前　言

世界主要国家和企业都认识到人工智能的重要性。首先，政府纷纷通过制定人工智能伦理准则、完善法律法规和行业管理等方式开展人工智能安全治理。自 2016 年起，先后有 40 余个国家和地区将推动人工智能发展上升到国家战略高度。特别是在新冠疫情冲击下，越来越多的国家认识到，人工智能对于提升全球竞争力具有关键作用，纷纷深化人工智能战略。其次，政府为企业发展人工智能商业活动提供了重要支持，全球人工智能市场规模未来会出现爆发式增长，预计 2025 年全球人工智能市场规模将达到 6.4 万亿美元。各国在战略层面大致分为两个方面：一是思考如何在已有的领域和产业加速融合，提升效率，促进生产力方面进行挖掘；二是加大投入刺激创新，寻找新的领域或者对现有领域实施颠覆性改变。

人工智能作为引领新一轮科技革命和产业变革的战略性技术，正成为世界主要国家推动科技跨越式发展、实现产业优化升级、赢得全球竞争主动权的重要战略抓手。首先，人工智能可以提高系统的鲁棒性，即即使在处理错误输入时，系统仍能保持其初始假定的稳定配置的能力，这归功于自我测试和自修复软件。第二，人工智能可以加强系统的韧性，即系统通过促进威胁和异常检测来抵御和容忍攻击的能力。第三，人工智能可以用来加强系统的反应，即系统对攻击作出自主反应的能力，识别其他机器的漏洞，并通过决定攻击哪个漏洞和在哪个点上进行战略操作，并发起更积极的反击。

但随着人工智能技术在各个领域的广泛应用，其安全问题也愈发突出。

为贯彻落实国务院《新一代人工智能发展规划》在人工智能安全方面的部署和要求，提升我国人工智能安全保障能力，支撑人工智能技术与实体经济安全融合，国家工业信息安全发展研究中心、深信服科技股份有限公司、中国科学院信息工程研究所、北京瑞莱智慧科技有限公司、北京交通大学、北京双湃智安科技有限公司、国网思极网安科技(北京)有限公司、科大讯飞股份有限公司、北京小桔科技有限公司、普华基础软件股份有限公司共同编制人工智能安全评估白皮书。

本白皮书梳理人工智能安全需求、挑战和威胁，介绍现有人工智能安全测评政策和标准，阐述十大人工智能安全评估技术，并提出人工智能安全评估建议。从技术和行业的角度以实践案例的形式展示应对人工智能安全风险的解决方案，以供相关人工智能技术、产品、服务提供方和应用方参考，提升自身安全风险防控能力，助力人工智能产业健康发展。