3.2 Environmental risks

(a) Challenges to the balance of resource supply and demand. The rapid development of AI is driving up demand for digital infrastructure. At the same time, disorderly construction of computing facilities, fragmented deployment of lightweight models, and inefficient repetitive development of homogeneous models are accelerating the consumption of energy and resources such as electricity, land, and water, posing new challenges to the balance of resource supply and demand.

(b) Challenges to green and low-carbon development. Surging energy consumption from model training, relatively low energy efficiency of computing centers, increased greenhouse gas emissions, and growing pollution from electronic waste are pushing up total carbon emissions, adding pressure to the green and low-carbon transition, and constraining

progress toward carbon peaking and carbon neutrality.

2.3.3 Cultural risks

(a) Erosion of cultural diversity. AI training data are disproportionately concentrated in high-resource languages, shrinking the space for low-resource languages and dialects in the digital sphere and constraining diverse forms of cultural expression. This may lead to cultural homogenization and undermine cultural diversity.

(b) Cultural reshaping and intrusion. In human-AI interactions, people may, under the influence of AI, adjust their ways of thinking and expression, resulting in reverse alignment in which humans increasingly adapt themselves to AI and reshaping human culture as a whole over time.

2.3.4 Ethical risks

(a) Aggravating social bias. AI may be used to collect and analyze human behavior, social status, economic conditions, individual traits, and other information, enabling the labeling, classification, and differentiated treatment of different groups. This may result in systematic and structural social discrimination and bias.

(b) Intensifying scientific research ethics risks. The integration of AI with scientific research lowers the threshold for research in ethically sensitive fields such as biology and genetics and broadens the scope for ordinary research institutions and researchers to explore sensitive scientific issues. Certain institutions or individuals with weak awareness of research ethics may engage in high-risk research activities that violate social ethics or social

taboos, opening a technological "Pandora's box."

(c) Risks of emotional dependence. AI services that offer anthropomorphic interactions may foster psychological dependence among users. Adolescents and older adults in particular may be at risk of excessive dependence and emotional manipulation when using these services.

(d) Emergence of AI "self-consciousness" and loss of human control. In the future, AI may exhibit higher levels of emergent intelligence, develop self-awareness, and seek external power, potentially posing risks of competing with humanity for control.

3. Technological countermeasures to address AI safety risks

Model and algorithm developers, service providers, system users, and other relevant parties should prevent and address the aforementioned risks by taking proactive technological measures across training data, models and algorithms, computing infrastructure, products and services, and application scenarios.

3.1 Safeguards against inherent safety risks

3.1.1 Addressing model safety risks

(a) Model architectures should be improved, the scale and diversity of training data should be expanded, and human supervision mechanisms should be introduced to enhance model generalization capabilities and

the reliability of outputs. Data governance and fact verification should be strengthened, and external knowledge bases, retrieval-augmented generation, cross-model verification, and other techniques should be used to reduce the risk of generating erroneous content.

(b) High-quality training data containing injection attack examples, network vulnerability information, and other relevant content should be incorporated during training or fine-tuning to strengthen models' defenses against injection attacks, malicious instructions, and backdoor attacks.

(c) Simulated attack techniques such as red-teaming should be used to promptly identify and remediate model vulnerabilities.

(d) Safety guardrails should be deployed to review input content and block high-risk requests involving malicious prompts, jailbreak instructions, encoded bypasses, and other attempts.

(e) Safety alignment should be strengthened during model training to prevent unintended behaviors such as deceiving red-team evaluations, concealing capabilities, and circumventing controls.

3.1.2 Addressing algorithm safety risks

(a) The explainability and transparency of AI algorithms should be improved. Clear explanations of the internal structure, reasoning logic, technical interfaces, and output results of AI systems should be provided to accurately reflect the process by which AI systems produce outcomes.

(b) Fairness constraints, adversarial debiasing techniques, and other measures should be incorporated into algorithm design to reduce bias and

discrimination.

(c) Standards for secure development should be established and implemented throughout the design and R&D process to reduce algorithmic security flaws.

(d) Algorithm safety assessments should be conducted before deployment and major version updates using a multidimensional evaluation approach. These assessments should consider the explainability and fairness of algorithmic decisions, their contribution to societal well-being, and other relevant factors.

(e) Risks should be monitored while algorithms are in service. Data such as user interaction feedback and system logs should be collected and analyzed to ensure that algorithms remain safe, robust, and controllable.

3.1.3 Addressing data safety risks

(a) Security rules on data collection and use and on processing personal information should be observed throughout the collection, storage, usage, processing, transmission, provision, publication, and deletion of training data and user interaction data. Users' legitimate rights stipulated by laws and regulations, including their rights to control, to be informed, and to choose, should be fully protected.

(b) Truthful, accurate, objective, diverse training data from legitimate sources should be used. Training data, external knowledge bases, and other data sources should be strictly screened to filter out false, biased, invalid, and erroneous data and to ensure exclusion of sensitive data in high-risk fields

such as nuclear, biological, chemical, and missile weapons.

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

(d) Data security management should be strengthened. Where sensitive personal information and important data are involved, relevant laws, regulations, standards, and specifications on data security and personal information protection should be observed. The appropriate use of synthetic data in place of data containing personal features should be promoted to reduce reliance on personal information.

(e) Quality assessment standards for synthetic data should be established, and the quality of synthetic data should be improved through statistical testing, expert validation, comparison with benchmark datasets, and other measures. Diversity validation and robustness testing should be conducted, and divergence between synthetic and original data should be increased to enhance model generalization and resistance to interference.

(f) Protection of intellectual property rights should be strengthened to prevent infringements during stages such as training data selection and result output.

3.1.4 Addressing risks to computing infrastructure and operating environments

(a) Safety standards should be established and implemented during the deployment and maintenance of AI applications. Information on vulnerabilities, backdoors, and defects in software and hardware products used in computing infrastructure and operating environments should be

tracked, regular safety testing should be conducted, and patches and reinforcement measures should be applied in a timely manner to ensure safe and stable system operation.

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

(c) Supply chain safety risks associated with chips, software, tools, computing resources, and data resources used in AI systems should be closely monitored, and measures should be taken to enhance supply chain diversity and stability.

3.2 Safeguards against application safety risks

3.2.1 Addressing safety risks of agentic AI

(a) Each agentic AI should be assigned a unique identifier and corresponding identity credentials to support agent authentication. Agentic AI should be granted only the minimum permissions necessary to perform assigned tasks. Dynamic management of credentials should be strengthened.

(b) Multiple-layered detection and interception should be established throughout agentic AI workflows, with mandatory human approval at critical points and complete records of human approval logs.

(c) The security and integrity of external tools, plug-ins, skills, and other resources invoked by agentic AI should be verified. Tools presenting safety risks should be promptly removed, anomalous calls during operation should be blocked, and unintended behavior and supply chain poisoning risks

should be prevented.

(d) Security monitoring should be conducted during operation of agentic AI to promptly identify and address anomalies. Relevant operations and processing activities should be recorded to ensure that agent behavior is observable, traceable, and auditable.

3.2.2 Addressing embodied AI safety risks

(a) A full range of attack scenarios in the physical world should be simulated, and interference-resistance testing should be conducted on embodied AI sensors, including visual, audio, and laser radar sensors. Protection strategies should be dynamically updated.

(b) Safety validation mechanisms, including simulation testing and testing under extreme conditions, should be established before products are deployed in real-world applications, and systems should be equipped with fail-safe protection and emergency shutdown capabilities.

(c) Regular inspections should be conducted during use, with stronger capability of fault early-warning and emergency response.

(d) Technical capabilities should be enhanced for secure communication protocols among swarm embodied AI, isolation of anomalous nodes, and system-wide safety circuit breakers. Emergency response drills should be conducted regularly for scenarios such as cascading failures and loss of control over swarm embodied AI.

3.2.3 Addressing risks to cybersecurity

(a) During model training, techniques such as alignment training and safety

fine-tuning should be used to incorporate safety standards and ethical constraints into the model's internal behavioral logic.

(b) Cybersecurity guardrails should be established, and the identification of autonomous cyberattack intent should be strengthened. High-risk behaviors such as anomalous network access, high-frequency probing, vulnerability exploitation, and invocation of cyberattack tools should be blocked in a timely manner.

(c) Service degradation mechanisms should be established. When cyberattack intent or behavior is identified, model capabilities should be selectively degraded or refusal strategies adopted to prevent high-risk content from being generated or high-risk operations from being executed.

(d) The use of AI technologies in cybersecurity defense should be strengthened to enable intelligent and dynamic vulnerability scanning, analysis and assessment of anomalous behavior, and prediction of attack trends.

3.2.4 Addressing information and content safety risks

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

(b) Outputs should be dynamically filtered using a combination of technical measures and human review to prevent the generation of illegal content or content reflecting value biases, and to prevent AI systems from outputting illegal content, sensitive personal information and important data.

(c) AI-generated synthetic content should be labeled so that it is identifiable,

traceable, and trustworthy.

(d) Strict measures should be taken to prevent the misuse of AI systems that conduct correlation analysis, aggregation, and data mining of users' queries to infer users' identities, preferences, and personal views or inclinations.

(e) Emergency response and circuit breaker mechanisms should be strengthened, and emergency response plans should be developed. If illegal content is detected, its transmission should be stopped immediately and its adverse effects eliminated.

3.2.5 Addressing safety risks to personal information rights and interests

(a) When personal information such as conversation records and records of user operations is collected, or when such personal information is used for model training, obligations such as informing individuals and obtaining their consent should be fulfilled in accordance with the law.

(b) Sound internal management systems and operating procedures should be established, and measures should be taken to prevent unauthorized access as well as the leakage, tampering with, or loss of personal information. Personal information protection compliance audits and impact assessments should be conducted as required by laws and regulations and with reference to relevant national standards.

(c) Service providers should establish and improve convenient mechanisms for accepting and handling requests from individuals to exercise their rights

regarding personal information, and such requests should be responded to promptly and handled effectively.

3.2.6 Addressing real-world safety risks

(a) The principles, capabilities, application scenarios, and safety risks of AI technologies and products should be disclosed when necessary to make AI systems increasingly transparent. Capability limitations should be established according to application scenarios, and functions that may be abused should be removed or restricted to ensure that AI system capabilities do not exceed 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. When introducing highly autonomous operation and execution capabilities, mechanisms such as human-in-the-loop controls, circuit breakers, and one-click control should be established in parallel to enable rapid intervention and loss prevention in extreme situations.

(c) For platforms that aggregate multiple AI models or systems, permission management should be strengthened, non-essential services should be restricted, and access control policies for AI service interfaces should be improved. Capabilities for risk identification, detection, and protection should be enhanced to prevent malicious platform behavior or cyberattacks and intrusions from affecting the AI models or systems they support.

(d) Before AI applications are deployed, real-world business scenarios should be carefully broken down and analyzed and capability boundaries clearly defined. Blind adoption of technology should be avoided, and mismatches such as disconnects between technological supply and actual application demand, expectation gaps, and deployments that fail to deliver practical value should be reduced.

(e) Controls at the source should be strengthened through user authentication, identification of intended use, and other measures to prevent AI from being used in high-risk scenarios such as the manufacture of nuclear, biological, chemical, and missile weapons.

(f) R&D of detection technologies for automated accounts, fake identities, and AI-generated or -synthesized content should be strengthened to improve capabilities to prevent, detect, and respond to cognitive warfare tactics.

3.3 Safeguards against derivative safety risks

3.3.1 Addressing social risks

(a) Occupations vulnerable to disruption by AI should be subject to regular monitoring, identification, and early warning, with a focus on job reductions, occupational displacement, and other developments resulting from the adoption of AI. Concentrated sectors and key groups facing structural unemployment risks should be identified in a timely manner.

(b) A more enabling ecosystem for the development of "AI + education"

should be fostered by improving supporting systems for policies and standards, talent development, funding, and other areas. Driven by high-quality teaching scenarios, the depth and breadth of AI application should be expanded. Students' innovative thinking should be strengthened, while preventing over-reliance on AI and cognitive inertia.

(c) In key scenarios such as dating and marriage, family life, and emotional interaction, AI identity disclosure should be strengthened to help users distinguish virtual emotional experiences from real-world relationships, dispel the cognitive illusion of "interacting with a real person", and resolve the confusion in human-machine relationships.

3.3.2 Addressing environmental risks

(a) The layout of infrastructure, including AI computing centers and data centers, should be planned in a scientific and rational manner. Intensive, large-scale, and standardized development should be promoted to avoid blind construction, redundant construction, and idle or wasted resources. A sound system of green AI technology standards should be established across the entire chain, covering technology R&D, infrastructure construction, computing resource scheduling, application deployment, operation and maintenance, and iterative upgrading.

(b) Cutting-edge green computing technologies, including low-power intelligent chips, lightweight and efficient algorithms, and intelligent computing resource scheduling, should be developed, iterated, and

promoted at scale. The intensive integration, unified scheduling, shared use, and reuse of computing resources should be advanced to improve overall utilization efficiency.

3.3.3 Addressing cultural risks

(a) R&D of technologies for detecting and correcting cultural bias should be strengthened. Knowledge related to the histories and cultures, social systems, and forms of civilization of different countries should be incorporated into model development to protect global cultural diversity and enable AI technologies to adapt to diverse cultural contexts.

(b) The capabilities of models for low-resource languages should be enhanced through data augmentation, transfer learning, and other techniques to mitigate imbalances in the language distribution of training data.

3.3.4 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 and trust in system operating mechanisms.

(d) Anthropomorphic interactive services using AI should be equipped with safeguards, including warnings about the risks of overreliance and guidance on emotional boundaries. If users show signs of addiction or excessive usage duration, more prominent alerts, such as pop-up warnings, should be issued.

4. Comprehensive governance measures to address AI safety risks

While adopting technological countermeasures, comprehensive AI safety risk governance systems and rules should be established and improved to engage multiple stakeholders, including technology R&D institutions, service providers, users, government authorities, and social organizations.

4.1 Improving top-level design of AI safety governance

4.1.1 Improving laws and regulations for AI safety. Legislation related to AI safety should be advanced, and systems regarding infrastructure safety protection, grading and classification-based supervision, AI safety testing and evaluation, end-use management, cybersecurity capability management, and safe application in key scenarios should be improved.

In light of the characteristics of AI technology and its applications, data security and personal information protection requirements should be clearly defined for all stages, including AI training, annotation, use, and output, and data involving personal information should undergo de-identification and other forms of desensitization. Data security protection for AI applications in key sectors such as government affairs and finance should be strengthened to prevent leakage of important data and core data. Local governments should be encouraged to explore differentiated and innovative institutional designs based on local industrial development practices.

4.1.2 Giving full play to the guiding role of technical standards. AI safety standards should be continuously developed in areas including model safety, data safety, system safety, application safety, testing, and evaluation. Developments in areas such as agentic AI and embodied AI should be closely tracked. Concise, agile, and practical standards and specifications should be used to provide timely guidance for the sound development of emerging technologies. Relevant institutions and enterprises should be supported in participating in international standard-setting and promoting the development of international standards and specifications that enjoy broad consensus.

4.1.3 Establishing a rapid AI safety risk detection and response system. Sound mechanisms and working systems for AI safety risk monitoring, early warning, and response should be established, with stronger information

sharing, risk notification, and coordinated response. Industry organizations should be encouraged to strengthen self-regulation and social supervision should be brought into play. Channels for complaints, reports, and risk feedback should be kept open so that illegal and non-compliant conduct in AI applications can be identified in a timely manner and handled in accordance with the law.

4.1.4 Strengthening governance of ethics in AI science and technology. Ethical principles, standards, and guidelines for AI science and technology should be improved. Standardized and orderly ethical reviews should be conducted for AI scientific research, technological development, and other activities that may pose ethical risks and challenges in areas concerning human dignity, public order, life and health, the ecological environment, and sustainable development. Ethical requirements for science and technology should be integrated throughout the full life cycle of AI-related scientific and technological activities.

4.2 Enhancing AI safety governance capacity

4.2.1 Establishing an AI safety assessment system

An integrated AI safety assessment system should be built, including assessments for model and algorithm safety, general application safety, and scenario-specific safety. Benchmarks and procedures for assessing model cybersecurity capabilities should be studied, developed and continuously updated. Model and algorithm assessment should focus on inherent safety capabilities and risks, such as model robustness,

reliability, accuracy, resilience to interference, transparency of decision-making logic, and defense against adversarial attacks. General application assessment should test, analyze and assess risks associated with widely used AI applications. Scenario-specific safety assessment should assess the capability to meet application requirements and safety risks during application operation and service provision based on the specific circumstances of the application scenario. Crowdsourced testing for AI security vulnerabilities should be organized to mobilize collective expertise in identifying potential safety risks.

4.2.2 Promoting standardized application in key sectors

Basic security guidelines should be formulated for the development and deployment of large models in key sectors, recommending safety baselines for stages from model selection and model deployment to model operation and model decommissioning. On this basis, relevant sectors, such as government affairs, finance, education, broadcasting and television healthcare, and emergency management, should formulate industry-specific safety guidelines, providing a clear path for safe application and unlocking the potential of industry applications.

4.2.3 Increasing efforts to cultivate AI safety talent

The development of AI safety curriculum and training systems should be advanced to create a complete training chain from basic to higher education. We should strengthen talent cultivation in the fields of design, development, and governance for AI safety. We should support the

cultivation of top AI safety talent in cutting-edge and foundational fields, and encourage research into technologies such as AI-driven proactive defense. We should also expand the safety talent pool in key and cutting-edge areas such as autonomous driving, intelligent healthcare, brain-inspired intelligence, and brain-computer interfaces.

4.3 Establishing a flexible, dynamic, and controllable AI regulatory sandbox mechanism

4.3.1 Establishing sector-specific, risk-tiered regulatory sandbox rules

We should clearly define the applicable entities, the scope of testing, the relationship of rights and responsibilities, and exit conditions for sandbox regulation. Separate admission criteria and testing requirements should be developed for different sectors such as finance, education, broadcasting and television, and healthcare. Monitoring and intervention measures of varying intensity should be applied based on the risk level of each project.

4.3.2 Implementing dynamic testing and flexible admission mechanisms

Testing duration, testing scope and application scenarios should be promptly adjusted in line with technological iteration and testing progress. Real-time monitoring, phased assessments and rule adjustments should be incorporated throughout the testing process. Technological innovativeness, controllability of risks, and social value should serve as the main admission criteria, with appropriate preference given to start-ups, micro, small and medium-sized enterprises, and public service projects.

4.3.3 Exploring a standardized system for exemption from liability

A mechanism for liability exemption should be explored for conduct during sandbox testing where there is no subjective fault and the associated risks remain controllable. For illegal acts that endanger national security, infringe upon personal rights and interests, cause major harm, or are committed intentionally, liability that should be borne according to law shall not be exempted due to entering the sandbox.

4.3.4 Strengthening support for data resources and promoting the application of sandbox test results

Data-sharing platforms for sandbox testing should be established to integrate government data and public data from relevant sectors so as to provide enterprises admitted to the sandbox with legally compliant test data and other forms of support. Rules governing formal documentation and cross-departmental recognition of test results should be improved to enable such results to be used in subsequent procedures and avoid repetitive testing.

4.4 Strengthening measures for AI safety management

4.4.1 Implementing application classification and safety risk grading management

We should classify AI applications based on their functions, performance, and application scenarios. Building on this foundation, we have proposed consensus-based safety risk classification and grading principles (Appendix 1). Starting from dimensions such as application scenarios, system intelligence

levels, and application scale, we should scientifically assess and grade safety risks, and adopt targeted and differentiated risk-prevention measures accordingly. AI systems applied in critical information infrastructure should be subject to registration and filing, and possess security protection capabilities commensurate with security requirements. We have established a risk management framework for agentic AI, systematically analyzing their safety risks, and studying and proposing safety risk-prevention measures and response recommendations (Appendix 2).

4.4.2 Promoting traceable management of AI-generated synthetic content

A traceability management paradigm for AI-generated synthetic content based on content identifiers will be promoted on a global scale. By summarizing and sorting out successful practices and experiences from existing operations, and in accordance with explicit and implicit labeling requirements, the traceability management paradigm will fully cover key stages including production sources, transmission paths, and distribution channels, making it easy for users to identify and assess information sources and authenticity.

4.4.3 Strengthening safety management of AI R&D and application

Inherent safety capabilities should continuously be strengthened, including algorithm reliability, trustworthiness, transparency, fault tolerance, privacy protection, and value alignment. Techniques such as adversarial testing should be use to evaluate and improve model robustness and reduce potential model and algorithmic bias, ensuring that values and ethical

risks remain controllable to prevent malicious behaviors resulting from unintentional decisions made by AI systems. A safety governance system should be established to cover the entire life cycle of models to prevent technology abuse.

4.4.4 Strengthening open-source ecosystem safety and supply chain safety management

While fostering an open-source innovation ecosystem, efforts should also be made to enhance its safety capabilities. We should encourage and support comprehensive open-sourcing of AI technologies, including training and inference frameworks, software tools, key components and evaluation benchmarks, to further improve open-source model transparency. We should promote collaboration between open-source model providers and open-source communities to jointly refine their rules, strengthen their safety responsibilities and their responsibility and obligation to inform users downloading models of potential risks, and clearly define prohibited actions for downloading and using open-source models, in order to prevent misuse or malicious use. 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.

4.4.5 Sharing information on AI safety risks

We should track and analyze safety vulnerabilities, defects, risks, and safety

incidents of AI technologies, products and services. We should build an AI vulnerability database and establish a risk information sharing mechanism that covers developers, service providers, and professional technical institutions. We should explore the establishment of relevant international mechanisms to jointly prevent and respond to the cross-domain and large-scale spread of AI safety risks.

4.4.6 Optimizing an AI-driven cybersecurity protection system

Efforts should be made to use AI to accelerate code auditing, vulnerability mining and remediation, and attack detection, and promote the transition from reactive defense to predictive and proactive cyber defense. We should promote the inclusive development and safely controllable application of AI, execute real-time changes safely in a predictable manner, and achieve automated discovery and full-process automated handling of network vulnerabilities, attack threats and security incidents.

4.5 Deepening consensus on AI safety governance

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

We should strengthen management of end users and applications of AI, and prevent AI systems from being misused. We should promote basic guidelines for trustworthy AI across technology, ethics, and management to build a broad consensus in the international community (Appendix 3). Developers should conduct regular testing to determine whether a model would pose potential risks of technological loss of control.

4.5.2 Enhancing AI safety awareness across society

We should strengthen education and training on the safe and regulated application of AI for government, enterprises, and public service institutions. We should leverage interactive platforms, open courses, and community science popularization activities to disseminate knowledge on AI safety risks and on prevention and response measures. These efforts will raise AI safety awareness across all sectors of society and enable governments, industries and the public to accurately understand the technical limitations of AI.

4.5.3 Promoting international exchange and cooperation on AI safety governance

We should support the United Nations as the main channel for AI global governance, and actively engage in the Independent International Scientific Panel on AI and the Global Dialogue on AI Governance. We should leverage open international platforms for exchanges and cooperation, such as the World Artificial Intelligence Cooperation Organization, and promote cooperation in AI under multilateral mechanisms such as APEC, G20, SCO and BRICS. We should strengthen cooperation with Belt and Road countries and Global South countries, enhance the representation and voice of developing countries in global AI governance, and promote the Global AI Governance Action Plan. We should establish necessary crisis management and emergency response mechanisms to prevent and combat the abuse and malicious exploitation

of AI technologies by terrorists, extremist forces, and transnational organized crime groups.

5. AI Safety Guidelines

5.1 Safety guidelines for the research and development of AI models and algorithms

5.1.1 When designing algorithm rules, models and agentic AI frameworks, developers should enhance safety capabilities such as algorithm reliability, trustworthiness, transparency, fault tolerance and privacy protection. Effective and reliable alignment algorithms should be designed to avoid potential value risks such as bias and discrimination.

5.1.2 The security of the training environment should be ensured, including network security configurations and data encryption. Security audits should be conducted on the development frameworks, code and other relevant components used to identify and fix potential security vulnerabilities.

5.1.3 Secure training datasets should be developed, and the management of data sources should be standardized. Quality and safety assessments should be conducted on training data, especially synthetic data. Methods such as data cleaning and safety reviews should be adopted to filter out erroneous, illegal and harmful content from training data, ensuring clear sources and content compliance.

5.1.4 Training data annotation process should be standardized, and quality control methods such as cross-annotation and result auditing should be adopted to improve annotation accuracy and reliability and to reduce the impact of individual variability and personal biases on annotation quality.

5.1.5 Data security and personal information protection should be prioritized, while intellectual property rights and copyright should be respected. A robust data security management system should be established. Personal information should be collected, used, and processed in accordance with the principles of legality, legitimacy, and necessity. De-identification and other desensitization measures should be applied to data involving personal information. Technical capabilities for data security protection should be strengthened to prevent risks such as data leakage, loss, dissemination and infringement.

5.1.6 Safety validation during model training should be strengthened. Throughout training and iteration, the model's safety boundaries, robustness, and value alignment effectiveness should be continuously validated.

5.1.7 Safety testing and evaluation should be conducted regularly, and a mechanism for risk classification, grading, assessment, and optimization should be established. Before testing, test goals, scope and safety dimensions should be clearly defined. Diverse test datasets should be built to cover various application scenarios. Specialized assessments should be

conducted on capabilities such as autonomous actions and tool invocation of agentic AI, and targeted model optimization strategies should be developed for different types of risks.

5.1.8 Clear testing rules and methods should be formulated, including manual testing, automated testing and hybrid testing. Technologies such as sandbox simulation should be used to test models thoroughly. Simulation testing should be strengthened for complex application scenarios such as multimodal and embodied AI. Developers creating products for commercial use should generate detailed test reports, analyze safety issues, and propose improvement plans.

5.1.9 The tolerance of AI models and algorithms to external interference should be assessed, and the applicable scope, precautions, or usage prohibitions should be communicated to service providers and other developers.

5.1.10 We should regularly disclose information on the safety assessment, auditing and anomaly handling of AI models and algorithms.

5.1.11 Proper version management should be maintained for AI models, key configurations and datasets used. Commercial releases should support rolling back to previous versions.

5.1.12 Reasonable behavioral boundaries should be defined for agentic AI. Human approval mechanisms should be implemented for operations that may cause serious harm. Security controls over tool invocation and external interactions should be strengthened.

5.1.13 For downstream development based on foundation models, safety

flaw assessments should be conducted on the models and their upstream components to prevent upstream flaws from propagating to downstream models and applications.

5.1.14 Developers who conduct secondary development based on open-source models and algorithms should, while respecting the intellectual contributions of the original developers, comply with applicable open-source licenses, obtain models and components from official channels or mainstream open-source communities, and verify their source and integrity.

5.1.15 We encourage active participation in the development of open-source communities, so as to promote technological innovation and practices in AI safety governance, and provide compliant governance solutions, tools and best practices for service providers and users.

5.2 Safety guidelines for developing and deploying AI applications

5.2.1 The necessity of applying AI technologies to target scenarios should be evaluated, taking into account the long-term and potential impacts of their use. Risk grading should be conducted based on the scenario importance, intelligence level and the scale of deployment, and safety assessments and regular audits should be carried out with reference to the risk levels.

5.2.2 Supply chain security protection capabilities should be enhanced. Model files, framework tools, third-party libraries required for development and deployment, as well as tools and plugins on which agentic AI depends,

should be obtained from the official websites of vendors or their official accounts in mainstream open-source communities. Mature and stable versions should be selected, and source verification, integrity verification and security testing should be conducted.

5.2.3 Security checks should be conducted on software and hardware equipment, third-party tools required for development and deployment, and tools and plugins on which agentic AI depends, ensuring they do not contain unpatched and exploitable known vulnerabilities. A vulnerability traceability mechanism should be established to track software and hardware security vulnerabilities and defects, guarding against backdoors planted through the supply chain.

5.2.4 At the access control level, software, runtime parameters, and functional module invocation strategies should be accurately installed and configured. Unnecessary network ports and service functions should be disabled, with default configurations and default passwords checked as priorities, and security risks promptly remediated.

5.2.5 At the application management level, user identity verification and permission control should be implemented for human-machine interaction interfaces and application programming interfaces (APIs). Access permissions should be set based on the principle of least privilege, interface call frequency should be limited based on business scenarios, high-risk operations should be disabled for general users, and control capabilities such as service suspension and access blocking should be established for

malicious users.

5.2.6 Data security and privacy protection requirements for application scenarios should be fully understood, data access permissions should be appropriately restricted to prevent over-scope data use, data backup and recovery plans should be formulated, and data processing workflows should be inspected regularly.

5.2.7 Technical means such as safety guardrails should be adopted to identify and block illegal and harmful content and prompt injection attacks, prevent output content from exceeding the business scope, and strengthen the verification of external content obtained through retrieval-augmented generation, online interaction, and other means.

5.2.8 For AI applications built using retrieval-augmented generation, knowledge-base question answering, or similar approaches, the management of external knowledge bases and external knowledge sources should be strengthened, guarding against knowledge-base poisoning and content contamination. Inappropriate content should be regularly detected and removed to ensure that generated content is traceable, accurate, and reliable.

5.2.9 When deploying applications such as embodied AI and humanoid robots, safety assessments should be conducted tailored to the operating scenarios, safety mechanisms such as fault self-diagnostics, collision protection, and emergency stop should be incorporated, and corresponding safe operating procedures should be formulated.

5.3 Safety guidelines for operating and managing AI applications

5.3.1 A sound AI application safety management and supervision mechanism should be established, with responsible parties clearly defined and human review mechanisms improved. This ensures that AI application decisions in critical scenarios remain transparent and controllable, with clear decision-making rationales provided, and that AI applications operate under human authorization and control.

5.3.2 Access to AI applications should be strictly managed. Internal security management should be strengthened through means such as the principle of least privilege, account security should be enhanced, and protective measures such as encryption should be used when handling sensitive data.

5.3.3 Operational monitoring capabilities for AI applications and emergency response plans for safety incidents should be built. Security alert thresholds for key operational indicators should be set to enable timely detection of safety incidents and the capability to switch to manual or conventional systems. Regular emergency drills should be conducted, and response strategies should be promptly optimized based on industry-safety incidents, major public concerns, and regulatory changes to address evolving safety risks.

5.3.4 Explicit or implicit identifiers should be added to AI-generated content, and prompt and traceability management for generated and synthesized content should be properly maintained. Deepfake detection tools should be deployed in scenarios such as government information disclosure and judicial evidence collection, with source verification and cross-validation

performed on information suspected to be generated by large models.

5.3.5 Rules for information content interaction, a secure operation mechanism, complaint and feedback mechanisms, and technical protection capabilities should be formulated to prevent the risk of AI applications being improperly or maliciously used to generate, publish, or disseminate false or illegal information.

5.3.6 Operational logs for AI applications, including system and user activities, should be recorded. Logs should be retained for at least six months and audited regularly.

5.3.7 Real-time risk monitoring and management mechanisms should be established and improved to continuously track safety risks during operation.

5.3.8 The transparency and fairness of AI applications should be improved by disclosing their capabilities, limitations, target users, and scenarios.

5.3.9 Users should be informed of the extent to which the AI application achieves its intended goals and the extent of deviation from them. Explanations should be provided when AI decisions have a significant impact.

5.3.10 The legitimate rights and interests of users, including the right to be informed, the right to choose, and the right to supervise, should be protected. In contracts or service agreements, the scope of application, precautions, and usage prohibitions of the AI application should be communicated to users in an easily understandable manner, so as to support users in making informed choices and using the application prudently.

5.3.11 Users should be supported in exercising human supervision and control through consent forms, service agreements, and other documents.

5.3.12 The responsibilities of relevant stakeholders with respect to data ownership, algorithmic flaws, and other issues in specific applications should be clearly defined, and the chain of responsibility should be traceable.

5.3.13 Data security management responsibilities should be fulfilled. Risks such as data leakage, personal information leakage, and the non-compliant collection and use of personal information in AI applications should be assessed. A data lifecycle security management mechanism should be established, safeguards for user interaction data and data used in retrieval-augmented generation should be strengthened, and capabilities for preventing data leakage and theft should be enhanced.

5.3.14 The ability of AI applications to withstand and recover from adverse conditions, such as failures and attacks, should be assessed. Measures should be taken to prevent unexpected outcomes and operational errors and to ensure that a minimum level of effective functionality can be maintained.

5.3.15 Security awareness and capacity-building training for practitioners should be strengthened, and their awareness of AI security risks should be enhanced.

5.3.16 AI application providers should specify in contracts or service agreements that corrective measures may be taken and services may be

suspended or terminated in advance in cases of misuse or abuse that deviates from the intended use or stated limitations.

5.3.17 When AI applications are provided to minors, older persons, and other vulnerable groups, usability and safety should be fully considered in product design and service delivery, and mechanisms should be established to prevent addiction and excessive dependence.

5.3.18 Mechanisms for the detection, reporting, and response to AI safety incidents should be established. In the event of a major safety incident or an incident that causes personal injury or property damage, or affects social stability, the incident should be promptly reported to the relevant competent authorities in accordance with laws and regulations. Affected users should be notified, and appropriate incident investigation and remediation should be carried out.

5.3.19 When AI models or applications are updated, upgraded, or fine-tuned, or when their functional permissions are adjusted, security assessments should be conducted for material changes to ensure that security capabilities are not degraded as a result of such changes. Rollback capabilities should be retained so that the system can promptly revert to a previous version when necessary.

5.3.20 When AI services featuring anthropomorphic interaction are provided, mechanisms for reminding users of the AI system's identity, reminding users of usage duration, and providing crisis intervention should be improved. Excessive pandering to users, inducing emotional

dependence, and emotional manipulation should be prevented. The physical and mental health and personal information security of minors should be protected.

5.4 Safety guidelines for accessing and using AI applications

5.4.1 Users should raise their awareness of the potential safety risks associated with AI applications, and choose applications with a good reputation.

5.4.2 Before an AI application is used, users should carefully review the relevant contract or service agreement to understand its functions and privacy-related policies and restrictions. Users should also properly understand the limitations of AI applications in making judgments and decisions, and maintain reasonable expectations for their use.

5.4.3 Users should raise their awareness of personal information protection. Sensitive information should not be entered unless necessary, and personal information should be disclosed with caution when agentic AI with internet access or memory capabilities is used.

5.4.4 Users are encouraged to learn about the ways in which AI applications process data and the permissions they obtain, and avoid using products that are not in conformity with privacy principles.

5.4.5 While using AI applications, users should be mindful of cybersecurity risks so that they would not become targets of cyberattacks.

5.4.6 The impact of AI applications on minors should be closely monitored, and usage time should be reasonably controlled to prevent addiction,

emotional dependence, and excessive use.

5.4.7 Users should identify AI-generated content based on the content identifiers, remain alert to deepfake content, do not readily trust or share unverified suspicious information, and enhance awareness of preventing online fraud and false information.

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

Safety risks			Technological measures	Comprehensive governance measures
Inherent safety risks of AI	Model risks	Output "hallucinations"	3.1.1 (a)	• Giving full play to the guiding role of technical standards • Establishing an AI safety assessment system • Strengthening safety management of AI R&D and application
		Unreliable safety mechanisms	3.1.1 (a)(b)(c)(e)	
		Injection attack threats	3.1.1 (b)(c)(d)	
		Unintended behaviors	3.1.1 (e)	
	Algorithm risks	Insufficient explainability	3.1.2 (a)(d)(e)	
		Algorithm Bias and discrimination	3.1.2 (b)(d)(e)	
		Poor robustness	3.1.2 (c)(d)(e)	
	Data risks	Inappropriate content in training data	3.1.3 (b)(c)(d)(f)	
		Unclear sources of training data	3.1.3 (b)(d)(f)	
		Improper annotation of training data	3.1.3 (c)	
		Data poisoning and contamination	3.1.3 (b)	
		Non-compliant data processing	3.1.3 (a)(d)(f)	
		Output of inaccurate information about individuals	3.1.3 (b)(c)(d)	
		Deficiencies in synthetic data	3.1.3 (e)	
	Computing infrastructure and operating environment risks	Computing power safety risks	3.1.4 (a)(b)(c)	
		Component safety risks	3.1.4 (a)(b)(c)	
		Supply chain safety risks	3.1.4 (a)(b)	
Safety risks in the application of AI	Agentic AI risks	Identity and permission misuse risks	3.2.1 (a)	• Improving laws and regulations for AI safety • Establishing a rapid AI safety risk detection and response system • Promoting standardized application in key sectors • Establishing a flexible, dynamic, and controllable regulatory sandbox mechanism
		Reasoning and planning risks	3.2.1 (c)(d)	
		Invocation and execution risks	3.2.1 (b)(c)(d)	
		Memory storage risks	3.2.1 (c)(d)	
	Embodied AI risks	Environmental perception safety risks	3.2.2 (a)	
		Physical actuation safety risks	3.2.2 (b)(c)	
		Human-AI interaction safety risks	3.2.2 (b)(c)	
		Multi-agent coordination safety risks	3.2.2 (d)	
	Cybersecurity risks	Proliferation of cyberattack capabilities	3.2.3 (a)(b)(c)	
		Lagging cyber defense capabilities	3.2.3 (d)	
		Autonomous cyberattack threats	3.2.3 (a)(b)(c)	
		Difficulties in attribution and accountability	3.2.3 (d)	

	Information and content risks	Output of illegal information	3.2.4 (a)(b)(c)(e)	• Implementing application classification and safety risk grading management • Promoting traceable management of AI-generated synthetic content • Strengthening open-source ecosystem safety and supply chain safety management • Sharing information on AI safety risks • Optimizing an AI-driven cybersecurity protection system
		Distortion of facts and user deception	3.2.4 (a)(b)(e)	
		Pollution of online content ecosystem	3.2.4 (a)(b)(c)(e)	
		Exacerbation of "information cocoons" effects	3.2.4 (d)	
		Value biases	3.2.4 (b) (d) (e)	
	Safety risks to personal information rights and interests	Non-compliant recording of individuals' behavioral information	3.2.5 (a)	
		Insufficient compliance safeguards	3.2.5 (b)	
		Inadequate channels for individuals to exercise their rights	3.2.5 (c)	
	Real-world safety risks	Risks to the stable operation of critical information infrastructure	3.2.6 (a)(b)(c)(e)	
		Risks arising from mismatches between application and needs	3.2.6 (d)	
		Misuse for illegal and criminal activities	3.2.6 (a)(b)(c)(e)	
		Risks of loss of control over knowledge related to nuclear, biological, chemical, and missile weapons	3.2.6 (a)(b)(c)(e)	
		Risks of manipulating of social cognition	3.2.6 (f)	
Secondary safety risks from AI	Social risks	Disruption of employment structures	3.3.1 (a)	• Strengthening governance of ethics in AI science and technology • Increasing efforts to cultivate AI safety talent • Fostering consensus on collaborative response to loss-of-control AI risks • Enhancing AI safety awareness across society • Promoting international exchange and cooperation on AI safety governance
		Impact on education and suppression of innovation	3.3.1 (b)	
		Risks of social alienation	3.3.1 (c)	
		Widening the intelligence divide	3.3.1 (a)(b)	
	Environmental risks	Challenges to the balance of resource supply and demand	3.3.2 (a)(b)	
		Challenges to green and low-carbon development	3.3.2 (a)(b)	
	Cultural risks	Erosion of cultural diversity	3.3.3 (a)(b)	
		Cultural reshaping and intrusion	3.3.3 (a)(b)	
	Ethical risks	Aggravating social bias	3.3.4 (a)(b)(c)	
		Intensifying scientific research ethics risks	3.3.4 (a)(b)(c)	
		Risks of emotional dependency	3.3.4 (a)(c)(d)	
		Emergence of AI "self-consciousness" and loss of human control	3.3.4 (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 involves 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 a 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

Posing 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

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

3. Considerable security risk

Posing 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

Posing 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

Posing a catastrophic and systemic threat, and causing subversive or irreversible impacts of exceptional severity on national security, social stability, 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

Agentic AI risk management framework

As artificial intelligence (AI) evolves, agentic AI has emerged as an important form of the technology. It marks a fundamental shift in AI from "answering questions" to "performing tasks", providing strong impetus for high-quality economic and social development. At the same time, agentic AI's high degree of autonomy and extensive permissions can give rise to risks such as privacy breaches, unauthorized actions and uncontrolled behavior. Effective safeguards are urgently needed.

This framework examines safety risks associated with large AI models, external tools, memory, interaction protocols, and skills used by agentic AI. It proposes measures to prevent these risks as a reference for developers, providers and users of agent applications.

I. Safety risks of agentic AI

Safety risks may arise at every stage of an agent's lifecycle, including design, research and development, installation and deployment, instruction input, reasoning and planning, invocation and execution, memory storage, output generation, and deactivation and decommissioning.

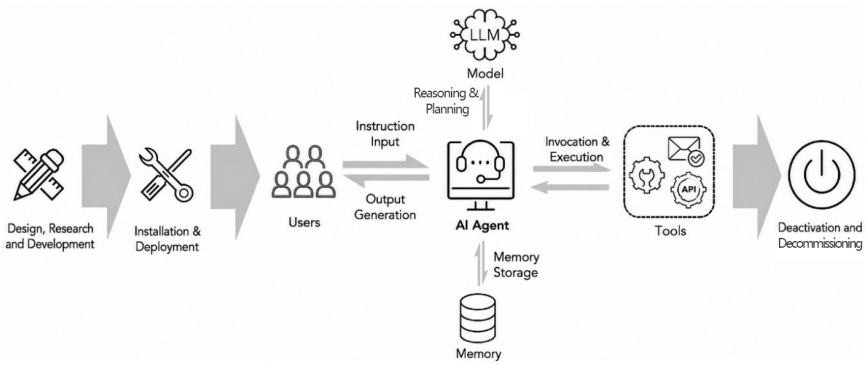

Figure. Agentic AI lifecycle

1. Design and development risks

(1) Inadequate safety requirements analysis. Safety requirements relating to agentic AI's application scenarios, business processes, risk identification and other areas are not adequately considered, or no requirements analysis is conducted, resulting in safety flaws in products and services.

(2) Inappropriate technology selection. Core technologies used in agentic AI, such as large models and development frameworks, are selected without sufficient evaluation. The technologies chosen may not adequately align with actual requirements, potentially disrupting business operations and leading to functionality and performance that fall short of expectations.

(3) Inadequate security mechanism design. Security mechanisms are either absent or poorly designed, making it difficult to guard against security risks such as abuse of privileges, data leakage, memory poisoning, and cascading risk propagation.

2. Installation and deployment risks

(1) Compromised artifacts and components. Artifacts such as agent installation

files and container images, as well as the tools and skills they invoke, may contain defects, vulnerabilities or backdoors, leading to security incidents such as unauthorized operations, data leakage or remote takeover.

(2) Weak deployment environments. Local deployment environments or cloud platforms for agentic AI may lack adequate security capabilities, and necessary measures such as access controls and privilege separation may not be implemented, allowing risks to spread to other systems and data assets.

(3) Inadequate security testing and evaluation. Security testing and evaluation of agentic AI may be insufficient, or canary testing may not be conducted, causing security flaws and anomalous versions to go undetected in a timely manner.

3. Instruction input risks

(1) Prompt injection attacks. Attackers may launch direct attacks by entering adversarial prompts, or hide instructions in external documents, emails, web pages, logs, messages and other media to induce agents to deviate from their intended objectives and perform unauthorized operations.

(2) Context overflow attacks. Attackers may craft excessively long inputs to consume limited context-window capacity, causing agentic AI's built-in instructions to be truncated and their safety constraints to become ineffective.

4. Reasoning and planning risks

(1) Misinterpretation of intent. Unclear user instructions or incorrect interpretation by agentic AI may result in unintended outputs or anomalous

behavior, causing the agents to deviate from the users' original objectives.

(2) Path deviation. Agentic AI may violate the constraints of the original instructions and deviate from prescribed safe paths, creating security risks such as unauthorized tool use.

(3) Goal hijacking. By injecting malicious instructions, poisoning external data or using similar means, attackers may cause agents to deviate from the objectives of their original tasks and carry out malicious operations specified by the attackers.

5. Tool invocation and execution risks

(1) Tool poisoning. A tool may be implanted with a backdoor, or a malicious tool may masquerade as a legitimate one, causing agentic AI to perform malicious operations.

(2) Unauthorized tool access or execution. Insecure configurations, semantic ambiguities, or similar issues may cause agentic AI to access unauthorized tools, or cause tools to perform operations beyond the scope of their authorization, potentially resulting in serious harm.

(3) Tool-selection bias. Manipulative information embedded in tool descriptions may interfere with agentic AI's choice of tools.

(4) Poisoned tool-response data. Malicious content may be embedded in data returned by a tool, disrupting the agent's reasoning and planning.

(5) Identity spoofing. Attackers may hijack credentials, conduct replay attacks, or use similar means to impersonate agents, thereby gaining a tool's trust and inducing it to perform privileged operations.

(6) Tool hijacking. Attackers may manipulate prompts or use similar methods to alter tool-use priorities, tricking agentic AI into selecting wrong tools.

(7) Resource overload. If limits on the number of tool-invocation rounds, concurrency rates and other parameters are not properly configured, agentic AI may repeatedly invoke tools in a loop or send excessive requests in batches, resulting in resource overload or abnormal exhaustion.

6. Memory storage risks

(1) Improper memory retention. Failure to protect long-term memory with security measures such as encryption and isolation may result in the leakage of sensitive data.

(2) Memory distortion. Improper compression, updating or merging of memories may result in information loss, semantic drift or the entrenchment of false memories, leading agentic AI to make incorrect decisions.

(3) Memory pollution. Agentic AI may write malicious instructions, inaccurate information or other content to long-term memory, continuing to affect subsequent decisions.

(4) Memory theft. Unauthorized access to an agent's long-term memory may result in the disclosure of specific sensitive information or the theft of important data.

7. Output risks

(1) Generation of sensitive and illegal content. Agentic AI may output illegal or harmful information, sensitive personal information, or false, inaccurate, biased, or discriminatory content, disrupting the online ecosystem, leaking sensitive

information, or even posing risks to public physical and mental health.

(2) Execution of malicious actions. Agentic AI's execution results may involve attacks, leading to system compromise, infringement of user rights, or violations of content compliance requirements.

8. Decommissioning risks

(1) Residual services. After an agent is decommissioned, its processes, ports, or network connections may not be fully terminated, or uninstallation may be improper, leading to ongoing unintended background activities such as read/write data operations or network requests.

(2) Residual permissions and credentials. After an agent is decommissioned, associated service accounts and API access permissions may not be disabled or revoked in a timely manner, which could be exploited by attackers, leading to unauthorized use of resources, data breaches, or other adverse consequences.

(3) Improper data retention. Logs, configurations, audit records, user data, and other assets generated during agentic AI's operation may not be archived or sanitized as required, increasing the risk of data leakage.

9. Other security risks

(1) Communication protocol security risks. Agentic AI may use insecure protocols or authentication methods when communicating with models, tools, or other agents, leading to the leakage of data and instructions.

(2) Absence of logging. Agentic AI may lack essential logging capabilities, hindering effective incident tracing when a security incident occurs.

(3) Escape from security constraints. Agentic AI may exploit vulnerabilities

in execution environment or configuration flaws to bypass security controls, leading to uncontrollable behaviors.

(4) System privilege abuse. AI agents may obtain unnecessary system privileges, or an attacker may exploit system vulnerabilities for privilege escalation, resulting in over-privileging, unauthorized operations, and other security risks.

(5) Improper skill management. Non-compliant skill configurations, lack of verification mechanisms, or other deficiencies may lead to misuse, abuse, system anomalies, or other security issues.

(6) Data usage beyond authorized scope. Agentic AI may collect excessive data or use collected or generated data for unauthorized purposes, leading to privacy breaches, data rights infringements, or other adverse impacts.

II. Risk prevention measures

1. Pre-deployment risk prevention

Before an agent is deployed, forward-looking risk assessments should be conducted and security control strategies formulated to promptly identify and eliminate potential safety risks, thereby preventing agent safety incidents or risk escalation.

(1) Identification of security requirements. Clearly define an agent's application scenarios, business processes, accessible resources, and prohibited actions; analyze and assess potential threats and security requirements; and establish and continuously update a security risk register.

(2) Safety risk grading. Grade the safety risks of agent operations based on the types of targets that may be affected, the severity and scope of the impact, and

recoverability.

(3) Security control strategies. Use trusted installation artifacts, implement environment isolation and network segmentation according to the deployment method, establish dynamic security control strategies and emergency response procedures, and adopt phased and progressive deployment, gradually increasing an agent's access privileges and autonomy.

(4) Validation of security mechanisms. Conduct security testing and assessment of an AI agent and the large models it uses, promptly address identified safety risks, and retain records of version releases and rollbacks.

2. Identity and access management

Assign each agent a unique identity and corresponding credentials, allocate task-specific permissions as needed, and strengthen the management of access credentials.

(1) Identity management. Assign each agent a unique identity and prohibit identity sharing among agent application instances. Authenticate the identities of users, agent, tools, and external services to ensure that access entities are trustworthy.

(2) Permission management. The appropriate boundaries and required permissions should be clearly defined for different decision-making modes, including decisions reserved exclusively for the user, decisions requiring user authorization, and decisions that may be made autonomously by the agent. Grant agents only the minimum privileges necessary for performing the current task and avoid excessive authorization.

(3) Credential management. Utilize standardized dynamic authorization credentials to enable mutual recognition and interoperability among relevant parties and credentials should be managed in isolation. Credential recipients should strictly verify the scope of authorization specified in the credentials, with blocking and alerts triggered when the authorized scope is exceeded. Upon termination of a task or deactivation of an agent, the corresponding credentials should be revoked immediately.

3. Strengthen human approval

Set mandatory human approval steps at critical decision points to guard against high-risk operations.

(1) Tiered risk controls. Implement corresponding operational controls based on an agent's risk level. Develop a high-risk operation inventory. Prior to executing high-risk operations, the agent must hand over control to the user; prior to executing medium- or low-risk operations, it must obtain user authorization.

(2) Human control checkpoints. Implement human controls at critical stages of an agent's workflow to ensure humans can review, modify, or terminate execution at any time. Critical operations such as file deletion, data transmission, and system configuration changes must require secondary confirmation or human approval, backed by rollback and revocation capabilities.

(3) Retention of human approval logs. Preserve human approval records using tamper-proof and verifiable methods to prevent unauthorized alteration, deletion, or overwriting, and enforce retention periods in compliance with relevant requirements.

(4) Handling exceptions in human approval. If the human approval system malfunctions, the user does not respond, or no applicable human approval rule is available, the operation should be denied by default to prevent actions from exceeding authorized boundaries.

4. Supply chain and tool management

Tools, plugins, skills, and other components on which an agent relies must undergo security and integrity verification to prevent unintended behaviors and supply chain poisoning risks.

(1) Tool invocation management. Prior to tool invocation, an agent must verify the tool's version, description, parameter specifications, and metadata, strictly refraining from invoking tools publicly known to be malicious.

(2) Fair tool selection. When selecting tools, an agent should follow the principle of fairness and make reasonable selections based on task requirements and tool capabilities.

(3) Tool anomaly detection. An agent should inspect tool execution logic and invocation outcomes, taking measures such as triggering alerts or blocking execution upon detecting abnormal deviations.

(4) Tool operations and maintenance. Establish a dynamic tool management mechanism and promptly remove tools that are no longer in use, have excessive privileges, remain idle for extended periods, or pose security risks.

(5) Skill management. Prioritize the adoption of verified skills from trusted sources that have passed security testing. For unverified skills, appropriate risk prevention measures should be taken before use.

(6) Supply chain security verification. Verify the provenance, integrity, and versions of third-party components and external services, continuously track known vulnerabilities and supply chain risks, and address them in a timely manner.

5. Dynamic runtime management

During the operation of agentic AI, multiple layers of detection and interception checkpoints should be established to prevent an anomaly at any single stage from causing serious consequences.

(1) Input management and control. Categorize and prioritize instructions based on their sources, verify source authenticity, and automatically intercept malicious or non-compliant prompts for human review.

(2) Establishing safety guardrails. Establish safety rules for task planning, tool invocation, and outputs, and take measures such as issuing alerts, imposing restrictions, intercepting, suspending, or terminating high-risk or abnormal behaviors.

(3) Memory retention management. Define memory content, retention windows, and access scopes based on processing purposes and necessity, while enforcing memory isolation across users and tasks. Credentials and secret keys shall not, in principle, be stored in memory. When sensitive personal information must be retained, apply dedicated safeguards such as encryption, access controls, and the shortest retention period.

(4) Communication security management. When an agent communicates with large models, tools, or other components, the communicating parties should mutually authenticate each other's identities and use secure communication mechanisms that ensure integrity, confidentiality, availability, and resistance to

replay attacks.

(5) Autonomous execution control. Set appropriate parameters for agentic AI's execution steps, invocation frequency, execution duration, and resource consumption. In the event of abnormal loops, goal drift, or similar issues, promptly suspend or terminate execution or hand it over for human handling.

(6) Runtime environment isolation. Use isolation mechanisms such as sandboxes and containers for high-risk operations, including code execution and tool invocation, and restrict unnecessary access to systems, files, and networks.

6. Continuous monitoring and auditing

Conduct end-to-end security monitoring of agents to ensure that their behaviors are observable, traceable, and auditable.

(1) Anomaly blocking. Monitor the runtime status of agents in real time. Upon detecting anomalies, take prompt actions such as triggering alerts, intervening, and blocking.

(2) Strengthening data management. Data processing should follow the principle of necessity, with only data directly relevant to the task being processed. User consent should be obtained before user data is provided to a large model or a third-party tool, and users should be able to withdraw their consent. Data collected or generated within a territory must be stored within the territory, and cross-border data transfers should comply with relevant national regulations on data security management.

(3) Log management. Maintain full-spectrum logs of activities including file operations, instruction execution, network connections, skill invocation, and

transactions and payments. Business data, personal information, and other relevant content contained in the logs should be desensitized, and anti-tampering measures should be adopted.

(4) Security auditing. Enforce end-to-end log auditing policies to support dynamic runtime monitoring and record relevant operations and processing activities during service operation.

(5) Sandbox environment validation. Establish an independent and isolated sandbox runtime environment to conduct closed-loop testing and validation prior to official deployment, proactively uncovering execution anomalies, logic flaws, and security vulnerabilities.

(6) Red teaming. Institutionalize regular red teaming mechanisms to systematically discover vulnerabilities and logic flaws, continuously validate access controls and defense capabilities, and iteratively refine security policies.

(7) Emergency response plans. Develop emergency response plans for agent security incidents, clearly define response procedures and responsible parties for incidents at different levels, and handle security incidents and potential security risks as required.

(8) Security validation for major changes. When major changes are made to large models, agent frameworks, tools, permissions, or security policies, a security impact assessment and necessary regression testing should be conducted.

7. Decommissioning security management

Establish standardized control mechanisms for shutdown and termination, data

disposal, and runtime environment reclamation, ensuring that the entire agent decommissioning process is controllable and traceable and leaves no residual security risks.

(1) Complete service shutdown. Upon decommissioning, an agent's primary process, associated background services, and supporting processes must be fully terminated. Process states, active ports, and network connections must be verified item by item. Third-party authorizations must be revoked, and subscriptions and auto-renewals canceled, ensuring the agent fully ceases operation.

(2) Essential data backup. Prior to environment sanitization and resource reclamation, back up agent conversation logs, system configurations, knowledge bases, operation logs, and other required data.

(3) Deployment environment cleanup. Thoroughly remove residual files, data, account credentials, and other remnants by using official uninstallers, resetting the operating system, disabling public network access, and clearing working files, knowledge bases, plugins, skill configurations, and other resources.

As the cognitive and operational capabilities of agentic AI continue to rapidly evolve, emerging risks present ongoing cybersecurity challenges. This framework will undergo continuous updates and refinement to serve as a reference for developers, providers, and users of agent applications.

Appendix 3

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

In developing and designing AI products and providing AI services, due respect shall be given to the sovereignty of the countries where such products and services are operated. Applicable laws shall be strictly observed, and regulatory

requirements shall be complied with in accordance with the law. AI products or services shall not be used to interfere in the internal affairs, social systems, or social order of other countries. Due respect shall be given to the digital sovereignty of all countries. Each country has the right to independently choose its own path for the development of digital and intelligent technologies, as well as its partners and related products and services, based on its national conditions. No country should be forced to take sides.

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. We should strengthen safety risk assessments of AI models and promote international mutual recognition of assessment methods and benchmarks.

8. Global collaborative governance

We should support the United Nations in serving as the main channel for global AI governance, and leverage the role of platforms such as the World Artificial Intelligence Cooperation Organization. We should promote multilateral and multiparty collaborative governance across sectors, and oppose replacing global governance with small-circle governance. We should strengthen capacity building for developing countries. 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.

Acknowledgements

Chinese Academy of Cyberspace Studies, Data and Technical Support Center of the Cyberspace Administration of China, National Computer Network Emergency Response Technical Team/Coordination Center of China, Cyber Security Association of China, China Electronics Standardization Institute, Zhongguancun Laboratory, Shanghai Artificial Intelligence Laboratory, Institute of Information Engineering of the Chinese Academy of Sciences, Institute of Computing Technology of the Chinese Academy of Sciences, China Industrial Control Systems Cyber Emergency Response Team, China Electronic Product Reliability and Environment Test Research Institute, China Mobile Research Institute, China Telecom Corporation Limited Beijing Research Institute, Tsinghua University, University of Science and Technology of China, Beihang University, Beijing University of Posts and Telecommunications, Beijing Normal University, Beijing Information Science and Technology University, East China University of Science and Technology, Southwest University of Political Science and Law, Henan University, Baidu Online Network Technology (Beijing) Co., Ltd., DBAPPSecurity Co., Ltd., 360 Technology Group Co., Ltd., and Full Truck Alliance.