人工智能安全风险及治理研究

国家工信安全中心信息政策所 叶晓亮

2021年9月

目　录

01

人工智能安全相关概念及风险分析

（一）人工智能安全相关概念

人工智能是推动新一轮科技革命和产业变革的重要技术，但人工智能决策能力给人们带来无限惊喜的同时也引发了人们对人工智能风险与安全性的思考

人工智能定义及发展

- 对于人工智能的定义，目前业界仍未形成统一意见
- 人工智能经过基础理论及算法研究阶段、专家系统中的初步应用阶段，迎来了第三个高潮期，我国将其当前发展阶段定义为“新一代人工智能”时期，在该时期人工智能正在给人类经济、社会与生活带来颠覆性影响

人工智能风险的界定

针对人工智能的安全性评估，业界往往围绕安全性、消费者基本权利等方面进行分析。本报告将人工智能风险分为三个层次：

- 技术本身脆弱性导致的风险
- 人工智能被攻击者利用产生的风险
- 人工智能随着在各领域广泛应用所产生的风险

人工智能安全的内涵

“人工智能安全”在不同场景下通常代表两种不同含义。在网络安全领域，人工智能安全指的是基于人工智能的网络安全防御技术及方法。在风险分析及应对方面，人工智能安全更多用来描述人工智能技术存在或引发的安全风险及应对能力。本文将聚焦于后一种含义的诠释

（二）人工智能安全风险分析

1.框架及算法等技术要素风险导致人工智能本身的脆弱性

人工智能学习框架易被攻击	每种深度学习框架依赖图像处理、矩阵计算、数据处理、GPU加速等众多基础库和组件，各个基础库和组件潜藏的安全隐患都会威胁学习框架本身以及上层应用的安全性
算法设计局限性仍难以避免	特征描述的局限、目标函数的偏差、计算成本的制约都是可能导致决策偏离预期甚至出现错误结果的原因
模型保密性和稳健性易受威胁	恶意攻击者可利用样本迭代对目标模型进行查询，根据其返回结果构建出相似模型，进而还原模型内部信息，或基于这个相似模型，构造对抗样本，使之做出错误决策
数据“投毒”易导致结果异常	攻击者通过修改样本、删除部分样本或加入精心设计的恶意样本等恶意操作，导致训练出的模型可用性和完整性遭到破坏
数据不均衡易引发数据“偏见”	人工智能决策结果的准确性、客观性很大程度依赖于数据本身。数据本身的分布偏差、技术人员本身对事物的认知的“主观性”，易导致人工智能决策结果往往出现偏见、歧视

（二）人工智能安全风险分析

2.技术两面性导致信息造假等被恶意利用的风险

人工智能技术降低了信息造假的门槛

深度伪造是指利用对抗网络实现对图像、音频及视频的生成或修改形成“虚假信息”。这不仅是对公民肖像权等个人权益的侵犯，若将其用于敲诈勒索、伪造罪证等恶意活动，还会严重危害人际关系、影响社会稳定。基于深度伪造的虚假新闻还可能对社会舆论生态造成恶劣影响，甚至威胁国家安全

人工智能技术被用于升级网络攻击手段

在传统网络攻击中，攻击规模和攻击效率难以兼顾，而人工智能技术的应用能够实现大规模的自动化网络攻击。一方面人工智能能够实现恶意软件编写和分发的自动化，大大提升渗透效率。另一方面基于被感染设备构建智能僵尸网络，利用人工智能技术可实现智能分析和主动攻击

（二）人工智能安全风险分析

2.技术两面性导致信息造假等被恶意利用的风险

人工智能技术降低了信息造假的门槛

深度伪造是指利用对抗网络实现对图像、音频及视频的生成或修改形成“虚假信息”。这不仅是对公民肖像权等个人权益的侵犯，若将其用于敲诈勒索、伪造罪证等恶意活动，还会严重危害人际关系、影响社会稳定。基于深度伪造的虚假新闻还可能对社会舆论生态造成恶劣影响，甚至威胁国家安全

人工智能技术被用于升级网络攻击手段

在传统网络攻击中，攻击规模和攻击效率难以兼顾，而人工智能技术的应用能够实现大规模的自动化网络攻击。一方面人工智能能够实现恶意软件编写和分发的自动化，大大提升渗透效率。另一方面基于被感染设备构建智能僵尸网络，利用人工智能技术可实现智能分析和主动攻击

（二）人工智能安全风险分析

3.不成熟技术广泛应用导致数据泄露及伦理道德等问题

人工智能对数据的大规模需求引发个人隐私问题和数据泄露风险

人工智能技术实现所依靠的数据采集设备，例如摄像头、手机等，已经能够实现数据无感采集，存在未经本人知情同意就进行个人信息采集的可能性，加之其强大的关联分析能力，能够实现人脸信息、身份信息、日常行踪甚至亲属关系的匹配，形成个人信息画像，公民隐私透明化风险尤其突出。同时，相关数据存储不当或遭遇黑客攻击，还将引发大规模信息泄露事件

技术的智能化发展将带来伦理问题及结构性失业等社会风险

由于训练数据或算法的局限性经常会导致系统输出带有偏见或错误的决策结果，例如，亚马逊人脸识别系统将黑皮肤女性错误识别为男性。由此可见，智能系统输出不准确的结果轻则影响社会公平正义，重则危害人身财产安全。此外，智能系统的普及推广将在部分行业实现“机器代人”，人工智能导致的失业问题也将逐步演化为更大的社会安全事件

02

人工智能典型应用场景安全风险分析

（一）人脸识别技术重准确轻安全

目前人脸识别准确率已基本实现了对人类肉眼的超越。作为当前人工智能技术普及程度最高的应用之一，人脸识别技术已广泛应用于公共场所安保、金融业务等场景中，但其尚未完全兼顾功能与安全

错误识别、易被欺骗等安全问题

- 由于当前技术限制和学习数据不足，在识别过程中受到年龄、性别、种族等因素影响易导致识别错误
- 针对人脸识别系统的欺骗手段和技术也逐步升级。除了图片、3D面具等手段外，图层叠加等对抗技术能够以更不易察觉的方式实现对识别结果的颠覆

隐私泄露问题

伴随着大规模人脸信息采集，隐私泄露问题不可避免。很多公共场所都可能存在未经本人知情同意的信息采集行为，海量人脸信息极易成为黑客攻击的目标，加之人脸识别系统的数据存储或处理安全保障措施缺乏，很可能造成大量隐私信息的泄露

（二）自动驾驶系统重体验轻避险

自动驾驶技术对数据高度依赖，使其在数据采集、信息交互、海量数据存储方面都面临安全威胁

1 数据采集方面

- 汽车通过各类传感器采集车速、油门、刹车、车窗、雨刷器等内部信息
- 基于视频采集设备及图像识别技术对交通信号、道路标识、行人及障碍物等道路情况进行识别

2 信息交互方面

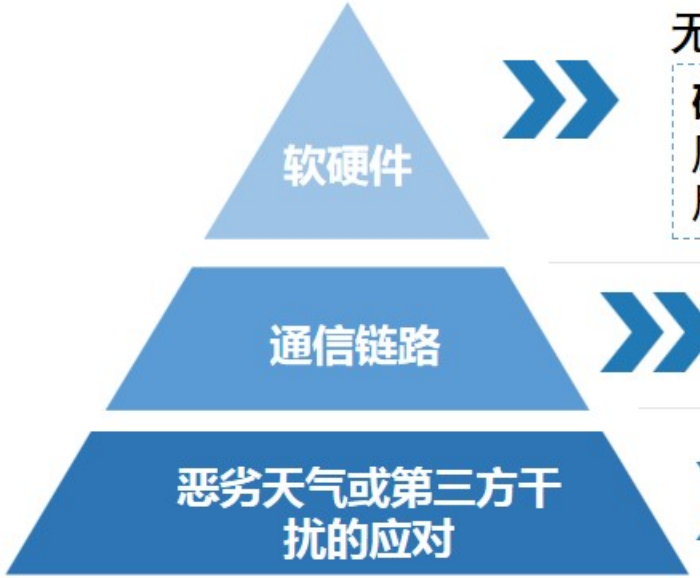
- 自动驾驶通过公共通信网络进行数据传输，数据在通信链路上可被窃听或遭受中间人攻击，甚至还可能被攻击者远程劫持
- 自动驾驶汽车还将面临外部网络生态风险，诸如道路基础设施和智能充电装置等，攻击者可以基础设施为“跳板”进而获取对大规模车辆的控制能力

3 数据存储方面

自动驾驶将采集数据存储在云端，其蕴藏的巨大价值也使得数据被窃取风险凸显。关联用户信息及其行动轨迹将对公民个人隐私造成极大威胁。

（三）无人驾驶飞机重功能轻稳定

无人机必须要达到很高的可靠性与稳定性，才能在更多商业用途中推广应用。虽然为避免坠机等安全事故，无人机一般都设有自主降落的应急预案，但其潜在的安全隐患仍对飞行安全构成威胁



无人机设备软硬件故障是影响其飞行安全的主要安全风险

硬件方面，由于成本、重量和性能等因素，致使元器件、原材料的选用以及冗余技术的应用等受到限制，易造成单点故障。**软件方面**，无人机传感器的故障、视觉识别技术的局限性可能导致无法有效识别障碍物

通信链路的薄弱环节将影响无人机系统的正常运行

通过对通信链路实施网络监听、欺骗攻击、拒绝服务攻击和信号干扰等，能够实现对通信链路的中断、飞行路线的操纵、对无人机的劫持，甚至造成无人机坠毁

无人机对恶劣天气或第三方干扰的应对能力仍然不足

五级以上大风、下雨、下雪、大雾等恶劣天气，以及飞行路线周边的高大建筑，都会对无人机传感器、卫星导航系统信号造成影响和干扰，导致无人机定位效果变差甚至失效

（四）专用型机器人重应用轻防护

智能机器人集合了多传感器信息融合、导航与定位、路径规划等技术，能够实现对多源输入信息的融合分析，根据环境的变化进行自主、动态的调整，具有很强的自适应能力、学习能力和自治功能

工业机器人	家用机器人及医疗机器人
➢ 传感器干扰方面，通过物理手段或网络攻击可实现对导航、避障等功能的干扰 ➢ 系统漏洞方面，机器人控制系统等安全保障不足 ➢ 网络通信方面，通信大多使用Wi-Fi、蓝牙等，缺乏加密机制，通信信息易被监听、窃取与篡改	在智能机器人运行过程中，会对个人信息进行持续、全面的搜集，其收集的信息远不止于文本类信息，还广泛涵盖了声音、图像、视频等信息。一旦机器人出现突发系统故障或遭遇网络攻击，极易出现隐私信息主动或被动泄露事件

03

国内外人工智能安全治理关注重点

（一）探索人工智能安全前沿技术

可解释人工智能框架有助于提升透明度

- 微软可解释机器学习框架InterpretML可为机器学习过程生成高准确性解释
- 脸书神经网络解释决策工具Captum，可帮助技术人员深入研究神经网络
- IBM的可解释AI工具包AI Explainability 360，通过使用可对比解释技术解释AI模型决策结果

模型“杀毒软件”预防对抗样本病毒

- 瑞莱智慧的RealSafe工具可针对AI在极端和强对抗环境下的算法安全性进行检测与加固
- Open AI通过训练人工智能系统相互辩论，以保障决策结果不会产生较大偏差
- 阿里的“AI安全诊断大师”可让AI模型自诞生初始便自带“免疫力”

多样化方法助力人工智能数据隐私保护

- 联邦学习可在多计算节点参与的情况下，实现参与主体平等性及高度独立性
- 差分隐私使得模型不会记录任何特定用户的细节
- 同态加密技术可实现机器学习全流程密文处理

人工智能技术助力网络安全防御系统

- 2021年3月美国人工智能国家安全委员会指出，政府应利用人工智能技术防范网络攻击
- 微软人工智能攻防对抗模拟工具CyberBattleSim，可提供真实的攻防演练环境

（二）开展人工智能应用规范探索

针对人工智能可能存在的安全风险，各个国家和国际组织积极展开约束性规则研制，确保人工智能在设计阶段将安全、信任、公平、道德规范纳入考虑

全球标准化机构积极开展人工智能安全基础共性标准研究

ISO与IEC成立人工智能可信研究组

重点关注可信人工智能标准制定

IEEE发布《人工智能设计的伦理准则》

旨在推动人工智能将人类福祉摆在优先位置

欧盟严格约束人工智能系统的数据应用

《通用数据保护条例》

为AI自动化决策的合法应用设置了严格条件

巴西《通用数据保护法》》

规定AI输入数据应遵循数据质量、透明度与非歧视性等原则

自动驾驶规范性文件突出人身安全保障要求

德国《自动驾驶伦理准则》

美国《自动驾驶法案》

韩国《自动驾驶汽车安全标准》

（三）重视人工智能数据安全防护

数据作为人工智能的重要驱动力，针对其安全性国内外一直积极出台治理举措

- 我国于2017年7月发布《新一代人工智能发展规划》提出须强化数据安全与隐私保护，2019年6月发布的《新一代人工智能治理原则—发展负责任的人工智能》要求人工智能发展应尊重和保护个人隐私

- 美国在2019年6月发布的新版《国家人工智能研究与发展战略计划》中，将确保人工智能系统安全可靠作为八大战略之一

- 欧盟2018年12月发布的《人工智能协调计划》提出必须遵从GDPR的关键原则。2019年4月发布的《可信赖人工智能伦理指南》指出AI系统必须确保隐私和数据安全

- 印度在 2018年6月发布《人工智能国家战略》强调在使用人工智能情况下采取更高标准的隐私保护规范

- 英国在2018年4月发布《产业战略：人工智能部门协议》，首次承诺开发公平、安全的数据共享框架

- 国际互联网协会发布的《人工智能与机器学习政策建议文件》认为，人工智能系统应当根据相关法律来收集、使用、共享和存储数据

（四）聚焦人工智能伦理原则建设

推动人工智能遵守伦理道德原则一直是世界许多国家、地区及国际组织的关注重点

中国

- 《新一代人工智能发展规划》
- 《新一代人工智能治理原则—发展负责任的人工智能》

美国

美国NIST发布《关于人工智能技术和道德标准指导意见》，指出应设计符合伦理的人工智能架构

欧盟

- 2018年将制定人工智能伦理准则作为战略重点
- 2019年发布《人工智能伦理准则》

韩国

2020年11月，韩国发布《国家人工智能伦理标准》，指出人工智能应以人为中心

巴西

2021年4月，巴西发布《国家人工智能发展战略》，强调人工智能须确保人权

国际组织

- 联合国教科文组织的《机器人伦理的报告》
- 电气与电子工程师协会的《人工智能设计的伦理准则》

04

我国人工智能安全发展对策建议

我国人工智能安全发展对策建议

01 加强AI自身及安全技术研究

- 加强人工智能核心关键技术基础研究，提高人工智能学习模型稳定性、成熟度及可解释性
- 开展AI相关安全风险产生机制研究，突破联邦学习、差分隐私、同态加密等人工智能安全防护关键技术的难点问题

02 完善政策法规及伦理道德规范

- 尽快建立AI安全上位法，完善部门规章
- 形成涵盖人工智能技术本身、数据、产品和系统的系列安全标准体系
- 对技术人员、管理人员、用户形成全覆盖的伦理道德规范

03 优化监管体系及监管力度

- 建立健全政府监管治理机制，对现有的监管体系进行完善优化
- 强化行业自律与企业自治，通过行业协会配合政府监管约束市场行为
- 准确把握监管力度

04 建立安全监测评估及应急处置机制

- 加强AI安全检测评估，完善检测评估指标体系及评估办法
- 构建覆盖政府部门、行业监管机构、企业的AI威胁信息共享、应急处置机制，构建人工智能攻防演练平台，开展应急处置演练

05 强化多领域实操性人才深度及广度

- 优化现有培养机制，建设跨学科人才培养体系
- 促进理论研究与实际应用相结合，鼓励企业加强与科研院所的实践合作
- 通过在线培训、在职培训以及日常宣传等方式，提高大众AI安全意识及技能

工信安全

感谢聆听