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

（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。

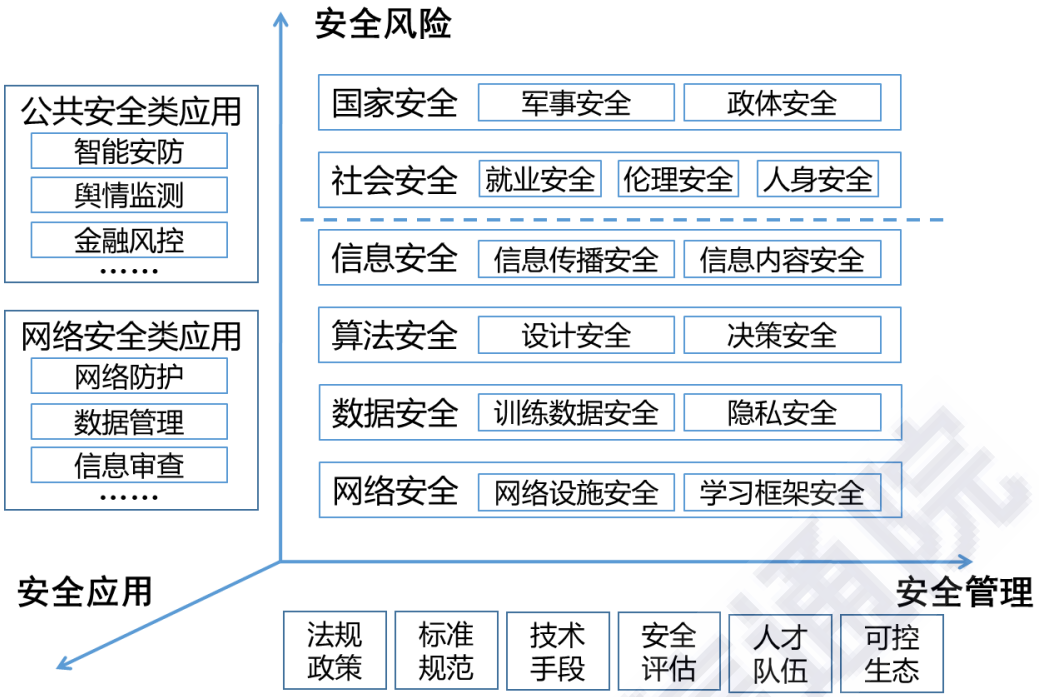

图 1 人工智能安全体系架构图

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

简单，误判情况较多。如对色情内容的识别，主要通过裸露的皮肤来进行判断，使得一些具有历史意义和艺术性的图片被误判。与国外公司相比，国内互联网企业在信息内容安全审查的自动化技术研发和产业化应用起步更早，特别是阿里、腾讯、百度、网易为代表的大型互联网企业，通过基于自身业务安全管理过程中所积累的海量标准样本库，开展对淫秽色情、涉恐涉暴等违法信息识别的建模训练，纷纷推出了基于人工智能的违法信息检测服务。据相关企业调研反馈，企业内部利用人工智能对图片和视频类违法有害信息进行识别的准确率达 99%，语音和文本类的识别准确率也高达 90%。

3、数据安全管理应用

基于人工智能的数据安全管理应用仍处在探索起步阶段。随着数据作为数字经济发展的关键要素，数据价值不断提升，数据安全保护的重要性进一步凸显。除通过人工智能防御网络攻击等外部威胁以避免发生数据泄露外，国内外企业积极探索基于人工智能的数据安全内部威胁防范。目前看，人工智能技术在数据分类、数据防泄漏等领域的应用取得初步成效。德国的 Neokami 公司利用人工智能技术帮助客户保护云端、本地或物理资产上的敏感数据，该公司所研发的数据分类引擎适用于多种业务场景，已被众多合作伙伴厂商所采用，在多家财富 500 强公司中创造价值。国内安全企业天空卫士通过综合运用统计学异常分析、双向循环神经网络等人工智能技术，有效融合数据内容安全和行为识别技术，实现针对企业内部高数据安全风险的人和设备的安全预警和实时控制。

（二） 社会公共安全应用

1、智能安防应用

基于人工智能技术的智能安防呈现全球高速发展的良好态势。传统安防产业主要解决光学器件分辨率、视频数据存储的技术问题，发展存在诸多局限性，例如，传统安防多为被动式应用，用于事后取证，对于事中响应、事前预防作用很小；非结构化的视频数据挖掘深度不够，无法有效利用等。与传统安防不同，基于人工智能的智能安防依托对海量视频数据的学习，可完成行为模式的推断和预测，已经从被动防御向主动判断、及时预警的智能化方向发展，目前已经应用于人脸识别、车辆识别等系统中，进行目标属性提取，实现对目标的智能检测、跟踪及排查。近年来，智能安防产业保持高速增长，已成为人工智能落地应用最好的行业之一，预计到 2020 年，智能安防的全球产业规模将达到 106 亿美元[12]。

国外芯片巨头把握行业发展良机，加紧在智能安防产业链上游布局。美国芯片巨头英特尔早在 2016 年就收购了具有领先技术的计算机视觉公司 Movidius，之后陆续推出多款植入独立神经运算引擎、支持边缘深度学习推断的视觉运算芯片以及神经计算 SDK 开发包，形成平台化设计，为世界范围内各大安防公司提供个性化的解决方案。近年来，英特尔进一步将智能安防作为主要业务增长点进行重点布局，借助公司在高端芯片的优势，积极研究图像处理、智能分析、云端存储等相关技术，不断完善在智能安防产业链上游环节的布局。

[12] 数据来源：中国电子学会

国内智能安防产业发展空间巨大。随着我国平安城市、天网工程、雪亮工程建设的不断推进，安防行业快速发展。“十三五”期间，安防行业正逐步向规模化、自动化、智能化转型升级，预计到 2020 年，安防企业总收入达到 8000 亿元左右，年增长率达到 10%以上[13]。与之相应的国内智能安防从 2016 年开始步入快速发展期，但受限于智能化产品价格偏高、场景应用局限性大等问题，大部分安防企业对人工智能还处在尝试使用阶段，超过 90%的市场份额还是传统安防占据[14]，但随着以公安、交通、金融为代表的社会治理领域进一步驱动智能安防快速应用，未来市场发展空间巨大。然而，必须指出的是，尽管**国内智能安防行业创新能力不断加强，但仍需向产业链上游努力迈进。**目前看，国内安防市场竞争格局以人工智能创新型企业和传统安防巨头两类企业为主。其中，基于人工智能的初创企业如云从科技、商汤科技和旷视科技等，依托在计算机视觉、数据深度分析等方面的技术积累，推出智能安防产品，进行产业布局；传统安防巨头海康威视、大华股份等近年来不断加大研发投入，加强技术创新能力，并且对初创企业进行投资收购，逐步提升安防产品智能化水平。两类企业在竞争与合作中驱动着国内智能安防市场的持续迅猛发展。但是，国内智能安防行业在处理芯片、传感设备等产业链上游环节受制于人，亟需利用行业规模化发展的良好机遇，以应用为牵引，加大行业整合力度，实现关键技术突破，努力向产业链上游迈进。

2、金融风控应用

[13] 《中国安防行业“十三五”(2016-2020 年)发展规划》
[14] 亿欧智库：《2018 年中国 AI+安防行业发展研究报告》

人工智能技术可用于提升金融风控工作效率和准确度。传统金融风控往往是基于评分卡体系，对银行借贷记录等进行建模。“金融+互联网”的发展使得金融业务覆盖更多收入群体，新增群体的显性征信数据往往大量缺失，金融机构不得不更多使用消费数据、运营商数据、互联网行为数据等交易信息进行分析。尽管这种底层数据的改变，对传统信用评分造成了巨大困难，但是给人工智能技术提供了用武之地。例如，面对海量异构数据，基于深度学习的特征生成框架已被成熟运用于风控场景中，对诸如文本、图片、影像等非结构化数据实现了深层特征加工提取，显现出对模型效果超出想象的提升。同时，人工智能可以极大提升基于大数据分析的决策效率，并且可以去除人为判定中的主观看法，使得决策判断会更加准确。

国外发展相对成熟，已应用于金融交易监管。美国 Neurensic 公司利用人工智能监控电子交易，基于机器学习识别对交易公司构成风险的行为，并可自动检测来自实际监管案例的高风险活动。2016 年底，纳斯达克和伦敦证券交易所启用人工智能投入市场监管；2017 上半年，华尔街两家交易所推出智能监管系统。智能监管能够有效降低监管成本和交易风险，提升监管效率，未来具有广泛发展空间。

国内处于起步阶段，仍需长时间的市场验证。国内融 360、好贷网、资信客等金融企业借助对企业市场影响力、产品口碑评价等广泛的数据采集和有效筛选，依托人工智能技术实现了对历史经营数据和实时市场信息的量化建模，进而实现了对各类资产风险的预测评估。但国内智能投顾所面临的市场探索和技术创新仍在摸索阶段，相关的

监管法规政策环境仍在持续变化完善过程中，基于人工智能的金融风控应用仍处于发展萌芽期，其后续发展仍有待观察。

此外，国内外均将人工智能技术应用于网络舆情监测分析领域。美国 911 事件之后，爱国者法案的签署促使美国网络舆情监控的合法化，允许政府监控潜在恐怖分子的所有通讯信息，包括电子邮件和互联网等[15]。美国基于人工智能技术对互联网数据的挖掘分析，加强了网络用户行为分析及预测能力，更好地维护本国国家安全。人工智能可预测网络事件发展方向，加强事件演变预警能力，事前采取舆情干预和引导，规避群体性舆情事件的发生，提升社会治理能力。目前，国内主要网络舆情监测系统均尝试在原有大数据分析基础上，加入自然语言处理、机器阅读理解等相关技术，提升系统的智能化水平。

四、 人工智能安全管理现状

（一） 主要国家人工智能安全关注重点

人工智能作为引领未来的战略性技术，已成为国际竞争的新焦点。世界主要国家把发展人工智能作为提升国家竞争力、维护国家安全的重大战略，加紧出台规划和政策，力图在新一轮国际科技竞争中掌握主导权。世界主要国家基于自身国际地位和发展战略，对人工智能安全的关注重点和重视程度不一。

1、美国：关注人工智能技术对国家安全的影响。

美国凭借其在人才储备、金融体系、IT 技术和互联网的优势，积极谋求在人工智能领域的全球领导地位，重点关注人工智能技术对

[15] 周松青，《中美网络舆情监控法律规制比较研究》

其国际领先地位的影响。2017 年 7 月，哈佛大学肯尼迪政治学院发布《人工智能与国家安全》报告，该报告系统总结了以往颠覆性技术对国家安全的重大影响及应对经验，提出在国家安全领域保持人工智能技术领先并有效管控风险的政策建议。2018 年 3 月 20 日，美国国会发起提案，建议成立“国家人工智能安全委员会”，并将制定“2018 年国家安全委员会人工智能法”。

2、欧盟和英国：关注人工智能对隐私、就业及伦理影响

2016 年底，英国发布《人工智能：未来决策制定的机遇与影响》，报告中关注人工智能对个人隐私、就业以及政府决策可能带来的影响，并就处理人工智能带来的道德和法律风险提出了建议。2018 年 3 月 27 日，欧洲政治战略中心发布《人工智能时代：确立以人为本的欧洲战略》，报告针对人工智能发展过程中可能遇到的劳动者被替代的问题和人工智能偏见的问题，提出了欧盟应该采取的对应策略。

3、俄罗斯、以色列、印度：重点关注人工智能国防领域应用以及对军事安全影响

俄罗斯在人工智能领域方面的成就虽然整体落后于其它科技强国，但在其国防需求的引导和国内工业的全面支持下，在军事无人平台上取得了不少成果。例如，早已被用于叙利亚来清除 ISIS 留在帕米尔的诱杀装置和爆炸装置的 Uran-6 排雷机器人、用于远程侦察和火力支援的 Uran-9 多功能机器人战车以及部署在俄罗斯太平洋舰队的 Platforma-M 侦查无人车等。

以色列在新世纪之后便在网络安防、通信、密码战、无人交通、

军用机械制造、航空航天等领域开展人工智能相关的研发与应用。例如，完全自动驾驶的军用车辆早在 2016 年便用于边境巡逻；能够理解和描述视频的人工智能算法已用于战场和边境线的监控；士兵智能手环可以让指挥官精确理解战场态势，并用人工智能技术分析战场回传数据做出明智决策。

印度政府在今年 5 月 22 日宣布将利用人工智能技术开发武器、防御和监视系统。早在今年 4 月，印度总理莫迪公开表示，人工智能和机器人将成为未来军事力量最重要的决定因素，印度将努力利用人工智能技术提升作战能力。目前，印度军方正在制定人工智能路线图，未来两年，将研究机器学习运用于空军、海军、陆军、网络安全、核、生物资源等领域，涉及自主化武器和无人监视等系统。

4、加拿大、日本、韩国、新加坡：侧重人工智能人才培养、技术研发和产业推进等，对人工智能安全关注较少

加拿大政府 2017 年 3 月推出了《泛加拿大人工智能战略》，将增加加拿大优秀的人工智能研究人员和技术毕业生人数定为战略目标之一。同时，拥有多伦多大学、蒙特利尔大学等人工智能研究重镇的加拿大拥有顶尖水平的研究团队，吸引了大量人工智能人才聚集，政府对人工智能的财政扶持力度大，整体教育质量高，工业体系成熟，拥有发展人工智能的深厚潜力。

日本政府在 2016 年 1 月颁布了《第 5 期科学技术基本计划》中，提出要建立以人工智能为核心的超智能社会 5.0。2017 年 3 月，日本政府制定了人工智能发展的路线图，明确了日本以 2020 年和 2030 年

为时间界限的人工智能发展进程，计划分 3 个阶段推进利用人工智能大幅提高制造业、物流、医疗和护理行业效率的构想。

韩国在人工智能技术方面同样有着深厚的基础。根据韩国信息与通信技术研究所数据显示，韩国在 2005 年 1 月至 2017 年第三季度期间，与人工智能相关的专利数量全球排名第三，仅次于美国和日本。2016 年 8 月，韩国政府提出了以人工智能为首的九大国家战略项目，作为发掘新经济增长动力和提升国民生活质量的新引擎。此外，韩国政府同样设定了人工智能的发展路线图和阶段性目标。

新加坡在 2017 年 5 月发布《新加坡人工智能战略》，计划未来五年投资 1 亿 5 千万美元，用于增强新加坡人工智能技术实力。新加坡结合“智慧国”建设目标，大力推动人工智能技术的产业化应用。目前，新加坡已经非常广泛地采用人工智能技术，至少有六分之一的组织在各种领域运用人工智能，其中信息技术领域高达 60%，供应链和物流 48%，客户支持 49%以及研发部门 41%[16]。

（二）主要国家人工智能安全法规政策制定情况

目前，世界主要国家的人工智能安全管理体系均处于构建初期，主要以思路和建议的形式体现在战略规划和报告中，落地实施的法律法规和管理政策相对较少。部分国家已在人工智能产业推进的先导领域试点制定人工智能应用规范要求，尝试开展安全管理，约束人工智能可能带来的安全风险。

1、美国：主张依托市场力量降低算法决策风险，在技术应用先

[16] 数据来源：希捷科技

导领域加强监管

一是主张发挥市场作用降低算法决策风险。在人工智能技术和产业发展中，美国鼓励企业加大创新，主张依托市场力量降低人工智能安全风险。美国权威科技创新智库数据创新中心在 2018 年 5 月发布《政策制定者如何推动算法问责》的报告中，提出了算法问责框架，旨在不牺牲数字经济发展的前提下控制算法风险的解决方案。报告指出，在大多数情况下，市场力量有能力阻止大部分有缺陷的人工智能算法产生，监管机构无需对算法进行干预；只有当算法应用带来的潜在危害较大，并足以动用监管审查时，才会触发算法问责。

二是对自动驾驶、刑事司法等先导领域加强监管。2017 年 9 月，美国众议院通过了《自动驾驶法案》，对自动驾驶汽车提出了包含系统安全、网络安全、人际交互、防撞性能等在内的 12 项安全要求。2017 年 12 月，美国纽约市通过《关于政府机构使用自动化决策系统的当地法》，对法院、警方等政府机构使用的人工智能自动化决策系统进行安全规制。

2、欧盟与英国：强化政府主导的伦理原则建设和法律法规约束

一是加强人工智能基本伦理道德原则建设。欧洲委员会下辖的欧洲科学与新技术伦理组织在 2018 年 3 月发布的《关于人工智能、机器人及“自主”系统地声明》报告中，提出了一套基于欧盟条约和欧盟基本权利宪章规定的价值观的人工智能基本伦理原则。该原则涵盖了“保障人类尊严”、“安全性、可靠性”、“可追责性”、“可持续性”等多个方面。2018 年 4 月 16 日，英国议会发布《英国人工智能发展

计划、能力与志向》，提出了包括“人工智能应为人类共同利益和福祉服务”、“人工智能应遵循可理解性和公平性原则”、“人工智能不应用于削弱个人、家庭乃至社区的数据权利或隐私”等在内的 5 项人工智能基本道德准则。

二是尝试建立人工智能自动决策应用规范。欧盟在 2018 年 5 月生效的《通用数据保护条例》中为人工智能自动化决策的合法应用规定了极其严格的条件：经用户明确同意，或是用户和数据控制者之间签订、执行合同所必需，又或被欧盟或成员国法案明确授权；不满足上述条件的无人工干预且对个人产生法律影响或者类似重大影响的人工智能自动化决策应用将被禁止。同时，欧盟《通用数据保护条例》明确要求数据控制者在收集数据时向数据主体告知以下信息“人工智能自动化决策的存在、有关自动化决策内部逻辑的有意义信息、对数据主体的重要意义和设想的后果”，并且鼓励数据控制者向数据主体解释某项人工智能自动化决策的具体原因。

3、我国：坚持规划引领和应用规范，探索构建人工智能安全管理体系

我国政府以战略规划为牵引，加大对人工智能安全的政策引导。2017 年 7 月国务院印发《新一代人工智能发展规划》提出：既要加大人工智能研发和应用力度，最大程度发挥人工智能潜力；又要预判人工智能的挑战，协调产业政策、创新政策与社会政策，实现激励发展与合理规制的协调，最大限度防范风险。同年 12 月，工信部印发《促进新一代人工智能产业发展三年行动计划（2018-2020 年）》提

出：完善发展环境，提升安全保障能力，实现产业健康有序发展；建立人工智能网络安全保障体系。随后，上海、北京、浙江、安徽、贵州、江西等省市纷纷结合自身产业发展实际和比较优势，发布专门针对人工智能的相关实施意见，并同步提出加强人工智能信息安全关键技术和网络安全架构等科研攻关、强化数据安全与隐私保护等重点任务。

同时，**围绕人工智能应用的先导领域，相关管理部门抓紧出台规范性引导文件**。例如，在金融领域，2018 年 4 月 28 日，中国人民银行、中国银行保险监督管理委员会、中国证券监督管理委员会、国家外汇管理局联合印发《关于规范金融机构资产管理业务的指导意见》，明确金融机构运用人工智能技术、采用机器人投资顾问开展资产管理业务应当经金融监督管理部门许可，取得相应的投资顾问资质，充分披露信息，报备智能投顾模型的主要参数以及资产配置的主要逻辑，积极防范人工智能应用于金融投资带来的安全风险。

综合看，我国政府部门对人工智能技术和产业发展的政策较为开放，以促进激励为主，积极利用人工智能发展优势，努力成为新一轮技术和产业变革的引领者。然而，如何有效管控人工智能安全风险尚处于摸索阶段，未来仍需进一步做好相关安全管理工作的前瞻研究与战略布局，以务实审慎的态度推进人工智能安全管理工作。

（三） 国内外人工智能安全标准规范制定情况

目前，国际上的人工智能现有标准主要是人工智能技术、应用领域的通用标准，而涉及人工智能安全、伦理、隐私保护等的安全相关

标准，大多仍处于研究阶段。

IEEE 正在开发人工智能伦理道德标准，规范人工智能安全设计。IEEE 标准协会正在努力确保人工智能技术的设计者在其工作中**优先考虑道德**。2017 年 3 月，IEEE 在《IEEE 机器人与自动化》杂志发表了名为“旨在推进人工智能和自治系统的伦理设计的 IEEE 全球倡议书”，倡议通过基于伦理的设计原则和标准帮助人们避免对人工智能技术的恐惧和盲目崇拜，从而推动人工智能技术的创新。目前，IEEE 工作组正在开发 IEEE P7000 系列中涉及道德规范的伦理标准，分别对系统设计中伦理问题、自治系统透明度、系统/软件收集个人信息的伦理问题、消除算法负偏差、儿童和学生数据安全、人工智能代理等方面进行规范。

ISO/IEC 成立人工智能可信研究组，开展人工智能安全标准研究。ISO/IEC JTC 1/SC 42 人工智能分技术委员会于 2017 年 10 月成立，其工作范围是人工智能领域的标准化。ISO/IEC JTC 1/SC 42 第二研究组为可信研究组，其研究范围包括：通过透明度、可验证性、可解释性、可控性等调查建立人工智能系统信任的方法；调查工程陷阱并用其缓解技术和方法评估人工智能系统的典型相关威胁和风险；调查研究实现人工智能系统鲁棒性、弹性、可靠性、准确性、人体健康和生产技术活动的安全性、社会政治性的安全性、隐私性等性能的方法；调查人工智能系统中偏倚的来源类型，以最小化为目标，包括但不限于人工智能系统和人工智能辅助决策的统计偏倚。

我国成立国家人工智能标准化总体组与专家咨询组，加强人工智

能安全标准研制工作。目前，我国人工智能安全标准主要集中在生物特征识别、智能网联汽车等少数应用领域的安全标准，以及大数据安全、隐私保护等支撑类安全标准，而与人工智能自身安全或基础共性相关的安全参考架构、安全评估、伦理设计、安全要求和测评方法等标准还很少。为落实《新一代人工智能发展规划》任务部署，加强人工智能领域标准化工作的统筹协调和系统研究，发挥标准化的支撑性、引领性作用，在相关部委指导下，中国人工智能产业发展联盟、国家人工智能标准化总体组与专家咨询组等行业组织相继成立，加强人工智能安全基础标准研究并继续深化应用领域安全标准化工作。我国通信标准化协会（CCSA）正在研制《人工智能产品、应用及服务的安全评估指南》和《人工智能服务平台安全评估要求》标准。

（四） 国内外人工智能安全技术手段建设情况

为快速及时地规避和防范人工智能安全风险，安全技术手段是安全管理体系中必不可少的组成，包括安全监管手段和安全防护手段等。依托安全监管技术手段，可及时发现人工智能产品和应用中的安全问题，采取应急处置，降低安全问题影响；依托安全防护技术手段，可增强安全防护能力，提升人工智能产品和应用的安全可靠性能。

国内外政府都重视人工智能安全监管技术手段建设工作，相关思路在规划报告中均有体现。2016 年 10 月，英国下议院科学和技术委员会发布《机器人技术和人工智能》报告，呼吁政府应该对人工智能进行监管。英国政府试图从检验和确认、决策系统的透明化、偏见最小化、隐私与知情权、归责制度与责任承担等方面，加强对人工智能

安全性的管控。我国《新一代人工智能发展规划》中提出：建立健全公开透明的人工智能监管体系，实行设计问责和应用监督并重的双层监管结构，实现对人工智能算法设计、产品开发和成果应用等的全流程监管。**目前人工智能安全监管技术手段建设工作主要依托企业自身开展。**由于人工智能尚处在产业化应用的初期，各国政府采取了促进发展为主的激励性政策，尚未形成完备的人工智能安全监管体系，政府主要通过对企业的规范引导和行业自律来加强人工智能安全监管，相关技术手段主要依托企业自身开展建设。在企业的具体监管方式上，主要围绕事前规范、事中监测和事后应急管控展开。例如，大疆科技通过事前环节的实名登记，实现所有无人机的实名使用，确保使用者有据可查；国内互联网信息内容服务企业通过对自动化推荐的智能算法进行事中环节的实时监测，来减少不良信息的传播；谷歌研究给人工智能系统安装“切断开关”，以在应急必要时刻触发其自我终结机制，规避人工智能系统运行中的常规监管手段的失效风险。

人工智能企业和网络安全企业日益重视人工智能安全防护手段建设。随着人工智能技术不断发展，应用不断推进，人工智能产品和系统逐步从实验室走向实际应用，人工智能企业致力于增强自身产品成熟度和可靠性，积极提升人工智能安全防护能力。例如，百度高度重视自动驾驶系统的安全性，Apollo 平台在基于隔离和可信的安全体系下提供了完善的安全框架及系统组件，免受网络入侵，保护用户隐私和汽车信息安全。同时，网络安全企业加大在人工智能软硬件平台、数据安全和算法设计的安全研究，360 安全研究院多次发现人工

智能技术体系架构中的安全漏洞，研究增强人工智能安全防护能力。

（五） 国内外人工智能重点应用的安全评估情况

人工智能产品和应用的安全评估评测涉及面广，各领域应用具有不同的安全评估指标、方法和要求，人工智能安全评估工作尚处于研究探索阶段，未能全面展开。目前，各国人工智能行业主要是围绕人工智能先导应用领域开展安全评估工作，重点包括自动驾驶、智能服务机器人等。

自动驾驶的安全测试验证受到各国高度重视，但未形成统一安全标准和评价体系。2016 年 9 月 20 日，**美国**交通运输部颁布《联邦自动驾驶汽车政策》，全球首次提出系统化的自动驾驶安全监管政策和安全评估要求框架，强调**安全性为第一准则**，要求技术创新必须在安全性能方面提供保障。**德国**通过修改现行的道路交通法以及制定自动驾驶伦理道德标准，明确**安全是自动驾驶的准入前提**，规定自动驾驶汽车要安装“黑匣子”，记录驾驶活动并保障数据安全。2018 年 4 月 12 日，**我国**工信部、公安部、交通部联合发布《智能网联汽车道路测试管理规范（试行）》，对智能网联汽车上路测试的主体、测试驾驶人及测试车辆，测试申请及审核，测试管理，交通违法和事故处理等进行了明确规定。但是，目前各国均缺乏自动驾驶相关安全标准，安全评估评测工作主要依托企业自身开展，各国政策文件更多是约束企业行为，尚没有形成第三方测试认证机构，无法统一开展自动驾驶的安全评估评测工作，未来可能会制约行业发展。

工业机器人相关安全标准较为完备，但智能服务机器人安全标准

体系和评估能力尚待完善。国际主要标准组织中，ISO 机器人检测标准分成两大类：工业机器人与服务机器人，具体由 TC299 承担机器人标准化工作，工作重点在于安全与性能测试标准，其中，ISO/TC299/WG2 完成了服务机器人领域第一个安全标准——ISO 13482：2014 《个人护理机器人的安全要求》；ISO/TC299/WG3 关注工业安全，目前工业机器人的标准已经比较完善，本体安全与集成安全的标准已修订完成。另外，IEC 标准化工作主要由 TC59、TC61、TC62、TC116 技术委员会承担，制定的标准主要涉及家用服务机器人的安全和性能、工业机器人的功能安全和医疗机器人安全等方面。**我国**于 2015 年 3 月，由国家发改委、工信部、国标委、认监委等部门联合指导成立了国家机器人检测与评定中心，逐步开展标准制修订、检测服务、认证服务等工作。2017 年 1 月，在上述四部委指导下，机器人检测认证联盟发布了《家用/商用服务机器人安全及 EMC 认证实施规则》，关注产品的安全性和电磁兼容性。国内外在智能机器人安全评估能力建设方面取得了一些进展，但是目前更多是对机器人的机械结构、电气特性、系统功能等指标进行安全测试，而针对与人工智能相关的信息数据、智能算法、决策模型等安全评估评测能力不足。

（六） 国内外人工智能人才队伍建设情况

人才是技术与产业发展的基石，是行业持续健康发展的前提。只有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，才有可能保障我国人工智能安全健康发展。

我国重视人工智能人才培养，取得较好成果。近年来，我国高校

积极布局人工智能相关学科建设，加大人才培养。截止 2017 年 12 月，全国共有 71 所高校围绕人工智能领域设置 86 个二级学科或交叉学科。2018 年 4 月，为落实《新一代人工智能发展规划》要求，教育部出台《高等学校人工智能创新行动计划》，积极推进人工智能创新行动和学科设置，加大基础研究力度，加强人才队伍建设。之后，清华、南开等多所 985 高校成立人工智能研究院。2018 年 7 月，清华、南大、西交大等 26 所高校联合签署《关于设置人工智能专业建议书》，申请设立人工智能本科专业。经过前期人才队伍建设，我国的人工智能论文总量和高被引论文数量从 2013 年起超过美国位居世界第一，我国在人工智能专利数量上也已经超过美国和日本，成为世界上人工智能专利布局最多的国家[17]。

我国人工智能人才队伍与美英国家相比还存在较大差距。尽管我国近年来我国人工智能人才培养取得较大成果，但是，同美英发达国家相比依旧存在较大差距。在学科建设上，美国、英国等人工智能人才培养体系扎实，研究型人才优势显著，很多著名院校早已设有人工智能相关专业和研究方向。目前，人工智能领域学术能力排在世界前 20 的学校中，美国占据 14 所，排名前八个席位的都为美国所占据[18]。在产业人才上，截止 2017 年 6 月，美国产业人才总量约是中国的两倍，美国 1078 家人工智能企业约有 78000 名员工，中国 592 家公司中约有 39000 位员工[19]。并且，资深人工智能从业者占比与美国差距

[17]清华大学，《2018 中国人工智能发展报告》
[18]腾讯研究院，《2017 全球人工智能人才白皮书》
[19]腾讯研究院，《中美两国人工智能产业发展全面解读》

明显（十年以上从业者占比 38.7%，而美国是 71.5%[20]）。

（七） 国内外人工智能产业生态培育情况

我国人工智能产业生态发展不均衡，基础环节薄弱。目前，我国在人工智能基础层、技术层和应用层均有布局，已经初步具备了相对完整的人工智能产业生态。但是，我国人工智能产业生态发展并不均衡，无法实现生态的可控发展。**应用层环节**，由于我国具备移动互联网市场优势，以及丰富应用场景，为人工智能应用层发展提供了便利条件，国内人工智能产业投资和技术研究主要集中在应用层环节，在自动驾驶、计算机视觉、语音识别等应用领域形成一定优势，甚至部分产业处于国际领先水平。然而，**基础层环节**，受投资周期长、收益风险高影响，我国人工智能基础层发展相对缓慢，缺少重大原创成果，在基础理论、核心算法以及关键设备、高端芯片等方面差距较大，尽管涌现出寒武纪、地平线等优秀创业公司，并已形成成熟产品，但同国外英伟达、谷歌等巨头相比差距明显，尚难以取得市场竞争主导权。

美国产业布局全面领先，主导全球人工智能产业发展。美国谷歌、IBM、微软等巨头企业正在加紧人工智能全产业布局，在人工智能基础层、技术层和应用层处于全面领先地位，尤其是在算法和芯片领域积累了强大的技术创新优势，从而主导全球人工智能产业发展。牛津大学研究报告提出“国家人工智能潜力指数(AIPI)”，中国产业生态位居世界第二，但得分仅为美国的四分之一[21]。

[20]领英，《全球 AI 领域人才报告》
[21]牛津大学，《解密中国 AI 梦》

五、 人工智能安全发展建议

当前，我国国家安全和国际竞争形势更加复杂，必须放眼全球，在国家战略层面对人工智能安全进行系统布局、主动谋划，坚持以加快技术和应用创新为主线，以完善法律道德规范为保障，以监管规范为牵引，大力推进标准建设、行业协同、人才培养、国际交流、宣传教育等工作，全面提升我国人工智能安全的综合能力，牢牢把握人工智能发展新阶段国际竞争的战略主动，打造竞争新优势、开拓发展新空间，有效保障我国网络空间安全和经济社会稳定发展。

（一） 加强自主创新，突破共性关键技术

一是立足自主创新，加大技术引进吸收，突破人工智能关键基础技术。我国人工智能产业目前呈现“倒三角”结构，体现是“重应用、轻基础”，为人工智能发展带来诸多不确定因素。因此，需要从云计算、大数据和机器学习等关键通用技术研究入手，破解基础安全风险。一方面，立足自主，以传感器、智能芯片、基础算法等重点技术安全可控发展为目标，实施重大技术攻关工程；另一方面，加大技术引进，以开放务实的态度开展对外技术合作，实现技术消化吸收和再创新。依托自主创新和技术引进相结合的发展模式，制定人工智能关键技术安全可控发展路线图，解决人工智能基础环节“卡脖子”问题。

二是加大人工智能安全技术研究，提升人工智能安全防护能力。针对人工智能安全研究落后于应用研究的现状，面向人工智能安全问题和风险痛点，引导多方加大投资，大力支持科研院所、人工智能企业和网络安全企业深化人工智能安全攻防技术研究，构建人工智能安

全攻防演练平台，在人工智能基础层、技术层和应用层设计具备自主知识产权的全方位、一体化安全防护技术架构，加快研发安全防护产品，探索并推广人工智能重点应用领域的最佳安全实践，保障人工智能安全技术研究与产业化应用进程同步推进。

（二） 完善法律法规，制定伦理道德规范

一是建立健全现行法律法规，以应对人工智能带来的隐私安全风险和主体责任问题。首先，推进个人信息保护法律法规建设。目前，我国部分现行法律法规中已涉及个人隐私保护相关内容，但是条款较为分散且不能形成完整体系，需加快统一立法，借鉴欧盟《通用数据保护条例》相关条款和实践经验，推进《中华人民共和国个人信息保护法》制定工作，明确个人信息范围，保障用户知情权，强化数据处理者责任，依法处理好数据开放利用和个人隐私保护之间关系，既保障人工智能对数据资源的合理需求，又防止个人信息的过度利用。**其次，**完善现行法律法规，明确问题责任主体。现行法律法规条文对人工智能产品和系统的应用缺乏约束，未明确界定人工智能产品和系统的设计、生产、销售、使用等环节的责任与义务，针对人工智能可能导致的公平公正问题、故障事故和违法违规行为，以及由此造成的财产损失、人身伤害和社会危害，需要加强研究、前瞻立法，在法律层面进行进一步明确的主体约束和责任划分。

二是研究制定伦理道德规范，以适应智能时代人机共生的社会行为模式。由政府引导，推动相关高校、研究机构和企业等成立人工智能伦理研究机构，加强人工智能伦理研究的统筹规划，跟踪人工智能

对社会伦理道德影响和安全风险，构建系统化的伦理道德规范来约束人工智能研究目标和方向，以及系统设计者和开发者的行为，倡导和强化算法伦理，提升人工智能与人类价值观的一致性，避免国家间的人工智能军备竞赛，保障全社会能够广泛共享人工智能创造的经济繁荣，促进人类社会健康发展。

（三） 健全监管体系，引导产业健康发展

一是完善政府监管体制，优化治理组织架构。人工智能与传统行业的日益融合会催生出大量新业态和新模式，相关产业的安全风险变得交织复杂，其监管工作涉及到政府多个管理部门，应根据实际发展情况优化完善政府监管体制，并提升监管技术手段的支撑能力，对智能推荐、自动驾驶、智能服务机器人、智能家居等人工智能应用的先行领域进行安全监管试点。同时，推动行政机构的治理组织架构的适时调整，在保证新技术发展对行业和社会影响处于可控范围的同时，给予工智能技术和产业创新发展的成长空间。

二是约束企业市场行为，强化企业自治责任。在大数据时代，人工智能企业（特别是大型互联网平台）能够获取海量数据，对数据的学习和利用将涉及到个人隐私、公共安全和社会治理等多个层面，因此，企业平台作为数据重要载体，其自身行为的合规合法性更为重要，训练数据和算法决策安全会直接影响用户权益和社会安全。应加大对企业的监管力度，界定政企责任边界，引导企业在重视自身经济效益的同时，强化社会责任意识，加强自律自治，保障数据采集、存储和流转的合法性和安全性，正确应用人工智能技术。

（四） 强化标准引领，构建安全评估体系

一是制定人工智能安全相关标准规范，弥补现有空白。联合研究院所、技术企业以及第三方测评机构，共同推动人工智能安全相关的国家标准、行业标准、联盟标准的制定工作，重点研究人工智能训练算法、决策模型等相关安全技术要求，形成涵盖人工智能产品和系统的网络安全、数据安全、算法安全和应用安全的系列安全标准体系，从而作为人工智能产品安全设计和测试验证的统一参考，提升人工智能产品的安全可靠性能。

二是以安全标准为引领，开展安全评估评测能力建设。以人工智能安全标准为引领，联合研究机构和科技企业共同开展人工智能产品、应用和服务的安全评估评测技术攻关，逐步积累安全检测样例库、测试样本库等知识资源，形成共享数据集，研发测试工具集，构建人工智能安全检测认证的公共服务平台，建立评估专家库和评估机制，实现人工智能安全的评估评测能力，以技术手段为支撑，切实规避人工智能产品和应用的问题缺陷与安全风险。

（五） 促进行业协作，推动技术安全应用

一是促进人工智能企业与网络安全企业协作，提升技术应用深度和产品成熟度。人工智能企业具有机器学习算法相关技术积累，网络安全企业具有漏洞库、案例库等数据资源和安全防护应用场景，应促进人工智能企业和网络安全企业深入合作，通过对现有网络安全知识库资源进行数据分析和特征学习，提升网络安全防护产品的漏洞挖掘、威胁预警、攻击检测等自主防御能力，并在应用场景中进行迭代升级

优化产品成熟度，共同推进人工智能技术在网络信息安全领域的深度应用。

二是促进人工智能企业与公共安全企业协作，扩大技术应用领域，提升社会治理能力。人工智能正逐渐发展成为新的通用技术，推动传统行业向自动化和智能化转型，在公共安全领域，人工智能的计算机视觉、语音识别与合成、自然语言处理等成熟通用技术已开始应用于安防监控、数据侦查、舆情管控等多个领域，应大力促进人工智能企业与传统的公共安全企业之间协同发展，共同发掘融合需求，瞄准行业发展痛点，形成集成解决方案，加速应用场景落地，推动人工智能在智慧公安、智能交通、智慧金融等涉及公共安全领域的广泛应用，提升国家社会治理的智能化水平。

（六）加大人才培养，提升人员就业技能

一是加强人工智能技术和产业人才队伍建设，降低人才短缺对行业发展的制约风险。首先，立足学校教育，大力贯彻落实教育部《高等学校人工智能创新行动计划》等相关文件，在具备条件的高校内增设人工智能相关专业，扩大招生名额，加强专业教育和职业教育，提供具备人工智能思维、技能和人机协作操作能力的人员，资助重点学科大学创建实验室和创新中心，增加研究型人才培养。**其次，**加大企业培养，针对人工智能人才短缺的现状，鼓励人工智能技术企业内部创办培训机构或与学校联合建设实验室，开展技术与应用研究，在实践中培养可用人才。**再次，**加强国外人才引进，制定人才政策引进专项人才，支持高校或企业引进世界一流领军人才；直接在国外设立研

发中心，吸收当地人才为我所用；鼓励行业收购，企业通过资本运作方式，控股或收购国外具备核心技术的公司团队。

二是优化人才培养体系，提升人员就业技能，降低人工智能引发的失业风险。面向人工智能产业发展引发的社会就业岗位变革，**首先，**应动态调整高校、职业学校等专业设置，逐步降低乃至取消可替代专业招生名额，保障在校学生能够学以致用，防止“毕业即失业”现象发生；**其次，**应引导在岗人员树立终身学习理念，完善在职培训和再就业培训体系，通过多元培训更新在岗人员的就业技能，促进在岗人员实现更高质量就业，降低人工智能引发的失业风险以及由此造成的社会冲击。

（七） 加强国际交流，应对共有安全风险

一是加强技术研究合作，解决现阶段人工智能技术瓶颈，促进人工智能成熟发展。针对当前深度学习技术瓶颈，如对抗样本鲁棒性差、不可解释性、不完全信息与不确定环境适应性弱等问题，通过在国外设立研究中心、组织国际技术交流等方式，开展技术研究国际合作，跟踪最新技术成果，共同加强迁移学习、类脑学习等新技术的研究，解决算法黑箱、算法歧视等安全隐患和监管难题，增强人工智能决策的鲁棒性和安全性，促进人工智能由专用智能向通用智能发展。

二是积极参与标准制定，共同应对人工智能带来的安全问题和伦理影响。积极参与 ISO/IEC JTC1 中 SC27、SC42 的数据安全、隐私保护、问题追责和人工智能可信相关标准制定工作，紧密跟踪 IEEE P7000 系列中人工智能安全、伦理相关标准，与全球主要标准化组织

ISO、IEC、ITU、ETSI、NIST 等加强沟通，与先进国家和领军企业建立交流机制，共享治理经验，促进我国人工智能持续、安全发展。同时，世界主要国家政府应建立人工智能发展交流和对话机制，在竞争中寻求合作共赢，制定国际间共同遵守的人工智能伦理道德规范，规避人工智能技术的恶意应用，切实保障人工智能真正造福人类。

（八） 加大社会宣传，科学处理安全问题

一是开展宣传教育，加强安全防范意识。人工智能技术本身是中立的，但是存在人为恶意应用的可能。例如，机器学习可用于个人信息挖掘，加速隐私获取；信息内容合成技术可丰富网络诈骗手段，使其更具迷惑性。因此，针对利用人工智能技术的新型安全事件，需要加强宣传，公示原因，开展全民范围的安全防范措施教育，培养公民的隐私保护意识和防范诈骗能力，降低人工智能恶意应用可能导致的个人财产损失和社会不良影响。

二是强化舆论引导，树立正确发展理念。现阶段人工智能技术发展虽然取得显著成果，但仍存在一些共性问题，其产业应用（如自动驾驶、智能机器人等）很多处于探索试验阶段，存在不成熟性，可能会酿成安全事件。针对当前人工智能技术在产业应用过程中暴露出的安全事件，应加强正向舆论宣传，降低社会焦虑，引导人们正确看待新技术发展过程中的安全问题，为人工智能技术发展和产业推进创造开放、宽松的社会环境。

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