CAICT 中国信通院

人工智能安全白皮书

（2018 年）

中国信息通信研究院

安全研究所

2018年9月

版权声明

前　言

人工智能作为引领未来的战略性技术，日益成为驱动经济社会各领域从数字化、网络化向智能化加速跃升的重要引擎。近年来，数据量爆发式增长、计算能力显著性提升、深度学习算法突破性应用，极大地推动了人工智能发展。自动驾驶、智能服务机器人、智能安防、智能投顾等人工智能新产品新业态层出不穷，深刻地改变着人类生产生活，并对人类文明发展和社会进步产生广泛而深远的影响。

然而，技术的进步往往是一把“双刃剑”，人工智能作为一种通用目的技术，为保障国家网络空间安全、提升人类经济社会风险防控能力等方面提供了新手段和新途径。但同时，人工智能在技术转化和应用场景落地过程中，由于技术的不确定性和应用的广泛性，带来冲击网络安全、社会就业、法律伦理等问题，并对国家政治、经济和社会安全带来诸多风险和挑战。世界主要国家都将人工智能安全作为人工智能技术研究和产业化应用的重要组成部分，大力加强对安全风险的前瞻研究和主动预防，积极推动人工智能在安全领域应用，力图在新一轮人工智能发展浪潮中占得先机、赢得主动。

本白皮书从人工智能安全内涵出发，首次归纳提出了人工智能安全体系架构，在系统梳理人工智能安全风险和安全应用情况的基础上，进一步总结了国内外人工智能安全的管理现状，研究提出了我国人工智能安全风险应对与未来发展建议。

目 录

一、 人工智能安全内涵与体系架构..... 1
（一） 人工智能基本概念与发展历程..... 1
（二） 人工智能安全内涵..... 2
（三） 人工智能安全体系架构..... 3
二、 人工智能安全风险分析..... 6
（一） 网络安全风险..... 6
（二） 数据安全风险..... 8
（三） 算法安全风险..... 9
（四） 信息安全风险..... 12
（五） 社会安全风险..... 13
（六） 国家安全风险..... 15
三、 人工智能安全应用情况..... 16
（一） 网络信息安全应用..... 17
（二） 社会公共安全应用..... 20
四、 人工智能安全管理现状..... 23
（一） 主要国家人工智能安全关注重点..... 23
（二） 主要国家人工智能安全法规政策制定情况..... 26
（三） 国内外人工智能安全标准规范制定情况..... 29
（四） 国内外人工智能安全技术手段建设情况..... 31
（五） 国内外人工智能重点应用的安全评估情况..... 33
（六） 国内外人工智能人才队伍建设情况..... 34
（七） 国内外人工智能产业生态培育情况..... 36
五、 人工智能安全发展建议..... 37
（一） 加强自主创新，突破共性关键技术..... 37
（二） 完善法律法规，制定伦理道德规范..... 38

（三） 健全监管体系，引导产业健康发展..... 39
（四） 强化标准引领，构建安全评估体系..... 40
（五） 促进行业协作，推动技术安全应用..... 40
（六） 加大人才培养，提升人员就业技能..... 41
（七） 加强国际交流，应对共有安全风险..... 42
（八） 加大社会宣传，科学处理安全问题..... 43

一、 人工智能安全内涵与体系架构

（一）人工智能基本概念与发展历程

1、人工智能基本概念

计算机之父阿兰·图灵在 1950 年的论文《计算机器与智能》中提出了“机器智能”以及著名的“图灵测试”：如果有超过 30%的测试者不能确定出被测试者是人还是机器，那么这台机器就通过了测试，并被认为具有人类智能。1956 年，在美国达特茅斯会议上，科学家麦卡锡首次提出“人工智能”：人工智能就是要让机器的行为看起来更像人所表现出的智能行为一样。在人工智能概念提出时，科学家主要确定了智能的判别标准和研究目标，而没有回答智能的具体内涵。之后，包括美国的温斯顿[1]、尼尔逊[2]和中国的钟义信[3]等知名学者都对人工智能内涵提出了各自见解，反映人工智能的基本思想和基本内容：研究如何应用计算机模拟人类智能行为的基本理论、方法和技术。但是，由于人工智能概念不断演进，目前未形成统一定义。结合业界专家观点，项目组研究认为，人工智能是利用人为制造来实现智能机器或者机器上的智能系统，模拟、延伸和扩展人类智能，感知环境，获取知识并使用知识获得最佳结果的理论、方法和技术。

2、人工智能发展历程

人工智能发展经历多次低谷，本轮发展呈现加速态势。人工智能自 1956 年诞生至今已有六十多年的历史，在其发展过程中，形成了符号主义、连接主义、行为主义等多个学派，取得了一些里程碑式研

[1]人工智能是计算机科学的一个领域，它主要解决如何使计算机感知、推理和行为等问题。
[2]人工智能是关于知识的学科－－怎样表示知识以及怎样获得知识并使用知识的科学。
[3]人工智能是人类智慧的部分模拟。

究成果。但是，受到各个阶段科学认知水平和信息处理能力限制，人工智能发展经历了多轮潮起潮落，曾多次陷入低谷。进入新世纪以来，随着云计算和大数据技术的发展，为人工智能提供了超强算力和海量数据，另外，以 2006 年深度学习模型的提出为标志，人工智能核心算法取得重大突破并不断优化，与此同时，移动互联网、物联网的发展为人工智能技术落地提供了丰富应用场景。算力、算法、数据和应用场景的共同作用，激发了新一轮人工智能发展浪潮，人工智能技术与产业发展呈现加速态势。

当前人工智能仍处于弱人工智能阶段，主要是面向特定领域的专用智能。从整体发展阶段看，人工智能可划分为弱人工智能、强人工智能和超人工智能三个阶段。弱人工智能擅长于在特定领域、有限规则内模拟和延伸人的智能；强人工智能具有意识、自我和创新思维，能够进行思考、计划、解决问题、抽象思维、理解复杂理念、快速学习和从经验中学习等人类级别智能的工作；超人工智能是在所有领域都大幅超越人类智能的机器智能。虽然人工智能经历了多轮发展，但仍处于弱人工智能阶段，只是处理特定领域问题的专用智能。对于何时能达到甚至是否能达到强人工智能，业界尚未形成共识。

（二）人工智能安全内涵

由于人工智能可以模拟人类智能，实现对人脑的替代，因此，在每一轮人工智能发展浪潮中，尤其是技术兴起时，人们都非常关注人工智能的安全问题和伦理影响。从 1942 年阿西莫夫提出“机器人三大定律”到 2017 年霍金、马斯克参与发布的“阿西洛马人工智能 23

原则”，如何促使人工智能更加安全和道德一直是人类长期思考和不断深化的命题。当前，随着人工智能技术快速发展和产业爆发，人工智能安全越发受到关注。一方面，现阶段人工智能技术不成熟性导致安全风险，包括算法不可解释性、数据强依赖性等技术局限性问题，以及人为恶意应用，可能给网络空间与国家社会带来安全风险；另一方面，人工智能技术可应用于网络安全与公共安全领域，感知、预测、预警信息基础设施和社会经济运行的重大态势，主动决策反应，提升网络防护能力与社会治理能力。

基于以上分析，项目组认为，人工智能安全内涵包含：一是降低人工智能不成熟性以及恶意应用给网络空间和国家社会带来的安全风险；二是推动人工智能在网络安全和公共安全领域深度应用；三是构建人工智能安全管理体系，保障人工智能安全稳步发展。

（三）人工智能安全体系架构

基于对人工智能安全内涵的理解，项目组提出覆盖安全风险、安全应用、安全管理三个维度的人工智能安全体系架构。架构中三个维度彼此独立又相互依存。其中，安全风险是人工智能技术与产业对网络空间安全与国家社会安全造成的负面影响；安全应用则是探讨人工智能技术在网络信息安全领域和社会公共安全领域中的具体应用方向；安全管理从有效管控人工智能安全风险和积极促进人工智能技术在安全领域应用的角度，构建人工智能安全管理体系。

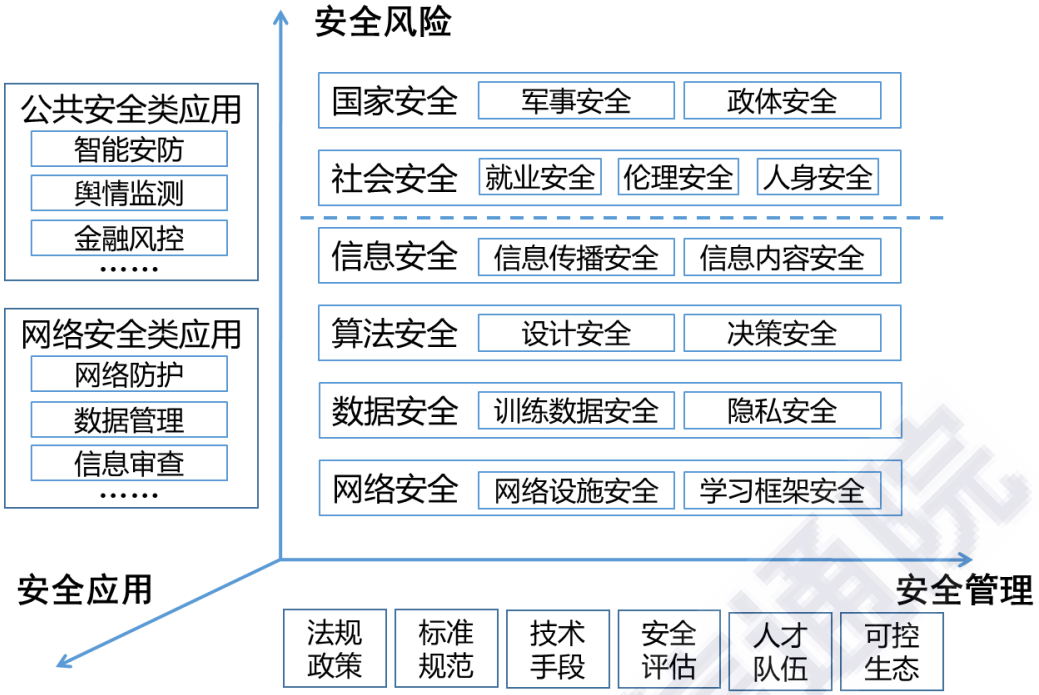

图 1 人工智能安全体系架构图

1、人工智能安全风险

人工智能作为战略性与变革性信息技术，给网络空间安全增加了新的不确定性，人工智能网络空间安全风险包括：网络安全风险、数据安全风险、算法安全风险和信息安全风险。

网络安全风险涉及网络设施和学习框架的漏洞、后门安全问题，以及人工智能技术恶意应用导致的系统网络安全风险。

数据安全风险包括人工智能系统中的训练数据偏差、非授权篡改以及人工智能引发的隐私数据泄露等安全风险。

算法安全风险对应技术层中算法设计、决策相关的安全问题，涉及算法黑箱、算法模型缺陷等安全风险。

信息安全风险主要包括人工智能技术应用于信息传播以及人工智能产品和应用输出的信息内容安全问题。

考虑到人工智能与实体经济的深度融合发展，其在网络空间的安全风险将更加直接地传导到社会经济与国家政治领域。因此，从广义上讲，人工智能安全风险也涉及社会安全风险和国家安全风险。

社会安全风险是指人工智能产业化应用带来的结构性失业、对社会伦理道德的冲击以及可能给个人人身安全带来损害。

国家安全风险是指人工智能在军事作战、社会舆情等领域应用给国家军事安全和政体安全带来的风险隐患。

2、人工智能安全应用

人工智能因其突出的数据分析、知识提取、自主学习、智能决策、自动控制等能力，可在网络防护、数据管理、信息审查、智能安防、金融风控、舆情监测等网络信息安全领域和社会公共安全领域有许多创新性应用。

网络防护应用是指利用人工智能算法开展入侵检测、恶意软件检测、安全态势感知、威胁预警等技术和产品的研发。

数据管理应用是指利用人工智能技术实现对数据分级分类、防泄漏、泄露溯源等数据安全保护目标。

信息审查应用是指利用人工智能技术辅助人类对表现形式多样，数量庞大的网络不良内容进行快速审查。

智能安防应用是指利用人工智能技术推动安防领域从被动防御向主动判断、及时预警的智能化方向发展。

金融风控应用是指利用人工智能技术提升信用评估、风险控制等工作效率和准确度，并协助政府部门进行金融交易监管。

舆情监测应用是指利用人工智能技术加强国家网络舆情监控能力，提升社会治理能力，保障国家安全。

3、人工智能安全管理

结合人工智能安全风险以及在网络空间安全领域中的应用，项目组研究提出包涵法规政策、标准规范、技术手段、安全评估、人才队伍、可控生态六个方面的人工智能安全管理思路。实现有效管控人工智能安全风险、积极促进人工智能技术在安全领域应用的综合目标。

法规政策方面，针对人工智能重点应用领域和突出的安全风险，建立健全相应的安全管理法律法规和管理政策。

标准规范方面，加强人工智能安全要求、安全评估评测等方面的国际、国内和行业标准的制定完善工作。

技术手段方面，建设人工智能安全风险监测预警、态势感知、应急处置等安全管理的技术支撑能力。

安全评估方面，加快人工智能安全评估评测指标、方法、工具和平台的研发，构建第三方安全评估评测能力。

人才队伍方面，加大人工智能人才教育与培养，形成稳定的人才供给和合理的人才梯队，促进人工智能安全持续发展。

可控生态方面，加强人工智能产业生态中薄弱环节的研究与投入，提升产业生态的自我主导能力，保障人工智能安全可控发展。

二、 人工智能安全风险分析

（一） 网络安全风险

人工智能学习框架和组件存在安全漏洞风险，可引发系统安全问

题。目前，国内人工智能产品和应用的研发主要是基于谷歌、微软、亚马逊、脸书、百度等科技巨头发布的人工智能学习框架和组件。但是，由于这些开源框架和组件缺乏严格的测试管理和安全认证，可能存在漏洞和后门等安全风险，一旦被攻击者恶意利用，可危及人工智能产品和应用的完整性和可用性，甚至有可能导致重大财产损失和恶劣社会影响。近年来，国内网络安全企业的研究团队曾屡次发现 TensorFlow、Caffe 等软件框架及其依赖库的安全漏洞，这些漏洞可被攻击者利用进行篡改或窃取人工智能系统数据和信息，导致系统决策错误甚至崩溃。

人工智能技术可提升网络攻击能力，对现有网络安全防护体系构成威胁与挑战。**一是人工智能技术可提升网络攻击效率。**人工智能技术可大幅提高恶意软件编写分发的自动化程度。过去恶意软件的创建在很大程度上由网络犯罪分子人工完成，通过手动编写脚本以组成计算机病毒和木马，并利用 rootkit、密码抓取器和其他工具帮助分发和执行。但人工智能技术可使这些流程自动化，通过插入一部分对抗性样本，绕过安全产品的检测，甚至根据安全产品的检测逻辑，实现恶意软件自动化地在每次迭代中自发更改代码和签名形式，在自动修改代码逃避反病毒产品检测的同时，保证其功能不受影响。2017 年 3 月，首个用机器学习创建恶意软件的案例出现在《为基于 GAN 的黑盒测试产生敌对恶意软件样本》的论文报告中，基于生成性对抗网络(GAN)的算法来产生对抗恶意软件样本，这些样本能绕过基于机器学习的检测系统。2017 年 8 月安全公司 EndGame 发布了可修改恶意软

件绕过检测的人工智能程序，通过该程序进行轻微修改的恶意软件样本即可以 16%的概率绕过安全系统的防御检测。**二是人工智能技术可加剧网络攻击破坏程度。**人工智能技术可生成可扩展攻击的智能僵尸网络。Fortinet 在其发布的 2018 年全球威胁态势预测中表示，人工智能技术未来将被大量应用在蜂巢网络（Hivenet）和机器人集群(Swarmbots)中，利用自我学习能力以前所未有的规模自主攻击脆弱系统。与传统僵尸网络不同的是，利用人工智能技术构建的网络和集群内部能相互通信和交流，并根据共享的本地情报采取行动。被感染设备也将变得更加智能，无需等待僵尸网络控制者发出指令就能自主执行命令，同时自动攻击多个目标，并能大大阻碍被攻击目标自身缓解与响应措施的执行。这在本质上标志着智能 IoT 设备可以被控制对脆弱系统进行规模化、智能化的主动攻击。

（二） 数据安全风险

逆向攻击可导致算法模型内部的数据泄露。人工智能算法能够获取并记录训练数据和运行时采集数据的细节。逆向攻击是利用机器学习系统提供的一些应用程序编程接口（API）来获取系统模型的初步信息，进而通过这些初步信息对模型进行逆向分析，从而获取模型内部的训练数据和运行时采集的数据。例如，Fredrikson 等人在仅能黑盒式访问用于个人药物剂量预测的人工智能算法的情况下，通过某病人的药物剂量就可恢复病人的基因信息[4]；Fredrikson 等人进一步针对人脸识别系统通过使用梯度下降方法实现了对训练数据集中特定面

[4] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: an end- to- end case study of personalized warfarin dosing

部图像的恢复重建[5]。

人工智能技术可加强数据挖掘分析能力，加大隐私泄露风险。人工智能系统可基于其采集到无数个看似不相关的数据片段，通过深度挖掘分析，得到更多与用户隐私相关的信息，识别出个人行为特征甚至性格特征，甚至人工智能系统可以通过对数据的再学习和再推理，导致现行的数据匿名化等安全保护措施无效，个人隐私变得更易被挖掘和暴露。Facebook 数据泄露事件的主角剑桥分析公司通过关联分析的方式获得了海量的美国公民用户信息，包括肤色、性取向、智力水平、性格特征、宗教信仰、政治观点以及酒精、烟草和毒品的使用情况，借此实施各种政治宣传和非法牟利活动。

（三） 算法安全风险

算法设计或实施有误可产生与预期不符甚至伤害性结果。算法的设计和实施有可能无法实现设计者的预设目标，导致决策偏离预期甚至出现伤害性结果。例如，2018 年 3 月，Uber 自动驾驶汽车因机器视觉系统未及时识别出路上突然出现的行人，导致与行人相撞致人死亡。谷歌、斯坦福大学、伯克利大学和 OpenAI 研究机构的学者根据错误产生的阶段将算法模型设计和实施中的安全问题分为三类。**第一类**是设计者为算法定义了错误的目标函数。例如，设计者在设计目标函数时没有充分考虑运行环境的常识性限制条件，导致算法在执行任务时对周围环境造成不良影响。**第二类**是设计者定义了计算成本非常高的目标函数，使得算法在训练和使用阶段无法完全按照目标函数执

[5] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures

行，只能在运行时执行某种低计算成本的替代目标函数，从而无法达到预期的效果或对周围环境造成不良影响。**第三类**是选用的算法模型表达能力有限，不能完全表达实际情况，导致算法在实际使用时面对不同于训练阶段的全新情况可能产生错误的结果。

算法潜藏偏见和歧视，导致决策结果可能存在不公。人工智能算法已应用于个性化推荐、精准广告领域，以及需要进行风险识别和信用评估的信贷、保险、理财等金融领域和犯罪风险评估的司法审判领域，可能产生具有歧视和偏见的决策结果。例如，使用 Northpointe 公司开发的犯罪风险评估算法 COMPAS 时，黑人被错误地评估为具有高犯罪风险的概率两倍于白人[6]。算法歧视主要是由两方面原因造成。**一是**算法在本质上是“以数学方式或者计算机代码表达的意见”，算法的设计目的、模型选择、数据使用等是设计者和开发者的主观选择，设计者和开发者将自身持有的偏见嵌入算法系统。**二是**数据是社会现实的反应，训练数据本身带有歧视性，用这样的数据训练得出的算法模型天然潜藏歧视和偏见。

算法黑箱导致人工智能决策不可解释，引发监督审查困境。当社会运转和人们生活越来越多的受到智能决策支配时，对决策算法进行监督与审查至关重要。但是“算法黑箱”或算法不透明性引发监督审查困境。算法黑箱或算法不透明性主要由三方面原因造成：**一是**拥有决策算法的公司或个人可以对决策算法主张商业秘密或者私人财产，拒绝对外公开。**二是**即使对外公布决策算法源代码，普通公众由于技

[6] 数据来源：ProPublica

术能力不足，也无法理解决策算法的内在逻辑。**三是**由于决策算法本身具有高度复杂性，即使是开发它的程序员也无法解释决策算法做出某个决定的依据和原因。因此，对决策算法进行有效监督与审查是非常困难的。

含有噪声或偏差的训练数据可影响算法模型准确性。目前，人工智能尚处于依托海量数据驱动知识学习的阶段，训练数据的数量和质量是决定人工智能算法模型性能的关键因素之一。在含有较多噪声数据和小样本数据集上训练得到的人工智能算法泛化能力较弱，在面对不同于训练数据集的新场景时，算法准确性和鲁棒性会大幅下降。例如，主流人脸识别系统大多用白种人和黄种人面部图像作为训练数据，在识别黑种人时准确率会有很大下降。MIT 研究员与微软科学家对微软、IBM 和旷世科技三家的人脸识别系统进行测试，发现其针对白人男性的错误率低于 1%，而针对黑人女性的错误率则高达 21%-35%[7]。

对抗样本攻击可诱使算法识别出现误判漏判，产生错误结果。目前，人工智能算法学习得到的只是数据的统计特征或数据间的关联关系，而并未真正获取反映数据本质的特征或数据间的因果关系。对抗攻击就是攻击者利用人工智能算法模型的上述缺陷，在预测/推理阶段，针对运行时输入数据精心制作对抗样本以达到逃避检测、获得非法访问权限等目的的一种攻击方式。常见的对抗样本攻击包括两类，逃避攻击和模仿攻击。**逃避攻击**通过产生一些可以成功地逃避安全系统检测的对抗样本，实现对系统的恶意攻击，给系统的安全性带来严

[7]Joy Buolamwini, Timnit Gebru,《Gender Shades: Intersectional Accuracy Disparities inCommercial Gender Classification》

重威胁，例如，Biggio 研究团队利用梯度法来产生最优化的逃避对抗样本，成功实现对垃圾邮件检测系统和 PDF 文件中的恶意程序检测系统的攻击[8]。**模仿攻击**通过产生特定的对抗样本，使机器学习错误地将人类看起来差距很大的样本错分类为攻击者想要模仿的样本，从而达到获取受模仿者权限的目的，目前主要出现在基于机器学习的图像识别系统和语音识别系统中，例如，Nguyen 等人利用改进的遗传算法产生多个类别图片进化后的最优对抗样本，对谷歌的 AlexNet 和基于 Caffe 架构的 LeNet5 网络进行模仿攻击，从而欺骗 DNN 实现误分类[9]。

（四）信息安全风险

智能推荐算法可加速不良信息的传播。个性化智能推荐融合了人工智能相关算法，依托用户浏览记录、交易信息等数据，对用户兴趣爱好、行为习惯进行分析与预测，根据用户偏好推荐信息内容。当前，个性化智能推荐已经成为解决互联网信息内容过载的一种必要手段。智能推荐一旦被不法分子利用，将使虚假信息、涉黄涉恐、违规言论等不良信息内容的传播更加具有针对性和隐蔽性，在扩大负面影响的同时减少被举报的可能。McAfee 公司表示，犯罪分子将越来越多地利用机器学习来分析大量隐私记录，以识别潜在的易攻击目标人群，通过智能推荐算法投放定制化钓鱼邮件，提升社会工程攻击的精准性。

人工智能技术可制作虚假信息内容，用以实施诈骗等不法活动。在拥有足够训练数据的情况下，人工智能技术可制作媲美原声的人造

[8] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time

[9] Nguyen A M, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images

录音，还可以基于文本描述合成能够以假乱真的图像，或基于二维图片合成三维模型，甚至根据声音片段修改视频内人物表情和嘴部动作，生成口型一致的音视频合成内容。目前，运用人工智能技术合成的图像、音视频等已经达到以假乱真的程度，可被不法分子用来实施诈骗活动。2017 年，我国浙江、湖北等地发生多起犯罪分子利用语音合成技术假扮受害人亲属实施诈骗的案件，造成恶劣社会影响。2018 年 2 月英国剑桥大学等发布的《人工智能的恶意使用：预测、预防和缓解》研究报告预测，未来通过合成语音和视频及多轮次对话的诈骗技术成为可能，基于人工智能的精准诈骗将使人们防不胜防。2018 年 5 月 8 日，谷歌在 I/O 开发者大会上展示的聊天机器人，在与人进行电话互动时对话自然流畅、富有条理，已经完全骗过了人类。

（五） 社会安全风险

人工智能产业化推进将使部分现有就业岗位减少甚至消失，导致结构性失业。人工智能作为公认的第四次工业革命核心驱动力[10]，在其与传统行业相融合的过程中，不再局限于替代人类的手足和体力，而且可以替代人类的大脑，使得重复体力劳动者、简单脑力从业者甚至咨询分析等知识型行业等都可能面临下岗威胁。据 Forrester Research 预测统计，人工智能技术将在 2025 年之前取代美国 7%的工作岗位，其中 16%的美国工人将被人工智能系统取代。《未来简史》作者尤瓦尔·赫拉利预言，二三十年内超过 50%工作会被人工智能取代。如果相关岗位人员不能通过新技能的学习，实现岗位转换，将会

[10] 李开复《人工智能》、王海峰《中国人工智能之路》等

造成大量失业，从而形成严重的社会问题。

人工智能特别是高度自治系统的安全风险可危及人身安全。传统信息系统主要用于个人日常生活和办公辅助等。然而，无人机、自动驾驶汽车、医疗机器人等人工智能产品和系统则在个人生活、工作中可替代人类进行决策和行为操作控制。因此，人工智能安全风险不仅会产生传统信息系统可能造成的数据泄露、影响网络连通性和业务连续性等问题，而且会直接威胁人身安全。自动驾驶、无人机等系统的非正常运行，可能直接危害人类身体健康和生命安全。例如，2016 年 5 月，开启自动驾驶功能的特斯拉汽车无法识别蓝天背景下的白色货车，在美国发生车祸致驾驶员死亡；2017 年年初，我国发生多起无人机干扰致航班紧急迫降事件。

人工智能产品和应用会对现有社会伦理道德体系造成冲击。一是智能系统的决策算法会影响社会公平正义。智能系统由于训练数据或决策算法带有偏见或歧视，其决策结果势必将影响人类社会的公平正义。例如，Kronos 公司的人工智能雇佣辅助系统让少数族裔、女性或者有心理疾病史的人更难找到工作。**二是人工智能应用缺乏道德规范约束，资本逐利本性会导致公众权益受到侵害。**企业具有天生的资本逐利性，在利用用户数据追求自身利益最大化时，往往忽视道德观念，从而损害用户群体的权益。例如：携程、滴滴等基于用户行为数据分析，实现对客户的价格歧视；Facebook 利用人工智能有针对性地向用户投放游戏、赝品甚至虚假交友网站的广告，从中获取巨大利益。**三是人工智能会让人类产生严重依赖，冲击现有人际观念。**例如，

智能伴侣机器人依托个人数据分析，能够更加了解个体心理，贴近用户需求，对人类极度体贴和恭顺，这就会让人类放弃正常的异性交往，严重冲击传统家庭观念。**四是人工智能产品和系统安全事件导致的财产损失、人身伤害等面临无法追责的困境。**人工智能系统在人机协同中可能产生不可预知的结果，造成财产损失或人身伤残。由于人工智能产品和应用自身不具备责任承担能力和法律主体资格，在问题回溯上又存在不可解释环节，这就给现有法律体系和伦理秩序带来严峻挑战。

（六） 国家安全风险

人工智能可用于影响公众政治意识形态，间接威胁国家安全。今年深陷 Facebook 数据泄露丑闻的剑桥分析公司，被多家媒体报道深度参与了 2016 年美国大选。该公司主要采用人工智能技术支撑的广告定向算法、行为分析算法和数据挖掘分析技术支撑的心理分析预测模型辅助进行“竞选战略”，帮助政客确定不同种类的选民在特定问题的立场，指导其在竞选广告中的语言语调等。美国伊隆大学数据科学家奥尔布赖特指出，通过行为追踪识别技术采集海量数据，识别出潜在的投票人，进行虚假新闻的点对点的推送，可有效影响美国大选结果。

人工智能可用于构建新型军事打击力量，直接威胁国家安全。智能武器的应用会使未来战争操控远程化、打击精准化、战域小型化、过程智能化。目前，主要国家都将人工智能作为影响未来世界格局的重要军事变革，纷纷从战略、组织架构、应用等角度加大人工智能在

军事领域的投入，或导致新一轮的军备竞赛。例如，美国国防部明确把人工智能作为第三次“抵消战略”的重要技术支柱。俄罗斯军队于2017 年开始大量列装机器人，计划到 2025 年，无人系统在俄军装备结构中的比例将达到 30%[11]。另外，随着人工智能的快速发展，智能产品价格将会下跌，获取更加容易，恐怖分子将越来越多地使用人工智能武器。例如，2018 年 8 月 4 日，委内瑞拉总统在公开活动中受到无人机炸弹袭击，这是全球首例利用人工智能产品进行的恐怖活动。

以上针对人工智能发展现状，梳理并分析了人工智能安全风险，整体而言，**从风险成因看**，人工智能带来的安全风险是由于其自身技术不成熟性以及技术恶意应用导致；**从发展阶段看，**人工智能安全风险尽管存在于网络空间和国家社会的多个领域，但部分安全问题尚处于前瞻性与苗头性阶段，未真正渗入产业生态环节。当前，人工智能技术发展呈现加速趋势，由于其自身的学科交叉性和垂直应用性，未来必将与传统行业进行深度融合。随着人工智能技术的创新突破和应用场景的日益增多，其安全风险也会动态演进，将越发具有泛在化、场景化、融合化等特点，对人类生产生活、国家政治经济等方方面面产生深远安全影响。

三、 人工智能安全应用情况

目前，人工智能技术由于能够感知、预测、预警关键信息基础设施和经济社会安全运行的重大态势，及时把握群体认知及心理变化，主动决策反应，对保障网络空间安全、有效维护社会稳定具有不可替

[11] 俄罗斯《2025 年先进军用机器人技术装备研发专项综合计划》

代的作用。因此，人工智能在安全领域的应用是当前国内外企业技术和应用创新的重点。结合人工智能安全应用的实践情况看，基于人工智能的网络信息安全应用创新活跃，同时，人工智能与传统社会公共安全的融合应用，也促进了安防监控、金融风控、舆情监测等向智能化发展。

（一） 网络信息安全应用

1、网络安全防护应用

基于人工智能的网络安全防护应用已成为国内外网络安全产业发展的重点方向。随着网络安全向动态防御和主动防御演进，人工智能以其对网络安全威胁的快速识别、反应和自主学习的巨大潜力，成为推进网络安全技术创新的重要引擎。在一定程度上，人工智能技术应用提升了网络防护的自动化与智能化水平，减轻了网络情报分析人员工作量，弥补了网络安全人才不足的现状。**从应用范围看，人工智能在网络安全的应用场景日益广泛。**当前，人工智能已从初期的恶意软件监测广泛应用到入侵检测、态势分析、云防御、反欺诈、物联网安全、移动终端安全、安全运维等诸多领域。例如，在入侵检测方面，以色列 Hexadite 公司利用人工智能来自动分析威胁，迅速识别和解决网络攻击，帮助企业内部安全团队管理和优先处理潜在威胁；我国山石网科公司研发智能防火墙，可基于行为分析技术，帮助客户发现未知网络威胁，能够在攻击的全过程提供防护和检测；在终端安全方面，美国 CrowdStrike 公司基于大数据分析的终端主动防御平台，可以识别移动终端的未知恶意软件，监控企业的数据，侦测零日威胁，

然后形成一套快速响应措施，提高黑客攻击的风险和代价；在安全运维方面，美国的 Jask 公司采用人工智能算法对日志和事件等数据进行优先级排序并逐一分析，以协助安全分析师发现网络中有攻击性的威胁，提高安全运营中心的运营效率。**从应用深度看，人工智能在网络安全的应用程度仍处于前期积累阶段。**除可提升部分网络安全防护产品性能外，基于人工智能技术的网络安全防护体系的创新仍在研究实践阶段。目前看，国外安全企业起步较早，如英国 DarkTrace 公司基于剑桥大学的机器学习和人工智能算法仿生人类免疫系统，致力于实现网络自动自主防御潜在威胁，能够帮助企业快速识别并应对人为制造的网络攻击，同时还能预防基于机器学习的网络攻击。相比之下，国内基于人工智能技术的网络安全防护整体解决方案尚处于研究阶段，对于利用人工智能技术实现整体网络安全防护体系和架构的创新优化仍需探索。

2、信息内容安全审查应用

基于人工智能的信息内容安全审查应用已进入规模化应用的初级阶段。近年来，在基于人工智能技术进行文本、图像和视频识别的应用日益成熟，以及全球信息内容安全管理日趋加强的双轮驱动下，面向违法信息的信息内容安全审查成为了人工智能在安全领域落地应用的前沿领域。美国互联网巨头 Facebook 不仅利用人工智能技术对互联网内容进行标记，而且利用机器学习开发了一款对用户的视频直播内容进行实时监控识别的工具，自动对直播中涉黄、涉暴或者自杀类别的视频内容进行标记。但从效果看，违法内容判定原则仍较为

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62304839

传真：010-62304980

网址：www.caict.ac.cn