冯登国，中国科学院院士，网络与信息安全专家，中国科学院软件研究所博士生导师。在 Theor.Comput.Sci、IEEE IT、CCS 等国际重要期刊和会议上发表学术论文 200 余篇，主持研制国际和国家标准 20 余项，出版《网络空间安全概论》《大数据安全与隐私保护》《可信计算理论与实践》等著作，荣获国家科技进步一等奖、国家技术发明二等奖等奖励。曾担任国家高技术研究发展计划信息安全技术主题专家组组长、国家信息化专家咨询委员会委员、国家重点基础研究发展计划项目首席科学家等。

卷首语 Foreword

人工智能已成为网络空间安全发展的关键变量

冯登国

以 ChatGPT 等为代表的人工智能 (AI) 技术不断走进人类社会的生活与生产活动中，并产生了越来越重要的影响。AI 应用给人类生活带来便利、提高生活质量与生产效率的同时，也带来了更加严重的安全问题。AI 安全问题已得到各国政府、产业界和学术界的广泛关注和高度重视，AI 已成为网络空间安全发展的关键变量。

首先，AI 已成为网络空间安全治理的重点对象。美国政府于 2019 年 2 月发布了《美国 AI 倡议》，强调 AI 对传统安全领域的重要意义；美国国家标准技术研究所 (NIST) 于 2023 年 1 月正式公布了《AI 风险管理框架》，对 AI 系统的开发和部署提供指导性建议，降低产生的安全风险，提升 AI 的可信度；美国政府于 2024 年 10 月发布了一份题为《关于巩固和推进美国在 AI 领域全球领导地位：利用 AI 实现国家安全目标，促进 AI 的安全性、可靠性和可信度》的备忘录，标志着美国在 AI 领域的国家战略迈上新台阶。2024 年 8 月 1 日，欧盟《AI 法案》正式生效，该法案注重基于风险来制定监管制度，从而平衡创新发展与安全规范。我国中央网信办等部门于 2023 年 7 月发布了《生成式 AI 服务管理暂行办法》，该办法明确了生成式 AI 服务应当满足的政策、法律、道德规范和约束；全国网络安全标准化技术委员会于 2024 年 9 月发布了《AI 安全治理框架》1.0 版，该框架涉及 AI 安全治理原则、AI 安全治理框架构成、AI 安全风险分类、技术应对措施、综合治理措施、AI 安全开发应用指引等方面。

其次，AI 作为网络空间的组成部分，其自身存在的安全问题必然对网络空间安全产生重要影响。AI 自身存在基础软硬件安全问题，包括硬件安全、芯片安全、软件安全及其产业链供应链安全；数据算法模型安全问题，包括数据的隐私性、可靠性以及算法模型的鲁棒性、公平性、可解释性…… 这些安全问题必然给网络空间带来安全风险，成为影响网络空间安全的重要因素。

再次，AI 在网络空间中的广泛应用，必将导致众多网络空间安全风险。AI 尤其是大模型的广泛应用必将带来针对关键信息基础设施的网络攻击变得更加自动化和智能化，国家重要敏感数据面临更大的泄露风险，深度伪造对国家政治、社会治安、金融秩序、伦理道德的危害加剧，大模型的“幻觉”缺陷导致大量错误信息涌现，AI 武器化带来极强的攻击优势，以及社会安全和伦理等网络空间安全问题。

最后，AI 赋能网络空间安全。AI 赋能攻击技术提升攻击的精准性、效率和成功率。例如，深度学习赋能恶意代码生成可提升恶意代码的免杀和生存能力；攻击者利用深度学习模型可提升识别和打击攻击目标的精准性；AI 赋能僵尸网络攻击可提升规模化和自主化能力；AI 赋能攻击目标侦察可获取更多攻击目标的有用信息；AI 赋能漏洞挖掘过程可提升漏洞挖掘的自动化水平；AI 可实现智能化和自动化的网络渗透；AI 可有效挖掘用户隐私信息。AI 赋能防御技术提升防御的能力和水平。例如，AI 可有效提高威胁检测与响应能力；AI 可克服人性的弱点抵御以人为突破口的攻击。

为了减小 AI 尤其是大模型的应用带来的安全风险，可以分别从政策法规制定、前沿技术探索、创新能力提升等方面展开工作，具体对策建议包括积极制定并实施 AI 伦理原则与框架，加强可信数字内容体系建设，加强隐私保护的模型推理研究，增加 AI 在网络攻防领域中的角色，建立大模型安全理论体系，以及构建 AI 安全测评体系。

©《智能系统学报》编辑部版权所有