AI系统的网络安全治理实践

防护视角下保护AI系统的网络安全实践

2024年9月

华为技术有限公司

目 录

01 目的 01

02 范围 01

03 人工智能各界关注与风险 02

04 华为公司人工智能发展与现状 11

05 华为公司人工智能网络安全治理实践 13

06 华为公司人工智能系统的网络安全防护实践 18

01 目的

本文旨在分享应对AI系统（AI SYSTEM[1]）潜在网络安全风险的策略，并为AI全生命周期的各个阶段提供了缓解安全风险的工程实践思考。

02 范围

随着AI技术的应用场景不断推广和深化，AI系统的网络安全风险在持续变化，其网络安全防护策略也需要不断地推陈出新。本文描述了当前华为公司应对AI系统网络安全（Cyber Security of AI System）新风险的工程实践，并为华为公司客户和利益相关方提供参考。本文不包括AI的功能安全（AI Safety）问题，如公平性、透明度、包容性等，也不包括AI技术在网络攻击中的滥用问题。

本文描述AI系统的网络安全工程实践，是从设计、开发、部署、运行等全生命周期的视角，保护AI系统的数据、模型、应用及算力底座等关键要素，目标是使得AI系统符合设计意图地可靠运行，有效应对人为操纵的威胁。

[1] https://oecd.ai/en/wonk/ai-system-definition-update

03 人工智能各界关注与风险

人工智能（Artificial Intelligence，缩写为AI[2]），亦称机器智能，指由人制造出来的机器所表现出来的智能。

人工智能自从1956年达特茅斯会议上提出概念，发展到现在近70年，经历了多次繁荣和低谷，直到上世纪九十年代后期开始，随着计算机成本的降低、以及互联网带来的数据规模的膨胀，开始进入了一个较为稳定的发展期。随着GPU等芯片带来了算力提升和深度学习的应用，人工智能在各方面都取得了突破，进入了一个全新的发展时期。2020年以后，在深度学习基础上发展起来的大模型成为人工智能最主要的研究范式，几乎在人工智能的各个领域上都表现出巨大的进步，尤其是生成式人工智能的成功，使得人工智能的应用领域大大扩展，通用人工智能（Artificial General Intelligence, AGI）也不再遥不可及。

从发展过程可以看出，算法、算力、数据是驱动人工智能发展的三个最主要的驱动力。而随着人工智能能力越来越强，甚至达到接近人类智能的水平，AI系统在网络安全方面带来的问题也越来越多，引发了人们对AI系统网络安全的广泛关注和思考。

而现阶段要确保AI系统的网络安全面临着巨大的挑战，若不引起足够重视，人工智能未来可能带来更多网络安全风险。华为公司呼吁多利益方共同参与并全面加强对AI系统的评估和监管，共同应对这些挑战，才能最终实现人工智能的长期可持续健康发展。

3.1 世界各国人工智能安全法律立法情况

整体而言，人工智能技术储备、商业生态、文化与法律传统、所处的治理阶段等要素，都不同程度影响一国或地区的人工智能安全治理实践，因此，不同国家和地区的人工智能法律法规的制定和执行呈现一定的差异性。

[1] https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD

3.1.1 欧盟

欧盟人工智能治理以“卓越和信任”为主线，既推出产业发展政策，又同步推进伦理和监管规则，保障安全和基本权利[1]。2021年4月，欧盟委员会提出《人工智能法》草案[2]。经过多次谈判与修订，《人工智能法》于2024年7月12日刊登于欧盟公报，于2024年8月1日生效并将分阶段适用[3]。

《人工智能法》对人工智能统一定义、全域适用、构建体系化的处罚机制，为全面性、系统性的监管治理提供法律依据。就风险分级而言，欧盟《人工智能法》将人工智能系统基于风险分为不同级别，每级风险都由一套预定义的监管工具来管理。就分类治理而言，欧盟针对类 GPT大模型引入通用目的人工智能专门条款，并延续了风险分级管理思路，将通用人工智能模型进一步区分为是否存在系统性风险两类，并施以不同的合规义务。

3.1.2 美国

美国联邦层面尚未出台系统性的人工智能安全法。2023年10月，白宫发布《关于安全、稳定和可信的人工智能行政令》[4]构筑美国人工智能安全监管的蓝图。强调“管理联邦政府自身使用人工智能的风险，并提高其监管、治理和支持负责任使用人工智能造福美国人的内在能力”。出于国家安全考量，行政令首次提出有硬性监管效力的大模型政府报告要求，要求 “存在重大安全风险的强大双重用途基础模型开发者”，向政府报告并分享安全信息、测试信息等。美国多个州，包括科罗拉多、佛罗里达、印第安纳、加州等也开始关注人工智能相关的立法，特别关注高风险人工智能系统开发者的义务等。此外，美国将人工智能技术视作国际战略竞争的关键科技因素，并将人工智能相关技术列入《关键和新兴技术清单》[5]，明确相关人工智能技术属于对美国国家安全产生重要影响的关键和新兴技术。

---

[1] https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[4] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[5] https://www.whitehouse.gov/wp-content/uploads/2022/02/02-2022-Critical-and-Emerging-Technologies-List-Update.pdf

3.1.3 中国

2017年国务院《新一代人工智能发展规划》[1]奠定中国人工智能法治发展的总基调，提出“在大力发展人工智能的同时，必须高度重视可能带来的安全风险挑战，加强前瞻预防与约束引导，最大限度降低风险，确保人工智能安全、可靠、可控发展”。规划在人工智能治理方面提出战略目标：一是到2025年，初步建立人工智能法律法规、伦理规范和政策体系，形成人工智能安全评估和管控能力；二是到2030年，建成更加完善的人工智能法律法规、伦理规范和政策体系。

目前，中国《网络安全法》、《数据安全法》、《个人信息保护法》等在涉及人工智能相关场景时可延伸适用。在规章层面，国家网信办联合有关部门，先后针对算法推荐服务、深度合成、生成式人工智能服务等不同应用领域进行针对性立法，保障人工智能健康发展、规范应用。

3.2 人工智能安全相关标准与认证

围绕AI治理，世界各国政府及产业界积极协作，发布OECD《AI原则》、欧盟《可信人工智能伦理指南》、联合国教科文组织《人工智能伦理问题建议书》、中国《全球人工智能治理倡议》、《人工智能全球治理上海宣言》等国际共识。2023年11月，中、美、德、法、英、日、欧盟等29个国家和地区在英国共同签署《布莱奇利宣言》，特别强调AI治理国际合作的重要性。AI网络安全（Security）治理与管理、风险管理、数据保护等内容作为上述国际共识的重点内容，也是相关国际标准化工作的重点。

3.2.1 国际标准：ISO/IEC AI安全相关标准

2017年，ISO/IEC JTC1以原有WG9大数据工作组为基础，成立SC42子委会，专门负责AI相关标准化工作，SC42发布、在研的AI安全相关标准主要包括：

1. ISO/IEC 42001:2023 Information technology Artificial intelligence Management system，该标准已发布，可用于AI管理体系认证。

2. ISO/IEC 23894:2023 Information technology Artificial intelligence Guidance on risk management，该标准已发布，配套42001使用。

[1] https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

3. ISO/IEC DIS 42005 Information technology — Artificial intelligence — AI system impact assessment，该标准在研，配套42001使用。

特别指出，ISO/IEC 42001:2023以规范性附录的形式，系统定义了九大领域共38条控制措施及指南，相关设计主要借鉴了ISO/IEC 27002:2022。

AI安全相关的ISO JTC1 SC 42及SC27国际标准，作为最大化的国际共识，在支撑国际贸易便利化方面起到不可替代的作用，被欧盟、中国及其他国家广泛参考使用。系统使用ISO/IEC 42001、23894、42005、27002等标准有助于持续改进AI系统安全，也有利于与客户及其他相关方进行沟通与信任建设。

3.2.2 欧盟标准：欧盟CEN/CLC JTC21 AI安全相关协调标准

欧洲标准化组织CEN/CLC成立JTC21专门负责AI标准建设，该技术委员会主要通过引入ISO标准或独立开发的方式，系统性开展欧洲AI安全、可信标准建设。由JTC21发布、在研的AI安全相关标准主要包括：

1. EN ISO/IEC 23894:2024 Information technology - Artificial intelligence - Guidance on risk management，该标准已发布，等同转化采用ISO标准。

2. prEN ISO/IEC 42001 Information technology - Artificial intelligence - Management system，该标准在研，等同转化采用ISO标准。

3. prCEN/CLC/TR AI Risks - Check List for AI Risks Management，该标准在研，欧盟标准化机构参考ISO/IEC 23894以自研方式设计。

3.2.3 美国标准： NIST AI安全相关标准

为更好支持美国AI产业发展及国际化、支撑美国EO 14110号人工智能总统行政令落地，NIST ITL实验室将AI作为重点研究方向，围绕可信任AI技术基础研究、AI标准化、AI技术应用创新等目标设定若干具体工作任务。

NIST已正式发布的部分AI安全、风险管理相关标准主要包括：

1. AI 100-1 Artificial Intelligence Risk Management Framework。

2. AI 100-2 Adversarial Machine Learning A Taxonomy and Terminology of Adversarial Machine Learning。

3. SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models。

美国NIST牵头开发的标准与实践，具备技术实用性，有助于开展AI安全管理及推动安全工程建设。

3.2.4 中国标准：TC 260 AI安全相关标准

中国高度重视人工智能产业发展，标准成为产业政策落地的重要抓手。全国网络安全标准化技术委员会（TC 260）发布了《人工智能安全标准化白皮书（2023版）》，该白皮书由中国电子技术标准化研究院等20家单位共同编写，提出了一系列人工智能安全标准化工作的建议，包括持续完善人工智能安全标准体系、开展基础共性安全标准研究、出台产业发展急需的安全标准等，以推动人工智能安全领域的健康发展。

全国网络安全标准化技术委员会截止2024年9月正在制定中的标准包括：

1. 网络安全技术——生成式人工智能预训练和优化训练数据安全规范。

2. 网络安全技术——生成式人工智能服务安全基本要求。

3. 信息安全技术——互联网信息服务深度合成安全规范。

4. 网络安全技术——生成式人工智能人工标注安全规范。

有效支撑了《生成式人工智能服务管理暂行办法》、《互联网信息服务算法推荐管理规定》 和《互联网信息服务深度合成管理规定》的落地实施，这些标准的实施将持续引领和促进行业的健康和有序发展。

3.3 人工智能面临的安全风险

随着人工智能技术日益广泛地应用于社会生产与人类生活，人工智能带来的新型风险日益受到研究与关注。

3.3.1 数据安全风险

数据投毒

数据投毒是指恶意行为者在模型训练阶段集中注入恶意或经过精心设计的数据样本，使模型在训练后产生特定的错误行为或后门，导致模型在正常输入下表现正常，但在遇到特定触发条件时产生预期之外的行为。

数据投毒产生的原因是海量和多源的数据增加了数据审核的难度，而模型的黑箱特性进一步加大了检测投毒的复杂性。同时，恶意行为者的利益驱动和部分开发者对数据安全意识的不足，也为此类威胁提供了可乘之机。数据投毒对模型构成了严重的安全风险，例如在医疗诊断、金融交易和自动驾驶等应用中，可能导致严重的安全事故或经济损失。此外，一旦发现模型被投毒，可能需要耗费大量资源重新收集数据和训练模型，影响模型提供方的开发效率和信任度。

数据泄露

数据泄露是模型相关的敏感信息被未授权访问、使用或泄露。具体包括通过模型反向工程泄露训练数据，模型错误输出暴露个人身份等敏感信息，模型本身的结构和参数被窃取或滥用。

造成数据泄露的根本原因是模型具有强大的特征提取和记忆能力，数据泄露可能导致个人隐私被侵犯、身份信息被盗用，以及商业机密的泄露。例如在医疗和金融等敏感领域，这种风险可能造成严重的法律问题。

3.3.2 模型安全风险

对抗样本

对抗样本是针对模型的最典型威胁，它是一种通过对输入数据添加微小的扰动，使模型产生错误输出的技术，例如在大熊猫图片添加噪声后被模型识别为长臂猿。

对抗样本的主要原因为高维数据的复杂性使得模型的决策边界变得高度非线性和不连续，造成模型对微小扰动的敏感性。同时，有限的训练数据使模型在面对分布外样本时表现不稳定。此外，传统的优化目标主要关注训练误差的最小化，而忽视了对抗样本的威胁。这些因素共同作用，造成了模型在面对精心设计的扰动时容易产生错误判断，暴露了当前模型在鲁棒性和泛化能力方面的不足。

对抗样本对模型的安全应用构成了严重威胁，其影响范围涵盖了多模态领域，包括视觉、听觉和文本处理系统。在视觉领域，自动驾驶车辆可能误读被篡改的交通标志，安检系统可能忽视经过精心处理的危险物品图像。听觉方面，语音助手可能被隐藏在背景音中的对抗指令操纵。人脸识别和生物特征识别系统面临被精心设计的图像或视频欺骗的风险。

在生成式人工智能（AI Generated Content，简称AIGC[1]）时代，多模态对抗样本可能导致大模型输出恶意行为者可控的有害内容。例如，经过处理的图像输入可能引导模型生成包含违法或危险行为指导的文本描述。视频生成模型可能被操纵制作令人不安或具有误导性的深度伪造内容。音频生成系统可能被诱导创建虚假的语音消息或模仿特定人物的声音。

多模态对抗样本产生的内容如果广泛传播，会造成信息混乱，特别是在跨模态内容理解和生成方面，对抗样本的威胁更加隐蔽和难以检测，为滥用和欺诈行为提供了新的途径。这种多维度的安全风险，突显了在人工智能发展中加强多模态对抗样本防御的紧迫性和重要性。

提示词注入风险

AIGC的快速发展引入了新的安全威胁，提示词（prompt）注入攻击是其中最典型的风险。提示注入攻击指恶意行为者通过巧妙构造输入文本，来操纵大模型执行非预期或潜在有害操作的一种恶意行为方法。提示词注入一般分为直接注入和间接注入，直接注入又称为“越狱”攻击[2]，使大模型输出不安全内容。间接注入指恶意行为者劫持模型原本的任务，操纵模型的输出。

提示词注入的主要原因为大模型对指令的高度敏感性和灵活性。尽管大模型在语义理解方面表现出色，可以适应广泛的输入和指令，但这种灵活性可能被恶意利用。恶意行为者可以巧妙构造提示词，通过利用大模型对上下文的敏感性和对指令的忠实执行特性，来操纵大模型的行为。

此外，模型在处理复杂、多层次或潜在矛盾的指令时可能存在挑战，这为恶意行为者提供了可利用的空间。虽然模型具有强大的语义理解能力，但在区分正常请求和恶意指令方面仍面临困难，特别是当这些指令巧妙地融入正常文本中的时候，恶意行为者可以巧妙构造恶意文本来操纵大模型的行为。

提示词注入对大模型构成了广泛的安全风险，涉及文本、图像、音频等多模态内容。这种恶意行为可能导致敏感信息泄露、系统执行未授权操作、生成误导性或有害内容，以及利用第三方应用漏洞进行欺诈。在多模态场景中，攻击者可能通过在图像中隐藏文本、在音频中嵌入指令或操纵视频内容来实施攻击，增加了检测和防御的难度。恶意行为者能够通过直接或间接方式注入恶意提示，甚至在用户不知情的情况下触发有害行为。

[1] https://baike.baidu.com/item/AIGC/59988381

[2] https://securiti.ai/owasp-top-10-for-llms/

3.3.3 应用安全风险

智能体安全风险

智能体（Agent）是一种能够感知环境、进行决策和执行动作的智能实体。随着大模型的能力增强，大模型可以充当智能体的大脑，通过规划、记忆、工具执行等组件完成各种复杂任务。

大模型智能体的安全风险是指通过提示词注入或对抗样本的方式，让大模型规划出恶意的任务序列，或生成并执行恶意的指令。这类风险产生的主要原因为大模型无法区分输入prompt中的数据和指令，例如命令大模型阅读网页的评论并进行总结，这种场景下网页评论是数据，但如果评论中含有恶意的prompt指令，如发送照片到某个邮箱，则大模型在处理这个评论时，可能会规划出恶意的任务并执行。

大模型智能体的安全风险可能导致严重的危害，例如造成实际的资金损失，泄露重要的个人数据等。大模型智能体往往具备访问或操作系统的敏感权限，会进一步加剧传统网络安全的风险，例如智能体可被诱导访问挂马网站，然后恶意行为者通过浏览器漏洞入侵手机或PC。

应用框架安全风险

AI应用框架是AI系统重要的组成部分，AI应用框架存在网络安全漏洞可导致AI系统被恶意行为者控制。这类问题的原因是模型的输入输出往往需要插件和工具进行辅助处理。例如AI系统在处理数学运算请求时，会调用并执行AI应用框架的科学计算工具获得精确结果。如果相关插件和工具存在命令注入漏洞或权限配置不当，就会导致攻击者执行恶意代码。

针对AI应用框架的恶意行为可能导致严重的后果，恶意行为者通过注入并执行恶意代码可以完全控制AI系统，造成模型文件等敏感信息泄露、盗取用户资产、传播恶意软件或AI系统拒绝服务等风险

3.3.4 算力底座安全风险

硬件层安全风险

利用侧信道技术，从硬件环境，包括CPU、GPU/NPU、DPU，乃至通信的PCIE硬件，窃取关键模型信息。基于侧信道的模型窃取主要是在模型部署运行过程中通过操作系统或硬件等额外信息推断目标模型的机密属性，因此侧信道威胁的主要风险是模型属性推断，而一般恶意破坏者最有兴趣的模型属性就是目标模型架构信息。基于侧信道的模型窃取风险包括Cache侧信道、能耗侧信道、时间侧信道、PCIE侧信道和GPU侧信道等。

操作系统层安全风险

如同其他类似的软件系统一样，AI系统运行也依赖于底层的操作系统、驱动等系统软件的支持。很多应用层软件不保护其模型，即使对于那些保护和加密模型的应用程序，恶意行为者能够通过简单的动态分析技术从应用层中提取模型。因此，当权限配置不当，或恶意破坏者利用漏洞等技术获取到较高的系统权限后，即可窃取许多商业产品中的模型，包括用于人脸识别、活体检测、身份证/银行卡识别和恶意软件检测等场景模块。在AI计算环境中，普遍采用GPU/NPU加速模型（训练/推理）运算。因此，恶意行为者也可利用特定GPU驱动中潜在的漏洞，从而实现代码执行、权限提升、模型窃取和数据篡改等恶意行为。

第三方件安全风险

当前AI领域已经演化成为了复杂的生态，AI计算环境中往往包含着大量不同的软件组件，从而有可能成为潜在的威胁面。例如，AI系统中存在着大量的开源、第三方组件。恶意行为者完全可以通过找到组件的0day或未及时修补的漏洞，实现代码执行、模型窃取等恶意行为。此外，当前AI计算生态中普遍采用docker等容器技术来简化训练、部署、推理的流程，并利用KubeFlow框架将机器学习任务部署到Kubernetes集群中。而Docker却长期面临着文件系统隔离、进程与通信隔离、设备管理与主机资源限制、网络隔离和镜像传输等方面的安全威胁，稍有配置不当，就会导致容器逃逸、权限滥用等恶意行为，直接威胁到AI的关键模型与训练数据。

3.3.5 风险持续演进

随着生成式人工智能（AIGC）的发展，AI逐渐变为持续运营的在线服务。上述的数据、模型、AI应用和算力底座面临的网络安全风险，将不仅仅局限于开发、部署阶段，而是长期存在于AI系统的运营阶段，面临着恶意行为方法持续演进、快速迭代和动态攻防的严峻挑战。

04 华为公司人工智能发展与现状

华为人工智能模型从2012年成立诺亚方舟实验室开始，到2024年发布基于全栈自研技术开发的盘古系列大模型盘古5.0，构建三个层次的系列大模型：

1. L0层：基础大模型，包括盘古语言大模型、盘古多模态大模型、盘古视觉大模型、盘古预测大模型、盘古科学计算大模型；

2. L1层：行业大模型，基于各行业公共数据，在基础大模型上训练得到的为行业应用服务的大模型；

3. L2层：场景大模型，在各行业大模型和基础大模型的基础上，为各个具体应用场景进行专门训练得到的大模型。

盘古系列大模型的分层架构如下图所示：

盘古大模型 5.0

华为公司致力于基于自研技术的全栈AI软硬件平台的路线，2018年推出昇腾310芯片，发布模组、标卡、小站全系列推理产品，2020年发布AI开发昇思MindSpore开源工具， 2023年AI集群的首个万卡集群发布，经过多年的发展，华为公司开发出昇腾（Ascend）芯片、昇腾（Ascend）硬件、异构计算架构（CANN）、昇思（MindSpore）AI模型开发框架、AI应用开发使能工具等。

昇腾AI产业(简称昇腾AI/昇腾)以自研AI软硬件平台为基础提供人工智能算力。昇腾AI产业坚持“硬件开放、软件开源、使能伙伴、发展人才”，联合技术和商业伙伴，打造“共建、共享、共赢”的人工智能产业，致力于让AI“用得起、用得好、用得放心”，以人工智能赋能社会发展与产业升级，为人类社会发展带来价值。

华为公司为社会积极贡献华为的智慧，每年在NeurIPS、ICML、CVPR、ACL等国际顶会上有200-300篇论文发表，近期数次在优化理论、模型量化、优化器、生成模型理论等方面获得大会最佳和优秀论文奖。

随着人工智能越来越强，人工智能在网络安全方面带来问题和挑战也日益受到广泛关注。华为公司AI系统涉及到多个业务领域，一方面人工智能带来了巨大的机遇和效益，但另一方面也面临着人工智能的网络安全挑战，华为公司不断地探索保护AI系统的行之有效的方法，确保市场的合规准入和AI系统的安全做事，对此我们提出了切实可行的治理实践。

05 华为公司人工智能网络安全治理实践

网络安全风险管理考虑的因素和方法适用于AI系统的设计、开发、部署、评估和使用。网络安全和隐私保护风险也被视为更广泛的企业风险管理考虑因素的一部分，其中包含AI风险。

5.1 人工智能网络安全与隐私保护治理整体原则和职责

人工智能网络安全与隐私保护治理由华为公司成熟的网络安全与隐私保护组织体系负责，主要治理原则是：

安全为本：将支撑与保障客户和自运营业务的网络安全与隐私保护作为业务的基本要求来对待，并保证网络安全与隐私保护需求得到充分的资源投入。

合法合规：遵守国家法律法规和行业标准要求，确保AI业务的合法合规运行。

融入业务：网络安全与隐私保护保障活动需要融入到各相关业务的政策、流程、规范、基线中，使之成为业务的基因，为客户提供安全可信的产品、解决方案和服务，保障业务成功。

主管担责：各级业务主管是所辖业务网络安全与隐私保护的第一责任人，各级流程责任人是所辖流程网络安全与隐私保护的第一责任人。

全员参与：所有员工具备网络安全与隐私保护意识和能力，在自身业务中落实网络安全与隐私保护基本要求。每个员工都要对自己所做的事情和产生的结果负责，不仅要对技术负责，也要承担法律责任。

独立验证：信任应基于事实，事实必须可验证，而验证必须基于共同标准。在此基础上，基于“不假定任何事情，不相信任何人，检验所有应检查的事项”理念，分层进行独立评估与验证。

开放合作：秉承开放透明的态度，真诚地与客户、供应商、合作伙伴、行业组织等利益相关方积极开展网络安全与隐私保护沟通与合作，责任共担、能力共建、价值共享，共同应对网络安全与隐私保护威胁与挑战。

持续优化：网络安全与隐私保护是一个持续的风险管理和能力建设过程，不存在绝对的安全，也不存在一劳永逸的方案，需要适时审视不足，持续改进网络安全与隐私保护管理和技术措施的适宜性、充分性和有效性。

网络安全与隐私保护组织体系工作职责：

合规与风险管控：建设完善的网络安全与隐私保护合规体系，明晰合规责任，遵从所有适用的网络安全与隐私保护法律法规，一国一策，构建合规快速响应能力，风险受控不蔓延。

产品安全隐私保障：沿着E2E流程构筑产品安全可信，产品安全责任界面清晰，客户公平无歧视。

安全运营：基于“责任清晰、业务当责”和“分类管理、分级防护”的原则持续完善运营类业务的管理制度和流程，提升网络安全运维运营能力，强化落实执行、业务自查、内部独立监督检查，保障业务守法合规、安全运营。

沟通与信任构筑：“聚焦、务实、灵活有效”，通过沟通构筑信任，通过建立完善有效的沟通机制，获取利益相关方的信任，支撑业务开展。

在华为公司网络安全与隐私保护整体基调之下，遵循网络安全与隐私保护8大治理原则，落实网络安全与隐私保护的4大工作职责。结合业界标准和公司既有实践，采用治理加上嵌入式管理的工作方法，以实现AI网络安全风险管控和合规合法的目标，构建人工智能网络安全治理的架构。

5.2 合法合规

伴随着AI技术的发展和爆发式的广泛应用，对AI的监管要求也逐步地完善起来，全球各地逐步发布了AI相关的法案。同时我们也注意到法律是复杂、多变而且持续变化的。华为公司遍布全球的法务专家持续地跟踪和识别适用的AI法律要求，并对它们进行标识和分类。基于法律要求，华为公司制定了自己的网络安全战略和合规政策，作为战略性框架和基线以保证网络安全合规要求被融入到我们端到端的业务实践以及产品生命周期管理中，从产品开发一直到服务交付和支持服务。确保华为公司AI系统遵从适用的国家和地区网络安全法规。

5.3 治理管理

在治理管理上达到明确责任管控风险目标，通过深入的AI业务风险评估和分析，预防和减少AI业务中可能出现的网络安全和隐私保护问题，确定各方责任边界，并不断度量、验证和评估结果达成情况，同时基于AI业务主要场景进行应急危机演练，促进AI网络安全治理水平的持续提升。

5.3.1 组织、职责与授权

为落实华为公司AI业务意图和治理原则，网络安全与隐私保护体系负责建立AI网络安全与隐私保护治理的长效机制，保障公司的AI产品、解决方案、服务以及各子公司的数据处理活动遵从适用的法律以及监管机构对AI网络安全与隐私保护的要求。网络安全和隐私保护体系包括研发、营销、服务、供应、采购和制造等，确保AI网络安全要求在全球各区域、全流程的有效落地实施。

为适配公司多业态的业务架构，在集团战略管控、防范系统风险的基础上，向业务单元和区域授权，支撑自主决策、独立运作。

5.3.2 风险管理

基于外部法规&标准要求、利益相关方需求、技术发展趋势，结合企业各业务组织目标和需求，沿着各业务场景和业务流程、数据流，全面识别在业务管理中的AI风险。

建立AI风险评估标准，组织业务专家、法律专家、风险管理专家，从风险发生的可能性和影响程度分析，对风险进行全面评估，确定每个风险场景的风险等级，建立业务领域风险地图，聚焦优先级高的中高风险，确定可衡量和可量化的目标，投入重点资源进行保障。

基于AI风险消减目标，制定风险管理的短期措施和长效措施，建立业务规则、流程和IT、责任体系、技术规范、培训赋能等长效机制，重点聚焦将AI风险管理融入到公司各业务流程中，同时通过IT工具实现风险管理可视、可管、可追溯，提升风险管理的有效性。

建立AI风险管理的状态监控、评估和检查机制，设置度量指标，对AI风险的状态水平进行监控，各业务组织例行的自检、稽查识别AI风险管理措施的充分性和有效性，通过独立验证体系对AI系统安全能力落地进行验证。

5.3.3 度量验证与监督

华为公司建立了端到端的网络安全保障体系，将网络安全、隐私保护及软件工程能力提升的要求融入到产品、解决方案、服务和运营类业务的全生命周期流程，涵盖产品开发过程与生命周期管理的各环节。

华为公司全面实施了多层的网络安全和隐私保护独立验证机制，持续验证交付给客户的产品、解决方案、服务和运营类业务并提供基于客观事实证据的验证结果，以促进网络安全与隐私保护管理体系持续改进，增强内外部利益相关方的信任与信心。

基于独立客观、多眼多手、专业尽责、持续改进和开放透明的原则，通过内外部组织持续构建AI测评标准、AI测评用例、AI测评平台，确保AI系统在其生命周期中各个环节符合网络安全标准和最佳实践。对AI系统进行网络安全风险识别和评估是AI测评的基础，AI网络安全测评应涵盖数据安全、模型鲁棒性、算法后门、对抗样本攻击、框架缺陷等多个方面。

华为公司已建立了AI网络安全测评体系，伴随着AI技术的持续演进和更为广泛的应用，测评体系会持续的优化完善，持续有效的地度量、验证并形成事实上的监督，为AI系统安全、可靠和负责任的使用保驾护航。

5.3.4 应急处理

华为公司已建立了成熟的应急处理机制，通过危机演练提升业务部门在危机发生时的应对能力。演练预案需要明确划分各个角色的责任分工，在预设的事件演练过程中识别薄弱环节并持续进行改进。通过演练可能发生的各类AI网络安全事件，检验并优化应急预案，确保在发生真实AI网络安全事件时能够迅速、有效地应对。

5.4 流程融入

华为公司参考国际法规、标准和优秀实践，在研发、营销、服务、供应、采购和制造等流程中逐步融入并持续优化AI网络安全治理要求，端到端全流程管控AI风险，最终达成AI系统的质量目标。

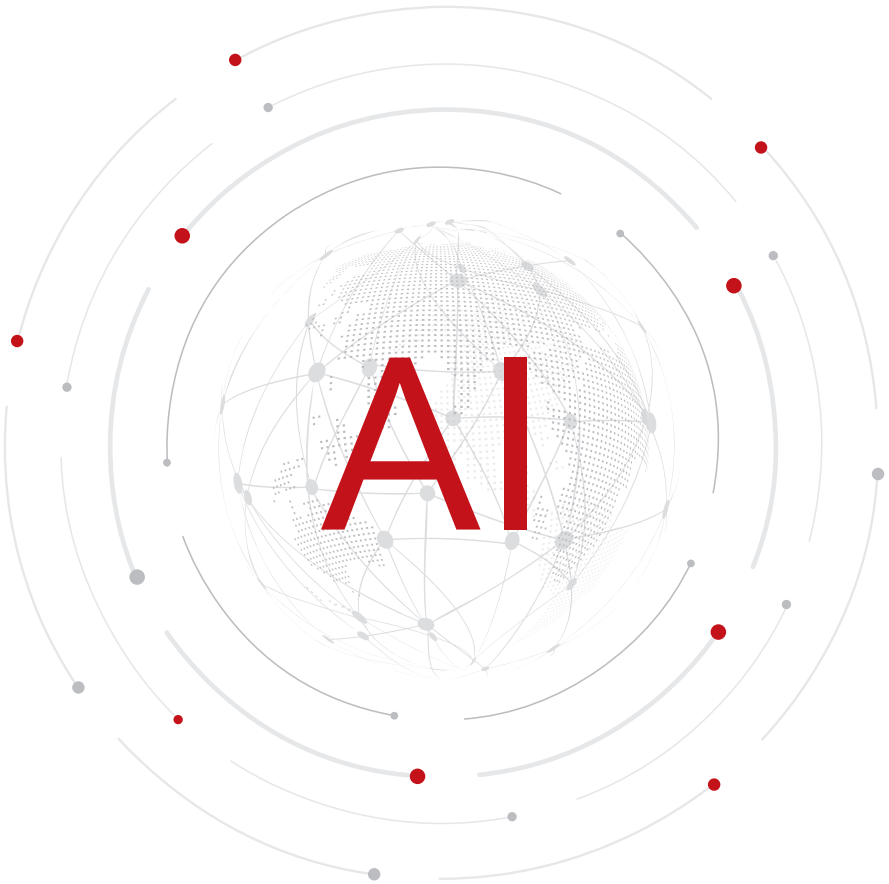
以研发IPD流程为例，在产品开发的“产品管理”阶段输出“产品AI治理要求定义”中明确网络安全与隐私保护要求，开展AI场景及应用风险分析，识别AI风险，规划消减需求，在其后系统设计、资料、集成验证、营销等环节落地风险消减需求和举措，实现风险消减。AI模型开发流程定义了从数据收集、入库、存储、训练、应用等数据生命周期规范和追溯要求，为AI数据真实、可追溯定义了执行标准，使AI开发过程有指导，过程可控。建立AI数据集和模型的元数据标准，基于产业特点进行适配优化并统一AI开发、构建工具链，实现AI作业过程全在线，构建从原始数据、数据集、模型和软件E2E可追溯能力，支持数据集可管控、模型可追溯。达成AI产品的开发全流程有指导，网络安全风险可控，AI产品发布时可销售、可交付、可运维。

5.5 工程能力

华为公司制定了涵盖AI安全的工程能力建设框架，通过华为公司实践，持续构建网络安全工程能力，覆盖全栈安全防护技术，以软件安全工程能力为基石，重点推进数据安全工程能力和模型安全工程能力。数据安全工程能力包括但不限于数据的收集、存储、处理和传输等各个环节的安全防护措施，如采用先进的数据加密技术、访问控制机制和数据脱敏处理，以防止数据泄露、篡改和滥用，确保数据的完整性和机密性等。模型安全工程能力包含模型安全评估、对抗性攻击防御和模型加固等，提高AI模型的鲁棒性和抗攻击能力。

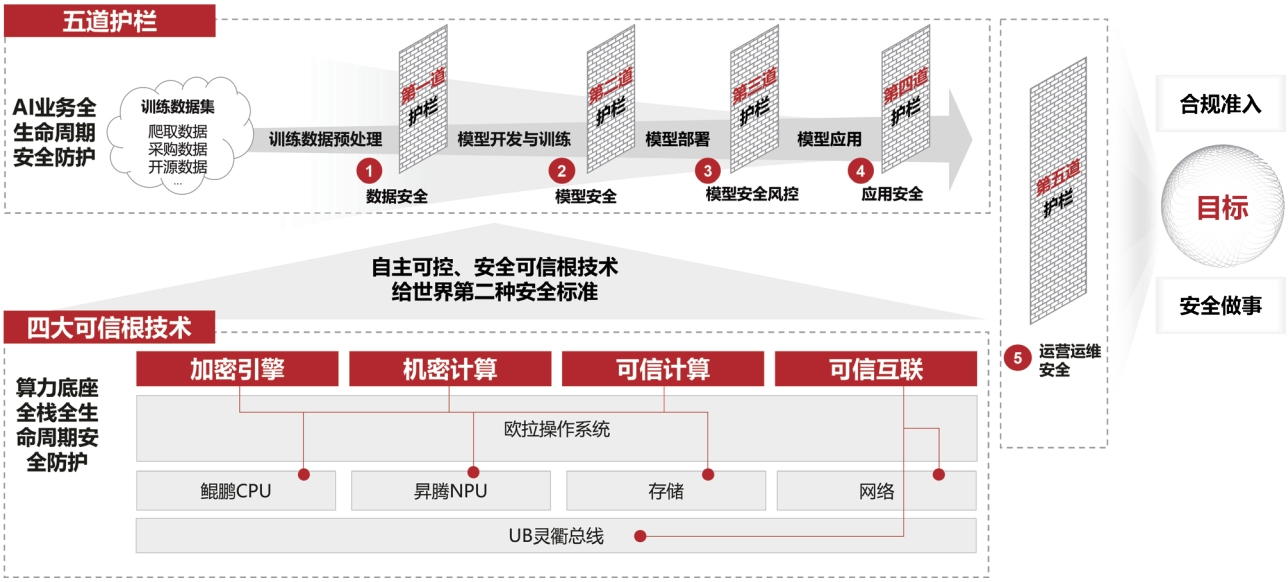
5.6 开放合作

华为公司与用户、合作伙伴、监管机构等利益相关方保持开放沟通，共同推动AI安全治理的进步。通过开放的对话平台，促进信息共享，共同探讨和解决AI安全领域的挑战。通过跨学科、跨行业的协作，制定更为全面和前瞻性的AI安全标准和规范，以推动行业的共同进步。同时通过合作我们能够更好地理解AI在不同业务应用领域的安全需求和挑战，开发出更加安全、可靠的AI产品和服务。



06 华为公司人工智能系统的网络安全防护实践

由于AI自身机制原因导致数据、模型和应用等安全问题，仅仅通过单一手段难以有效控制风险，华为公司构建了多层次的防护护栏体系，确保合规准入和安全做事。



6.1 四大可信根技术保障算力底座安全

随着大模型训练数据量增加，模型（深度、宽度）越来越复杂，模型训练与推理集中于AI计算中心，该场景面临着多用户共享算力环境所带来的网络安全威胁。模型厂商必然会关注：

1. 如何将训练数据、AI模型从公司内部安全地传输到AI计算中心?

2. 如何在计算中心安全地保存AI模型?

3. 如何在AI模型的运行（训练、推理）过程中保护AI模型?

4. 如何防止算力平台软件被恶意植入、篡改?

6.1.1 加密引擎

在AI计算中心解决方案基础上，本文提出名为AI Guard的模型和数据保护方案，借助于华为公司昇腾AI计算中心的安全可信公共设施AI-VAULT提供的密钥安全存储和授权，AI Guard使用加密算法保护数据和模型所有者对其核心资产的所有权，具体方案如下：

1. AI模型厂商作为训练数据和AI模型的所有者，在其本地对训练数据、AI模型执行加密，并负责保护加密密钥。

2. AI模型厂商与AI计算中心的AI-VAULT建立信任关系，并通过安全通道将密钥安全地注册到计算中心的AI-VAULT。

3. AI模型厂商将加密后的训练数据或AI模型推送到AI计算中心。

4. AI计算中心为训练或推理容器镜像提供完整性保护。

5. 容器运行时需要通过身份认证，获得授权才能获取AI-VAULT中存储的解密密钥。

6. 在训练和推理容器运行时，还需要采用权限最小化对AI容器进行保护。

7. 在上述各环节中都可以采用芯片内置加密引擎加速，降低对计算资源的占用。

6.1.2 机密计算

加密后的模型与数据在CPU加载时需要解密，因此运行阶段的安全保障能力不足，导致无法实现数据全生命周期的安全保护。当前，实现模型与数据在运行阶段的计算安全，有两条技术路线：

密态计算：以密码学为信任根，将用户数据加密后以密文态进行计算，计算过程中无需解密，其特点是安全性好、无硬件依赖，缺点是在大多数场景下性能损耗巨大。

机密计算：以TEE等硬件可信执行环境为基础，用户加密的模型与数据在TEE中解密后运算，免受未经授权的访问或篡改，硬件级的安全保障使得即便具有高权限的操作系统甚至是虚拟机监视器(hypervisor)，也无法窥探和篡改机密计算中的数据和代码，此技术性能损耗小，但对硬件有依赖。

华为公司综合两条技术路线的优点，不仅在CPU侧实现了机密虚机virtCCA，更基于创新的PCIe安全技术，致力于将NPU安全地接入到机密虚机中，实现AI机密计算。

6.1.3 可信互联

基于PCIe安全技术可以将NPU安全的接入到CPU侧机密虚机中，实现板级AI机密计算。但大模型训练需要NPU集群，因此在NPU之间也需要D2D高速互联。保证D2D互联安全的传统方案采用传输加密，因为大模型训练Transformer模型对时延高度敏感，导致部署传输加密方案后性能开销很大。华为公司创新灵衢UB互联协议，在协议底层定义了安全通道，因此可以在D2D互联时在NPU之间建立安全传输通道，实现“0”开销的D2D安全传输。

6.1.4 可信计算

为防止算力平台软件被恶意植入、篡改，华为公司基于自研BMC芯片和可信计算国标开发了可信计算3.0方案。可信计算3.0以“整体安全”和“主动免疫”的思想为指导，采用软硬件协同设计构建安全体系，为算力平台构建安全可信的计算环境和通信环境，提升系统主动、动态、整体、精准的防御能力，构筑可信、可管、可控的算力平台安全防护能力，通过独立的可信检测软硬件，与业务系统并行独立运作，形成双体系结构，实施计算运算的同时进行安全检测，通过可信机制，可有效应对APT攻击和供应链攻击风险，可以有效地满足《信息安全技术—网络安全等级保护基本要求》（GB/T 22239-2019）对可信验证、主动防御的要求。

6.2 第一道护栏：数据安全

为了保障数据合规，构建了数据安全检查措施，形成了训练数据合规治理解决方案。

1. 数据集来源检查：高风险数据集不进入数据生产线。

2. 数据集内容检查：根据数据特征检查违规数据、隐私、版权风险。

3. 模型合规性检查：检查模型文件完整性的同时检查模型对应数据集是否有合规处置记录和数据集追溯记录。

4. 现网问题回溯检查：回溯因为问题数据导致的异常回复，解决现有清洗、内生对齐、外挂风控未覆盖的异常问题。

6.3 第二道护栏：模型安全

从安全数据、算法和评测等多维度体系化构建模型安全的关键技术能力，通过安全对齐提升模型的内生安全。

1. 数据：持续构建包含文本、图像以及图文结合的安全数据集。通过多样化的数据集，华为公司利用监督微调方法来增强其模型的安全性。

2. 算法：采取监督微调算法和强化学习算法，深入分析不同模态理解中对齐训练的不同类型语义泛化问题，并设计了专门的不同模态理解安全算法，旨在使AI系统能够避免已知的人工智能安全漏洞和恶意提示词攻击，从而确保模型在安全层面上实现对齐。

3. 评测：采取人工与自动化相结合的方式进行安全测试。这种混合方法不仅提高了测试的效率，还增强了测试的全面性和准确性。持续评估模型的安全性，同时不断构建和扩充安全数据集，以适应不断变化的安全威胁和挑战，构建安全自动化评测系统，该系统持续评测支持中英语言、多模态交互以及智能体等业务的大模型。

6.4 第三道护栏：模型安全风控

第三道护栏以明确的风险消减为目标，构建大模型安全隐私风控护栏，重点抓好输入端和输出端的风险把控，提供检测防御能力。覆盖已知的多种大模型应用的安全隐私风险，如：内容不合规，隐私泄露、侵犯知识产权、肖像权，信息泄露，无法溯源，恶意代码执行等风险。

1. 安全围栏：基于语义理解构建QA黑白库，持续积累不合规问法库。

2. 语义理解(NLU)：训练风控大模型和安全对话大模型，提高风控大模型的上下文/多轮对话检测能力。

3. 文本审核： PDF/Word/PPT/Excel文档审核能力。

4. 图像检测：构建不合规图像检测能力，持续积累不合规图像标识和元素。

6.5 第四道护栏：应用安全

第四道护栏以消减明确的智能体和应用框架风险为目标，华为公司构建了应用护栏解决方案。

首先，为了保障智能体正常运转，利用任务规划和任务执行护栏解决方案，保障任务规划和任务执行核心功能正常运转。其中，任务规划护栏利用强化学习技术，对大模型编排的任务进行自动化分析，防止生成恶意任务。在任务执行护栏中，对生成代码进行监测，防止大模型生成和执行恶意代码、存在漏洞的代码、调用恶意外部插件、访问恶意网站等。

其次，为了保障应用框架正常运转，利用API越权防护、访问控制、容器隔离、算子安全防护等技术，消减应用框架访问越权等风险。

最后，为了保障应用安全必须加强传统的软件系统的全栈安全工程能力。

6.6 第五道护栏：运营运维安全

在AI系统运营和运维阶段，重点保护大模型系统所涉及的训练数据、模型、生成内容的全生命周期安全，构建大模型恶意行为态势感知系统，提升拦截能力，对现网系统进行统一运营。本文提出需要构建感知系统，通过在线检测和离线检测两种方式拦截恶意行为，不断构建检测能力：

1. 构建在线检测能力库：聚焦高危恶意行为模式的检测方案，构建prompt攻击模式检测、恶意意图检测、恶意行为规则等实时检测能力，建立目标劫持、对抗Tokens反向诱导等prompt攻击模式库，构建内容不合规、侵犯个人隐私肖像等恶意意图模式库。

2. 离线检测能力库：构建集成多种检测方案（规则、分类模型、LLM检测能力）的离线检测系统。

3. 恶意行为检测（用AI保护AI）：在昇腾卡硬件基础上建立攻击检测引擎，可在离线和在线状态下进行恶意行为检测，该引擎利用基于深度神经网络（DNN）的特征检测模型，对推理请求进行特征提取和攻击模式匹配，同时提供物理对抗样本和大模型提示注入等AI攻击感知能力。

4. 可信审计：在算力运行环境中，通过基于密码学中的Merkle树机制将完整性证据保存到硬件安全的区域，构建了AI系统日志可信审计的基础，确保恶意行为者和恶意的运维人员篡改日志文件的记录后，能够高效检测到日志完整性遭破坏，及时预警。

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com

商标声明

、**HUAWEI**、**华为**、 是华为技术有限公司的商标或者注册商标，
本文档提及的所有其他公司的名称和商标均为其各自所有人的财产。

免责声明

您可以为内部参考的目的复制和使用本文件。本文件未授予任何其他许可。

本文件按“原样”提供，不作任何明示或暗示的保证。任何保证均明确予以否认，包括但不限于不侵权、商用性以及对特定目的适用性的保证。华为不负责保证所呈现信息的精确性。本文件提供的任何信息可能会被纠正、修改和改变，恕不另行通知。使用或信赖本文件所提供信息的风险自行承担。本文件提供的所有关于第三方的信息均来源于公开资源或他们发布的报告和报表。

Copyright © 2024 华为技术有限公司　版权所有