DOI 10.15302/J-SSCAE-2021.03.004

基于人工智能的内容安全发展战略研究

朱世强，王永恒

（之江实验室，杭州 311121）

摘要：内容安全是指对信息内容的保护，以及信息内容符合政治、法律、道德层次的要求。人工智能的发展对内容安全产生了非常重要的影响。本文从基于人工智能的内容安全重大战略需求出发，梳理了国内外的研究现状与发展趋势，凝练了基于人工智能的内容安全的关键技术问题。研究提出，按照三步走的策略构建世界领先的基于人工智能的内容安全体系；在对抗性机器学习、可解释人工智能、混合增强智能、知识驱动的内容安全等方面开展技术创新突破，同时应注重政策法规和监管机制建设；建设面向内容攻防的网络靶场、面向舆情攻防的大规模社会系统模拟装置等内容安全重大基础设施。

关键词：人工智能；内容安全；体系建设

中图分类号：TP309　　**文献标识码：**A

Development of Content Security Based on Artificial Intelligence

Zhu Shiqiang, Wang Yongheng

(Zhejiang Lab, Hangzhou 311121, China)

Abstract: Content security refers to the protection of information content and that the information content meets the requirements at political, legal, and moral levels. The recent development of artificial intelligence (AI) has had a very important impact on content security. In this article, we summarize the research status and development trends of AI-based content security in China and abroad based on the major strategic demand therefor, and presents the key technical issues regarding AI-based content security. This study proposes to build the world's leading AI-based content security system through a three-step strategy. Innovation and breakthroughs should be made in areas such as adversarial machine learning, explainable AI, hybrid enhanced intelligence, and knowledge-driven content security. Meanwhile, the construction of policies, regulations, and regulatory mechanisms should be emphasized. Furthermore, major content security infrastructure such as cyber ranges for content attack and defense and large-scale social system simulation devices for public opinion attack and defense should be established.

Keywords: artificial intelligence (AI); content security; system construction

一、前言

随着移动互联网、数字媒体、人工智能（AI）等技术及应用的飞速发展，内容安全的重要性越来越突出。内容安全的内涵包括两个层面，第一个层面是指对信息内容的保护，如防窃取、防篡改等，这涉及信息内容保密、知识产权保护、信息隐藏和隐私保护等诸多方面；第二个层面是指信息内容符

收稿日期：2021-01-11；**修回日期：**2021-02-20

通讯作者：朱世强，之江实验室教授，主要研究方向为机器人、人工智能；E-mail: zhusq@zhejianglab.com

资助项目：中国工程院咨询项目“新一代人工智能安全与自主可控发展战略研究”（2019-ZD-01）

本刊网址：www.engineering.org.cn/ch/journal/sscae

合政治、法律、道德层次的要求。内容安全从多个方面对国家和社会产生重要影响，从反动、暴恐等影响国家安全的高危内容，到色情、赌博等影响社会民生的风险内容，再到垃圾广告等影响企业业务和个人生活的内容等。同时，虚假新闻、网络谣言等内容造假已经成为影响目前内容安全治理最大的问题之一。从信息审查的角度，数字信息来源多样、内容复杂、数量庞大、传播速度迅猛，对事实核查高效、快速、准确能力的要求越来越高。新型冠状病毒肺炎疫情初期，虚假新闻和网络谣言曾经给疫情防控工作造成严重阻碍。随着当前网络直播的流行，一些不法分子开始利用直播进行各种违法犯罪活动。对海量的直播数据进行实时的内容监管也是当前亟待解决的问题。

近年来，人工智能的快速发展给内容安全带来深刻的影响。基于人工智能的内容安全算法都可能遭受数据样本污染和对抗性算法攻击，从而导致决策错误。基于深度学习的伪造图像、虚假新闻、语音诈骗等内容欺骗技术，已经达到以假乱真的效果。智能推荐算法被不法分子利用，使不良信息的传播更加具有针对性和隐蔽性。另一方面，人工智能的发展也给内容安全带来了新的机遇。人工智能，特别是深度学习和知识图谱等技术的发展，能够有效提高内容鉴别、保护及违规审查等能力，加速将内容安全治理向自动化、智能化、高效化、精准化方向推进。

2019 年 3 月，中国工程院启动了“新一代人工智能安全与自主可控发展战略研究”重大咨询项目，本论文作为其中“基于人工智能的内容安全与攻防战略研究”课题研究成果的学术性展示，从基于人工智能的内容安全重大战略需求出发，梳理了相关关键技术及应用，总结了国内外的研究现状与发展趋势，提出了我国基于人工智能的内容安全发展战略建议。

二、基于人工智能的内容安全关键技术及应用

（一）基于人工智能的内容安全关键技术

1. 基于人工智能的内容伪造与保护

人工智能特别是深度学习的发展给内容伪造提供了极大的便利。深度伪造技术是一种利用人工智能程序和深度学习算法实现视频、音频模拟和伪造的技术 [1,2]。深度伪造涉及的技术主要有自编码器 [3] 及生成式对抗网络 [4] 等。目前，深度伪造技术不仅能伪造人脸，更可以模拟真人声音及创造出实际不存在的人物图像。结合基于人工智能的自然语言生成技术及社交网络传播，深度伪造大幅度促进了虚假新闻的发展 [5]。这种具有数字自动化特征的深度伪造技术，借助各类媒体传播虚假信息，具有极强的传播势能，可实现大规模、潜伏性的政治操纵和控制，因而大大加剧网络空间带来的政治安全威胁的影响力和对抗复杂性 [6]。

与内容伪造技术相对应，近期涌现出大量虚假内容检测技术。在基于人工智能的深度伪造内容检测方面，特征提取主要有生成式对抗网络流水线（GAN-Pipeline）、深度学习、隐藏分析（Steganalysis）等技术，分类器主要有支持向量机（SVM）、卷积神经网络（CNN）等 [7]。在基于人工智能的虚假新闻检测方面，有基于知识库、写作风格、传播特性、发源地等多种方法，涉及深度学习、知识库、图数据挖掘等多项技术 [8]。当前在新类型虚假新闻检测、虚假新闻早期检测、跨域检测及可解释检测等方面还存在很大挑战。

2. 面向内容分析的人工智能模型与算法安全

基于人工智能的内容分析涉及文本、图像、视音频处理的各类机器学习模型和算法，这些模型和算法本身的安全性对内容安全有至关重要的影响。机器学习模型和算法安全主要涉及以下几个方面 [9]。

（1）投毒攻击及防御。投毒攻击的方法是在训练模型时有意污染训练数据，从而破坏模型的可用性和完整性。训练数据的污染一般通过注入一些精心伪造的恶意数据样本来实现。

（2）后门攻击及防御。后门攻击以数据和模型两种方式在神经网络模型中植入后门，当模型得到特定输入时被触发，然后导致神经网络产生错误输出，因此非常隐蔽不容易被发现。

（3）对抗攻击及防御。机器学习模型和神经网络模型很容易受到对抗样本的影响。通过对原始样本添加特定的扰动，可以使分类模型对新构造的样本产生错误的分类判断。

（4）模型窃取及防御。模型窃取技术是指通过黑盒探测来窃取模型或者恢复训练数据成员，比如

窃取股票市场预测模型和垃圾邮件过滤模型。

目前每年都有大量新的机器学习算法出现，这些算法的安全已经成为普遍关注的问题。模型与算法安全问题可以看作防御和攻击方在信息缺失情况下对对方进行建模的技术之间的博弈。在新型鲁棒性模型及训练算法、多学习器安全问题、信息缺失下的系统建模和推理问题等方面还有待深入研究。

3. 面向内容分析的可解释人工智能

以深度学习为代表的人工智能技术面临可解释性难题，将其应用于敏感领域内容分析时，缺乏透明度和可理解性的“黑盒”算法很难获得人们的安全感和信任感。

人工智能模型的可解释性研究主要有三个方向 [10]：①深度解释，即采用新的深度学习模型去学习可用于解释的特征。很多相关工作与可视化技术相结合，提供更直观的解释。②可解释模型。传统的贝叶斯、决策树等模型具有很好的可解释性。当前也有很多研究者针对深度学习模型进行改进，使其具备更好的可解释性。③模型推理。这种方法把机器学习模型看作一个黑盒，通过大量实验在外部建立一个新的可解释模型。一种新型的研究方法是构建一套机器学习技术，该技术能够自动生成可解释模型，并且保持较高的学习效率。

尽管模型可解释性研究已取得一些瞩目的研究成果，但其研究还处于初级阶段，依然面临着许多挑战且存在许多的关键问题尚待解决。其中，可解释性研究当前面临的一个挑战是如何设计更精确、更友好的解释方法，消除解释结果与模型真实行为之间的不一致；另一个挑战是如何设计更科学、更统一的可解释性评估指标，以评估可解释方法的解释性能和安全性 [11]。

（二）基于人工智能的内容安全重要应用

1. 网络舆情分析与监管

舆论是“社会的皮肤”，是反映社会形势的晴雨表。大数据及人工智能技术为舆情分析和研判提供了全新的资源、方法与范式 [12]。网络舆论的内容复杂，且对舆情分析实时性的要求也比较高。人工智能技术能够使网络舆情分析更高效和准确，可大幅减少人工工作的成本。

近年来基于人工智能的网络舆情分析与监管获得广泛应用。百度公司的媒体舆情分析工具面向传统媒体和新媒体行业，针对内容生产、观点及传播分析、运营数据展示等业务场景，提供舆情分析能力。其政务舆情分析工具依托网页内容挖掘能力与中文语义分析技术支持国内外风险情报的深度挖掘及城市公众舆情态势的实时感知。腾讯计算机系统有限公司的 WeTest 舆情监控工具，通过分布式爬虫 7×24 h 抓取主流应用市场（应用宝等）评论评星和主流论坛（百度贴吧等）的用户发帖讨论，并智能汇总用户评论，进行智能分类。通过情感分析加情感维度提取技术，智能分析并定位到具体问题。2020 年中国信息通信研究院等单位发起的人工智能产业发展联盟发布了《人工智能助力新冠疫情防控调研报告》，报告中指出人工智能和大数据在新冠舆情分析中发挥了重要作用。

2. 多媒体内容分析与审核

即使是挪威、日本、意大利等标榜自由的国家，互联网内容审查也在加强力度。密歇根大学的一个团队使用其开发的“审查星球”（Censored Planet）工具（2018 年启动的自动审查跟踪系统），在过去的 20 个月中，从 221 个国家收集了超过 210 亿次内容审查的测量数据 [13]。近期多媒体数据特别是视频数据得到前所未有的增长并将持续这种增长趋势。海量的多媒体数据远远超出了人类的处理能力，基于人工智能的内容分析和审核获得了广泛应用。

基于人工智能的多媒体内容分析主要包括智能审核、内容理解、版权保护、智能编辑等。其中内容审核功能包括鉴黄、暴恐涉政识别、广告二维码识别、无意义直播识别等，利用识别能力对网络上没意义和不健康的内容进行排查和处理。内容理解功能包括内容分类、标签，人物识别、语音识别，同时也包括对图像和视频中的文字进行识别。版权保护功能包括内容相似性 、同源内容检索和音视频指纹等功能。内容编辑层面可以实现视频首图、视频摘要、视频亮点的生成，同时支持新闻拆条。

目前短视频、图片等成为多媒体审核的主要内容。基于海量标注数据和深度学习算法，可以从多维度精准识别多媒体内容中的违禁内容，如色情、暴恐等。2019 年阿里巴巴集团推出“人工智能谣言粉碎机”支持对新闻内容可信度的智能判别，在特定场景中的准确率已达到 81%。中国信息通信研究院初步实现基于人工智能技术的不良信息检测能力，支持对淫秽色情、涉恐涉暴等违法信息的识别，

识别准确率比传统方式提升了 17%，达到 97% 以上，识别速度达到传统方式的 110 倍。2021 年 2 月百度发布了《2020 年信息安全综合治理年报》。百度内容安全中心在 2020 年利用人工智能技术挖掘各类有害信息共 515.4 亿余条，通过人工自主巡查打击各类相关有害信息 8000 万余条。大幅提升审核速度，并制定了暴恐、政治敏感、水印、标签、公众人物、恶意图像等 6 个审核维度。

三、基于人工智能的内容安全国内外发展现状

（一）国外发展现状及最新进展

1. 美国发展现状分析

作为内容产业最为发达的国家，美国在内容安全与攻防战略方面主要有以下几个特点：第一，针对日益严峻的国际形势以及国内意识形态安全需要，对互联网内容产业加大监管力度，尤其是关于歧视，偏见等内容；第二，高度重视人工智能算法在内容安全方面的应用，Google 等公司与政府合作密切，政府对算法的安全性提出相关的审查要求；第三，美国国会参众两院高度重视利用人工智能算法进行内容造假问题，并通过召开听证会或提出相关法案的方式，将内容造假问题纳入立法程序。

2019 年，美国政府发布了《美国人工智能倡议》，其中强调人工智能对传统安全领域的重要意义，通过人工智能来确保美国的领先地位以应对来自“战略竞争者和外国对手”的挑战。2021 年 3 月，美国人工智能国家安全委员会（NSCAI）发布最终报告：积极维持美国在人工智能领域的统治地位，报告系统论述了美国如何在人工智能激烈竞争的时代赢得主动，维持全球领导地位，并详细阐述了联邦各机构今后改革的行动路线。美国国防部高级研究计划局（DARPA）于 2017 年启动了“算法战跨部门小组计划”（即“Maven 计划”），将机器学习算法集成到情报收集中。DARPA 还推动了其他和人工智能内容安全相关的研究，包括媒体取证、可解释人工智能等。在基于人工智能的内容安全技术方面，美国也处于全球绝对领先的位置。美国人工智能论文引文影响力、专利数量、企业数量和融资规模等指标都居全球第一。

2. 欧洲国家发展现状分析

欧洲国家在人工智能安全方面更注重伦理。欧盟委员会于 2019 年 4 月发布了人工智能道德准则《可信赖 AI 的伦理准则》，提出了实现可信赖人工智能全生命周期的框架。框架提出了实现可信赖 AI 的七个关键要素：人的能动性和监督，技术鲁棒性和安全性，隐私和数据管理，透明性，多样性非歧视性和公平性，社会和环境福祉，问责。

准则中特别强调了隐私和数据管理，要求隐私和数据保护必须贯穿人工智能系统的整个生命周期。人工智能系统使用的算法和数据应该更具透明性且可追溯。在 AI 系统对人类造成重大影响时，人工智能系统能够对决策过程进行合理解释，同时避免不公平的歧视。英国、法国、德国、俄罗斯等欧洲国家都出台了人工智能安全的纲领性文件，都涉及到了人工智能内容安全的策略和发展规划。

3. 日本发展现状分析

在日本的互联网内容产业中，最为发达的是游戏、动漫、音乐等产业，因此，在人工智能时代下，日本的内容安全战略主要针对人工智能生产内容的知识产权归属问题，并推进相关具体法案的推出。2018 年 12 月，日本内阁府发布《以人类为中心的人工智能社会原则》的报告，是迄今为止日本为推进人工智能发展发布的最高级别的政策文件，从宏观和伦理角度表明了日本政府发展人工智能的立场。

在报告中和人工智能内容安全相关的部分，主要强调了保护隐私和保障安全两部分内容。在保护隐私原则中，人工智能能够依据个人行动等数据高精度推断其政治立场、经济状况、兴趣爱好等。在保障安全原则中，报告强调人工智能系统需要把握利益与风险之间的平衡。报告建议深入研究人工智能风险及降低风险的方法，重视人工智能使用的可持续性。

4. 其他国家发展现状分析

除了美国、欧洲、日本、中国等人工智能发达国家和地区，很多其他国家也纷纷推出了人工智能安全的国家战略，而内容安全是其中的一部分。2018 年 6 月，印度发布了《人工智能国家战略》，其中提出利用人工智能促进经济增长和提升社会包容性，并试图寻求一个适用于发展中国家的人工智能发展战略。2019 年 11 月，以色列透露其国家级

人工智能计划，对人工智能的发展、如何满足政府军方的需求及安全性等进行了规划，主要目标是确定教育系统和学术机构能够提供足够的人工智能工程师，以满足政府、国防军事和产业界的人力需求。加拿大政府近年来在人工智能研究和开发方面加大了投入，致力于形成一个极其丰富的人工智能生态系统，包括多个专门研究机构和数千名人工智能研究人员。研究人员对面向内容安全的多媒体分析、可解释人工智能等技术进行了广泛研究。

（二）国内现状及最新进展

1. 国家战略与技术研究

我国高度重视人工智能产业的发展，2017 年，国务院发布《新一代人工智能发展规划》，作为新一轮产业变革的核心驱动力和引领未来发展的战略技术，对人工智能产业进行战略部署。《中国新一代人工智能发展报告 2019》显示，中国人工智能论文发文量居全球首位，企业数量、融资规模居全球第二。2018 年 1 月，我国成立了国家人工智能标准化总体组和专家咨询组，并发布了《人工智能标准化白皮书（2018 版）》，提出建立统一完善的标准体系。2019 年，我国成立了中国人工智能学会人工智能与安全专委会，为解决网络空间安全面临的挑战性问题提供了新的途径。2020 年 3 月国家正式施行《网络信息内容生态治理规定》，以网络信息内容为主要治理对象，以建立健全网络综合治理体系、营造清朗的网络空间、建设良好的网络生态为目标，抵制和处置违法和不良信息。中国科学院专门成立了信息内容安全技术国家工程实验室，围绕国家网络信息安全的重大需求，开展基础理论和网络信息获取、分析及挖掘等核心关键技术研究。

2. 新兴产业的创新发展

目前，我国各地政府紧跟国家步伐，分别根据当地经济发展状况，结合国家内容安全与人工智能相关政策，提出相关的人工智能发展行动计划，加强人工智能监管力度，并推动人工智能在内容安全方面的深度应用。

商汤科技开发有限公司自主研发的原创深度学习平台 SenseParrots，已经在人脸识别、图像识别、视频分析、无人驾驶、医疗影像识别等应用层技术落地，为基于人工智能的内容安全提供了技术支撑。腾讯科技有限公司注重网络安全能力建设，设立了七大网络安全实验室，专注安全技术研究以及安全攻防体系搭建。利用腾讯优图的 DeepEye 识别技术引擎，对内容进行置信度分析，依托腾讯社交的海量样本优势进行深度识别训练，并基于多模型匹配技术进行文本识别，助力内容安全。阿里云平台采用自然语言理解算法识别文本垃圾和恶意行为，采用深度学习算法结合独有的情报、舆情、预警和分析体系及实时更新的样本图库，快速定位敏感信息。华为技术有限公司在机器学习算法安全方面做了大量研究工作，对数据投毒、模型窃取、后门攻击等模型和算法攻击给出了解决方案，为基于人工智能的内容分析模型和算法提供安全保障。此外，我国还有一些人工智能创新企业，在将人工智能应用到内容安全方面做了大量的研究工作。

3. 人工智能 2.0 时代的新发展

潘云鹤院士提出人工智能 2.0，其主要特征是通过大数据和群体智能，拓展、管理和重组人类的知识，为经济和社会的发展提供建议，在越来越多专门领域的博弈、识别、控制和预测中达到甚至超过人类的能力 [14]。在人工智能 2.0 时代，首先是和人机交互紧密结合；其次是和大数据云计算的结合，大数据和云计算是人工智能发展的重要推动力；再次是人工智能和智能监控的紧密结合；最后是人工智能和先进制造的结合。

人工智能 2.0 时代的大数据智能、跨媒体智能等将对内容安全产生重要影响。基于多模态数据的深度融合、知识库、跨媒体分析与推理，结合类脑计算、群体智能等技术，能够实现更智能、更精准的内容分析。我国以阿里巴巴集团、腾讯科技有限公司、百度集团、华为技术有限公司等为代表的科技企业，已经开始积极探索人工智能 2.0 的相关技术及在内容安全治理方面的应用。如百度内容安全中心的“网络生态治理 2019”专项行动，综合运用了自然语言甄别、音视频智能识别、内容智能挖掘等多种人工智能技术。

四、我国基于人工智能的内容安全发展建议

（一）总体发展战略

总体上采取“三步走”的发展战略：到 2025 年基于人工智能的内容安全发展初见成效，到 2035 年实现世界一流水平，到 2050 年达到世界领

先水平。

到 2025 年，人工智能内容安全的发展环境和基础设施基本完善，基于内容安全的重点前沿理论和应用技术进步明显，在内容安全人工智能模型与算法研究方面取得初步成效，人工智能内容攻防关键技术研究取得关键性突破，人工智能内容安全领域涌现一批优秀企业，集聚一批安全领域的领军人才和专家，面向个人、企业和国家的三级人工智能内容安全体系基本建成。

到 2035 年，人工智能内容安全的发展环境优势明显，国家大力投入基础设施建设，规模和水平并肩世界第一梯队，基于内容安全的理论研究，学术水平进入世界一流行列，在人工智能安全领域提出创新理论和方法，人工智能内容安全审核机制成型完善，人工智能内容安全应用技术广泛应用，在内容安全人工智能模型与算法研究和内容攻防关键技术研究上达到世界一流水平。

到 2050 年，在人工智能内容安全领域实现基础层、技术层和应用层的全面世界领先，总体创新能力和理论体系达到国际领先水平，拥有一批主导人工智能内容安全发展潮流的国际顶级专家，在网络空间内容安全领域拥有一批专业的，稳定的具备世界领先水平人才队伍和创新创业团队。人工智能内容安全法规伦理规范和政策体系完整完善，人工智能安全评估和管控能力均居于世界领先水平。

（二）内容安全发展政策保障

1. 政府主导人工智能发展路线

国家在维护网络信息安全方面所处的高度和发挥的作用是无可比拟的，在规范互联网行为、打击网络犯罪等模式治理方面有着先天的优势。政府要主导人工智能安全的发展战略，将人工智能安全上升到国家高度，作为国家未来发展的核心竞争力，同时把内容安全作为人工智能安全的一个重要部分。在人工智能应用中深入分析内容安全需求，强化顶层设计，提出基于国家网络空间安全的总体规划，建立内容安全治理的实施细则，建立安全准入制度和检测评估方法、机制。以企业为主体推进人工智能内容安全发展，政府在法律法规、安全风险、政策指导、资源配置、行业准则等方面提供保障，制定分阶段发展战略，目标清晰，从科研立项、智能经济到智能社会全面布局，加强指导性和执行力。

2. 建立健全合法有效的监管机制

制定人工智能内容安全风险管控制度，从系统安全、算法安全、应用安全多层次制定安全防护措施。保障用户的数据安全，避免算法设计对公众产生的危害，明晰算法动机和可解释性，克服算法设计和数据收集引发的不公正影响。内容安全风险管控制度中针对社交网络、短视频、线上直播等重点应用的内容安全进行详细规定。通过建立可审查、可回溯、可推演的监管机制，确保目标功能和技术实现的安全统一。建立人工智能数据安全监督机制。依照国家法律法规，政府部门针对数据过度采集、数据偏见歧视、数据资源滥用等人工智能数据安全风险，通过线上线下多种方式实施监督检查，及时发现和防范安全隐患。

3. 构建人工智能内容安全标准体系

优化我国人工智能内容安全标准化组织建设，促进国家、行业和团体标准化组织联合有序推进人工智能内容安全标准出台。在人工智能产品、应用和服务等多个环节，制定内容安全检测评估方法和指标体系，通过检测评估强化内容安全与隐私保护。按照内容安全承载模式分类，建立图形 / 图像内容、文本内容、视频内容、音频内容安全指标体系，按照内容安全行为模式分类，建立智能鉴黄、暴恐涉政识别、敏感人脸识别、不良场景识别、广告识别过滤、Logo 识别、反垃圾等安全指标体系。

（三）开展面向内容安全的人工智能技术创新

新形势下的内容安全面临巨大挑战，需要在技术层面进行创新突破，主要包括以下方面。

1. 人机协同的混合增强智能

在内容安全方面，很多时候当前的人工智能还无法独立完成任务，如在视频直播中识别非法活动。因此需要在人机协同、脑机协作、认知计算等技术上进行创新突破，充分融合人类智能和机器智能，实现人工智能的增强，同时基于人类的指导和反馈实现人工智能的持续改进。

2. 知识驱动的内容安全

基于人工智能的内容安全复杂应用需要知识的辅助，因此需要大力推进知识驱动的内容安全创新。技术方向包括跨媒体知识获取、内容安全知识库构建、大规模知识库的管理及知识演化、面向内容安

全的知识推理等。

3. 高性能内容安全分析

有害内容一旦传播出去，可能会造成国家和社会的重大损失，因此很多内容监管应用要强调实时性。需要研究高性能的内容分析算法，特别是在视频直播这样的场景下，需要处理海量视频数据流，同时需要关联多通道的历史数据和知识。

4. 对抗性机器学习

对抗性机器学习直接影响到人工智能的模型和算法安全，从而直接威胁到内容安全。需要在数据投毒的防御、决策时攻击的防御、深度学习模型和算法的鲁棒性等方面进行技术创新。

5. 可解释人工智能

人工智能模型和算法的可解释性直接影响内容安全分析和监管应用的可信度。需要在可解释机器学习模型、基于深度学习和可视化的模型解释、基于推理的模型解释等方面进行技术创新。

（四）完善内容安全基础设施

以技术创新促进基于人工智能的内容安全发展，需要建立和完善一批国家重大基础设施来满足新技术实验需要，以及监管政策和策略评估的需要。

1. 面向内容攻防演练与研究的网络靶场

构建大规模、开放式、共享式、增长式的国家级内容安全网络靶场，为用户提供内容安全攻防体系验证、应用系统及安全产品安全性检测、风险评估及应急响应等高端服务，创新突破复杂网络属性、行为、交互式动态等高度仿真，复杂业务模拟和节点重构等大规模仿真，全景捕获复现和应激反制等对抗性仿真，多层次多维度攻击效能效果的仿真评估等重难点技术，构建大规模、高逼真、对抗性网络靶场。并在技术验证、战略预判、内容检验、情报分析，舆情预警等方面构建攻防演练模型，通过靶场的系统性、基础性、开拓性工作，使国家内容安全能力得到实质性提升。

2. 面向舆情攻防演练与研究的大规模社会系统模拟装置

模拟装置以虚实结合的方式进行建设，用真实数据驱动虚拟模型，对虚实数据进行一体化分析。基于最新人工智能技术，建立智能拟合模型，实现大规模舆情攻防模拟推演交互式可视化分析。支持多用户在模拟推演平台开展实验分析，支持对模拟系统运行进行实时干预，并提供可视化数据展示验证效果。支持政府对舆情信息了解得更加全面、对舆论动态认识得更加深刻、对敏感变化捕捉得更加及时，增加理政治国的主动性、施政为民的科学性。

参考文献

[1] Kietzmann J, Lee L W, McCarthy I P, et al. Deepfakes: Trick or treat? [J]. Business Horizons, 2020, 63(2): 135–146.

[2] Stamatis K. Artificial Intelligence in digital media: The era of deepfakes [J]. IEEE Transactions on Technology and Society, 2020, 1(3): 138–147.

[3] Tavakoli M, Baldi P. Continuous representation of molecules using graph variational autoencoder [C]. CA: 2020 AAAI Spring Symposium on Combining Artificial Intelligence and Machine Learning with Physical Sciences (AAAI-MLPS 2020), 2020.

[4] Liu H, Zheng X Y, Han J G, et al. Survey on GAN-based face hallucination with its model development [J]. IET Image Processing, 2019, 13(14): 2662–2672.

[5] Brashier N M, Schacter D L. Aging in an era of fake news [J]. Current Directions in Psychological Science, 2020, 29(3): 316–323.

[6] Jeffcao. 人工智能时代下的“烦恼”：美国国会听证会探讨“深度伪造（deepfake）”风险及对策[R/OL]. 北京：腾讯研究院，(2019-07-02) [2020-11-15]. https://www.tisi.org/10852.
Jeffcao. The “worries” in the era of Artificial Intelligence: US congressional hearings discuss the risks and countermeasures of “deepfake” [R/OL]. Beijing: Tencent Research Institute, (2019-07-02) [2020-11-15]. https://www.tisi.org/10852.

[7] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection [J]. Information Fusion, 2020 (64): 131–148.

[8] Zhou X Y, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities [J]. ACM Computing Surveys, 2020, 53(5): 1–40.

[9] Joseph A D, Nelson B. Adversarial machine learning [M]. Cambridge: Cambridge University Press, 2019.

[10] Mueller S T, Klein G. Explanation in Human-AI systems: A literature meta-review synopsis of key ideas and publications and bibliography for explainable AI [R/OL]. (2019-02-05)[2020-11-15]. Florida: Institute for Human and Machine Cognition Pensacola United States, https://deepai.org/publication/explanation-in-human-ai-systems-a-literature-meta-review-synopsis-of-key-ideas-and-publications-and-bibliography-for-explainable-ai.

[11] Karlo D F, Mario B, Nikica H. Explainable artificial intelligence: A survey [C]. Opatija: the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), 2018.

[12] 喻国明, 马思源. 人工智能提升网络舆情分析能力 [J]. 网络传播, 2017 (2): 85–87.

Yu G M, Ma S Y. Artificial intelligence improves online public opinion analysis capabilities [J]. Internet Broadcast, 2017 (2): 85–87.

[13] Raman R S, Ensafi R. Censored planet: An Internet-wide, longitudinal censorship observatory [R/OL]. Ann Arbor: University of Michigan, (2020-11-10)[2020-11-15]. https://censoredplanet.org/censoredplanet.

[14] 中国人工智能2.0发展战略研究项目组. 中国人工智能2.0发展战略研究[M]. 杭州：浙江大学出版社, 2018.
Strategic Research on Artificial Intelligence 2.0 in China Team. Strategic research on Artificial Intelligence 2.0 in China [M]. Hangzhou: Zhejiang University Press, 2018.