doi:10.11887/j.cn.202103010 http://journal.nudt.edu.cn

运用科学计量学的人工智能安全技术评估*

吴 集,梁江海,刘书雷
(国防科技大学 前沿交叉学科学院,湖南 长沙 410073)

摘 要:面对智能化时代的巨大进步,人工智能安全逐渐成为备受关注的议题。以Web of Science数据库收录的核心论文为研究对象,运用科学计量学方法和可视化技术,对包括国家、机构在内的研究力量分布以及热点、前沿和所涉学科在内的主题内容进行分析和总结。针对潜在的人工智能安全议题,构建定性分析框架,对人工智能安全进行定性定量结合的分析,对人工智能安全技术发展进行初步探索与评估。

关键词:人工智能;安全;科学计量;分析框架
中图分类号:TN95 **文献标志码**:A **文章编号**:1001-2486(2021)03-073-06

Artificial intelligence security technology evaluation applying scientometrics

WU Ji, *LIANG Jianghai*, *LIU Shulei*
(College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China)

Abstract: Artificial intelligence security is an important issue caused by rapid development of modern science and technology. Based on core papers obtained from the Web of Science database, a scientometrics method and the visualized tool were used to explore the research distribution of countries, institutions, and research hot spots. A qualitative analysis framework was constructed for analyzing artificial intelligence security in integrated qualitative and quantitative perspectives. A preliminary exploration and evaluation of the development of artificial intelligence security technology was presented.

Keywords: artificial intelligence; security; scientometrics; analysis framework

全球人工智能技术迅猛发展,正引领科技、产业新变革,推动智能经济、智能社会转型发展,塑造人类安全与发展新环境。人工智能具有先进科学技术"双刃"属性。面对智能化时代的巨大进步,必须坚持安全发展理念,审视潜在的机遇和风险,防范人工智能引发的系统性破坏和结构性风险,确保人工智能健康发展、造福于民。

本文运用科学计量学方法和可视化技术,基于对人工智能安全研究的文献情况进行计量分析,以Web of Science数据库收录的核心论文为研究对象,运用科学计量学方法和可视化技术,对包括国家、机构在内的研究力量分布以及热点、前沿和所涉学科在内的主题内容进行分析和总结,针对潜在的人工智能安全议题,构建定性分析框架,对人工智能安全进行定性定量结合的分析,在此基础上通过风险与技术映射,初步对若干人工智能安全前沿技术进行梳理和评估,希望能够为相关进一步研究提供参考。

1 人工智能安全研究的状况

过去20年间,全球众多国家与地区广泛参与到人工智能领域的研究中。在人工智能研究掀起新一轮浪潮的同时,波音自动飞行控制系统失效、优步无人驾驶车致命、人工合成表情产生等安全事件频发,无人驾驶、社会伦理、系统灾难、隐私侵犯等方面问题不断突显,国际人工智能安全研究不断受到关注。

针对"人工智能安全研究格局是什么,关注点在哪里,应对技术和措施有哪些研究?"等问题,采用可视化工具CiteSpace[1-2]对2007—2017年国际人工智能安全领域的期刊文献进行了国家分析、研究机构分析、来源出版物共引分析、领域研究热点分析与知识群聚类,从微观和宏观的层面分析人工智能安全研究热点,为后续开展人工

* 收稿日期:2019-11-04
基金项目:国家社会科学基金资助项目(2019-SKJJ-B-020)
作者简介:吴集(1977—),男,广西崇左人,副研究员,博士,E-mail:wwujji@163.com;
梁江海(通信作者),男,助理研究员,博士,E-mail:wxs@sina.com

智能安全技术评估提供依据。

1.1 人工智能安全研究的国家和机构分析

采用文献检索方法为主题：（artificial intelligence * secur * or robot * secur * or auto * secur *）时间跨度为 2007—2017 年，返回检索结果18 762篇。从发表人工智能安全文献的数量来看，地区呈现不均衡的分布。美国和中国是发表人工智能安全论文最多的国家，美国为 512 篇，中国为 310 篇，印度、德国为 150 篇左右。

在 CiteSpace 中“node type”选择参数“Country”，以 2 年为一个时间片，生成人工智能安全研究的国家分布网络图谱，如图 1 所示。图 1 中的每个节点代表了一个国家，用不同大小的年轮进行表示。年轮的大小与该国家的文献数量成正比。相邻节点的边的粗细与节点之间的联系程度成正比。图 1 中，每个年轮最中心的圆表示文献的中心性，中心性是在知识图谱网络中起连接作用大小的度量，年轮中心圆的直径越大，则中心性越大，说明该节点在网络中越重要，与其他节点的联系更紧密。从文献的中心性来看，美国排在首位，中国排在第二位。

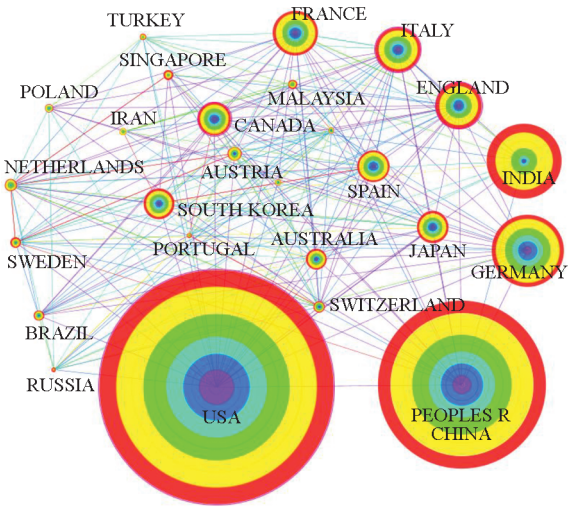

图 1 人工智能安全论文影响力国家网络

Fig. 1 National influence network of AI security papers

在 CiteSpace 中，“node type”选择参数“Institution”，以 2 年为一个时间片，得到了人工智能安全研究的学术机构分布图谱，如图 2 所示。图 2中，每个节点代表了一个发表论文的学术机构，用不同大小的年轮进行表示，年轮的大小与该机构的文献数量成正比。文献产出数量方面，发表人工智能安全文献最多的学术机构是 Chinese Acad Sci、Tsinghua Univ、Univ Illnois、Nanyang Technol Univ、Georgia Inst Technol 等，National Univ Def Technol 排在第 14 位；文献的中心性方面，在人工智能安全研究网络中起到关键连接的学术机构排序依次为 Chinese Acad Sci、Univ Illnois、Tsinghua Univ、Purdue Univ、Carnegie Mellon Univ 等，National Univ Def Technol 排在第 11 位。

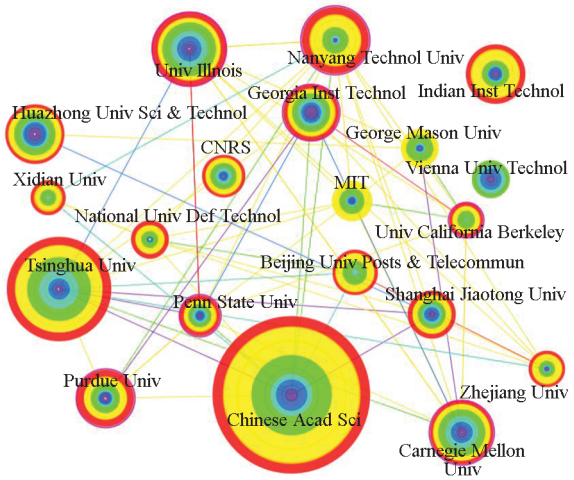

图 2 人工智能安全研究的机构分布网络

Fig. 2 Network of AI security research facilities

1.2 人工智能安全研究的热点分析

文献题录中的关键词是对主题的高度概括和集中描述，可以用于确定某一学科领域的研究热点。学科领域在每一时期都有研究热点，进而构成该学科的主要知识领域。基于 CiteSpace 对 Web of Science 核心合集文献进行分析，可用可视化的形式展现主题词或关键词的频次高低、聚类关系，得出研究热点。

将最终精炼整理得到的18 762 篇论文全记录信息导入 CiteSpace，设置参数为：时区分隔（time slicing）=（from 2007 to 2017）（2 years per slice）；主题词来源（term source）为标题（title）、摘要（abstract）、作者关键词（author keywords（DE））、扩展关键词（keywords plus（ID）），即全部勾选；节点类型（node types）选择关键词（keyword）；阈值（selection criteria）为每个时间片前 35 个高频或高被引节点；选择最小生成树（minimum spanning tree）算法进行剪枝（pruning）；视图方式（visualization）选择 timeline，进行图谱绘制。得到如图 3 所示的高频关键词随着时间演化的时间线演化图，以及如表 1 所示的高频关键词、高中心性关键词。

从表 1 中可以看出，国际人工智能安全研究领域十分广泛，不仅围绕身份认证、算法、隐私、网络安全、入侵检测等，且在无线传感网络、智能电网、自动化等领域也有数量众多的文章。从关键词的影响看，算法、自动化、可信性、验证、入侵检测、管理、协议、智能电网、机器学习是备受关注的议题。

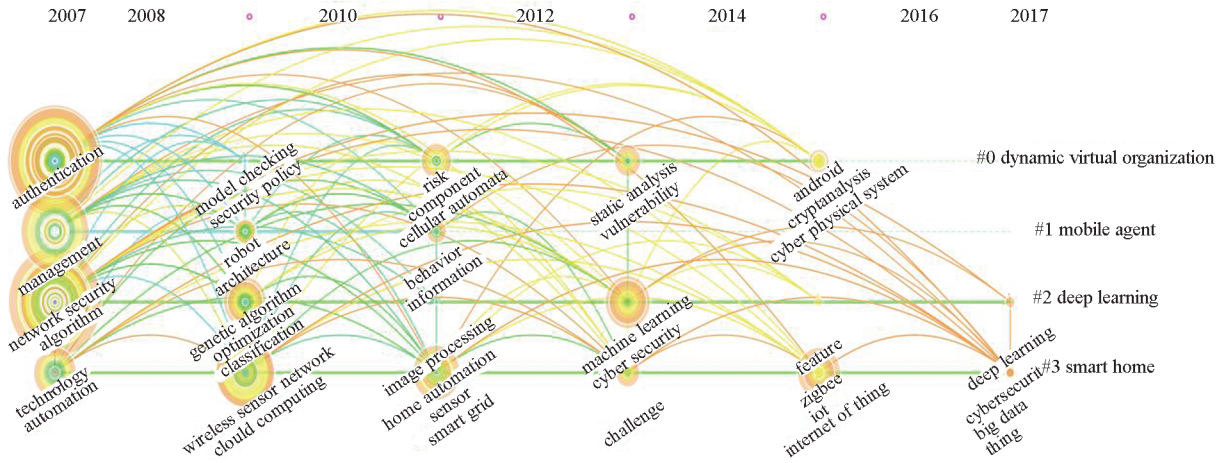

图 3　人工智能安全研究的高频关键词演化图

Fig. 3　Timeline diagram of AI security high frequency keywords

表 1　AI 安全高频关键词、高中心性关键词

Tab. 1　High frequency and centrality keywords of AI security

高频关键词	频次	高中心度关键词	中心度
authentication	361	algorithm	0.22
algorithm	344	automation	0.22
privacy	335	trust	0.21
design	321	verification	0.16
network security	275	intrusion detection	0.14
management	259	management	0.11
IDS	238	protocol	0.11
verification	231	smart grid	0.10
WSN	219	machine learning	0.10
cloud computing	213	authentication	0.08
recognition	183	attack	0.07
smart grid	183	sensor	0.07
automation	182	architecture	0.07
protocol	181	network security	0.06
cyber security	168	cryptography	0.06

2　人工智能安全的定性分析框架

依据文献分析的结果可以看出:以往关注的人工智能安全与信息、网络、自动化等紧密相关,如身份认证、算法、隐私、网络安全、入侵检测等,但对新一代人工智能技术如深度学习、植入芯片、深度伪造等安全议题的研究体现得并不明显,对新一代人工智能安全技术的研究仍处于起步阶段。为此,从定性分析角度,建立人工智能安全分析框架,结合各国宏观政策和学术研究、社会关注情况,梳理人工智能安全应关注的议题。

2.1　政策与研究综述

2017 年,中国发布《新一代人工智能发展规划》,提出“建设安全便捷的智能社会,提高社会治理智能化水平”“构建泛在安全高效的智能化基础设施体系”,从法律法规、伦理规范、安全监管与评估等方面推动人工智能安全建设[3]。2017 年初,在美国阿西洛马召开的 Beneficial AI 会议提出“阿西洛马人工智能原则”[4];2018 年,中国《人工智能安全发展上海倡议》强调,人工智能安全发展要遵循“面向未来、以人为本、责任明晰、隐私保护、算法公正、透明监管、和平利用、开放合作”八大理念[5]。

2019 年 4 月,欧盟发布人工智能的七个道德准则:受人类监管,技术的稳健性和安全性,隐私和数据管理,透明度,多样性、非歧视性和公平性,社会和环境福祉,问责制[6]。2019 年 6 月,中国国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》,提出和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理八项原则[7]。

根据国内社会调查和国际学者研究,人工智能安全关注的第一类议题是未来人工智能对人类生存的挑战。依据奇点预测,2045 年人工智能将超越人类智能,将首次对全体人类在生存、生产上带来终极挑战。第二类议题是人工智能可能引发的安全问题。其一是操作权失控或恶意使用导致的安全威胁。黑客可能通过智能方法发起网络攻击。其二是技术失控或管理不当所致的安全问题。第三类议题是对已有社会体系破坏的问题。人工智能技术越发达,信息鸿沟就越深,进而演变成为服务鸿沟、福利鸿沟。第四类议题是人工智能对伦理规范的挑战。从数据隐私、安全威胁、算

法歧视和人工主体权利伦理方面，人工智能技术的应用正在将一些生活中的伦理性问题在系统中规则化。第五类议题是系统的研发设计必须要与社会伦理匹配对接，机器规范和人类规范必须兼容。

2.2 人工智能安全分析框架

为此，鉴于网络空间安全的复杂性，人工智能安全涉及因素更为广泛深入。参考 OSI、CC 等网络安全评估标准模型[8]，从技术、应用、社会三个层次，构建人工智能安全分析框架，如图 4 所示。依据人工智能安全分析框架，结合科学计量的高频关键词等启发式信息，新一代人工智能安全应关注的维度、领域和潜在风险包括以下方面。

图 4 人工智能安全定性分析框架

Fig. 4 AI security qualitative analysis framework

在技术因素维度，支持人工智能的“计算、算法、数据、网络”四大核心，本身仍存在无法验证的内在安全。在传统的计算、算法、网络中，尽管在安全供应链、可信计算、算法验证、网络安全方面开展了大量工作，但仍缺乏全面的突破，对用于检验和维护深度学习算法的数据集合安全仍缺乏研究。

在系统应用维度，从已经开始部署的无人驾驶汽车、无人机、先进制造、网络服务，以及未来潜在的金融、运输、媒体、教育、医疗等，系统能力不足、设计存在漏洞、缺乏安全机制、产业监管滞后等引发了隐患甚至事故，技术和系统设计不确定性风险广泛存在。

在安全治理维度，恶意代码、野生智能、机器失控、操纵智能手段犯罪等将对治安、司法、反恐等带来新的问题，人与机器共存、异构智能共存、社会角色转换等将对社会伦理带来需要应对的不确定性。

3 人工智能安全技术评估初探

3.1 应关注的人工智能安全技术

基于文献计量和政策分析获得的两类安全关注点，依据人工智能安全分析框架，通过建立“风险－场景－技术”关联，以图 4 辨析的风险和场景为需求牵引，通过网络信息安全、系统与工业安全、社会安全领域的国内外技术扫描和技术预测，针对新一代人工智能安全潜在的分析，梳理出 15 项应关注的下一代人工智能安全技术，如图 5 所示。

在使能技术维度，针对数据、算法、计算、网络领域存在的潜在风险，研究提出隐私数据完美加密、机器学习训练数据集防攻击等 6 项应关注的技术。在系统与应用维度，针对 AI + 产业、AI 系统、AI 应用领域潜在的风险，提出人机混合智能系统安全设计、安全防伪的生物特征识别与验证（应对深度伪造）等 5 项应关注的技术。在安全与治理维度，针对社会与伦理、公共安全领域潜在的风险，提出应对机器智能的人类效能增强、以人为中心的人机共生社会治理等 4 项应关注的技术。

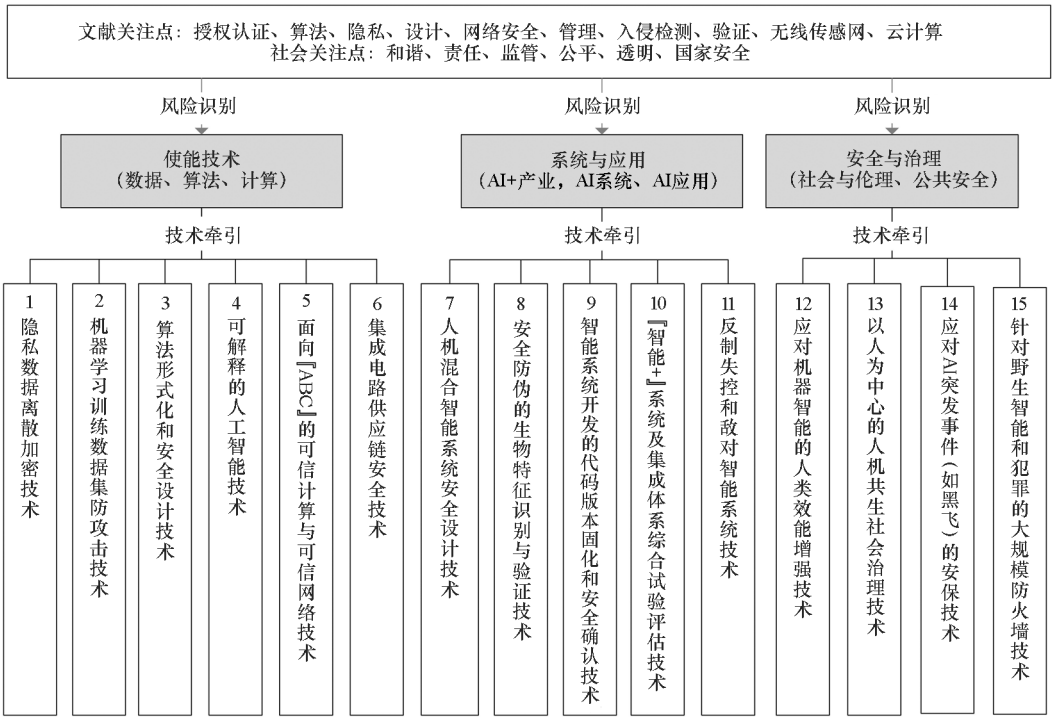

图 5　应关注的 15 项人工智能安全前沿技术

Fig. 5　Selected 15 AI security advanced technologies

3.2　人工智能安全技术发展评估

由人工智能潜在安全风险牵引和梳理出的 15 项人工智能安全技术发展程度各不相同。算法形式化和安全设计、可信计算与可信网络、反制失控和敌对智能系统等技术国际均有一定程度的发展,隐私数据完美加密、集成电路供应链安全、应对机器智能的人类效能增强等国外机构正在开展研究。其中,隐私数据完美加密指通过数据脱敏、匿名化、差分隐私和同态加密技术,防止智能数据挖掘、网络搜索对隐私数据的侵害;集成电路供应链安全,如 DARPA“电子供应链硬件完整性”等项目对集成电路芯片进行防伪监测;应对机器智能的人类效能增强,指应用生物植入式芯片、混合显示、脑机接口等增强人类反应、认知、行动效能的技术。依据国内外技术发展动向,对 15 项人工智能安全技术进行初步的“技术成熟度”和“技术重要度”两项评估,结果如图 6 所示。

技术成熟度评估主要依据 NASA 技术成熟度从 1 到 9 级的等级划分[9]。技术重要度主要考虑每项人工智能安全技术针对的潜在风险频度、化解风险的支撑程度,即技术针对的风险频发、影响越大,则技术越重要。技术重要度具体的量化公式为:

$$TA = C \cdot \boldsymbol{F}_{\text{safety}} \cdot \boldsymbol{F}_{\text{critical}}$$

其中:C 为调节常数,$\boldsymbol{F}_{\text{safety}}$ 为技术应对的潜在风险频度向量因子,$\boldsymbol{F}_{\text{critical}}$ 为技术应对的潜在风险致命度向量因子。实际计算中,$\boldsymbol{F}_{\text{safety}} = [fs_1, \cdots, fs_8]_{1\times 8}$,$fs_i$ 为技术对应某类风险发生概率,取值为[0, 1]间的实数。$\boldsymbol{F}_{\text{critical}} = Trans([fc_1, \cdots, fc_8]_{1\times 8})$,为技术对应某类风险的致命程度,取值为[0,1]间的实数。调节常数取 $C = \lceil 9 \times TA / \max(TA) \rceil$,$\lceil\ \rceil$ 为向上取整运算。

以图 5 中第 12 项技术“应对机器智能的人类效能增强”为例,TA 计算过程为:

$$TA = C \cdot [0.6,0.6,0.3,0.2,0.4,0.2,0.3,0.1] \times [0.2,0.2,0.3,0.3,0.4,0.5,0.5,0.6]^{\mathrm{T}} = 4$$

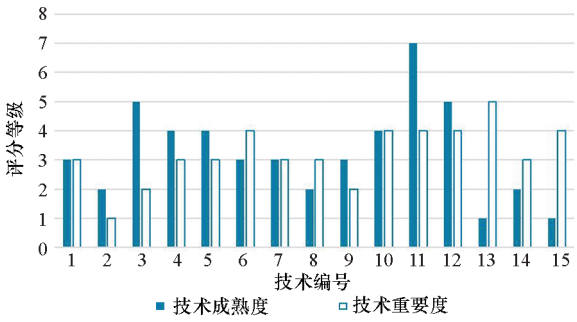

图 6　人工智能安全前沿技术评估

Fig. 6　Evaluation on AI Security advanced technologies

4　结论

人工智能安全研究是坚持安全与发展协调统

一,应对智能“终极挑战”的新兴重大问题。目前,针对新一代人工智能安全的研究仍处于起步阶段。借鉴以往信息化发展经验,面对新一代人工智能技术快速发展、广泛渗透,从安全技术、安全政策、治理体系等方面有待于创新,为应对各领域包括军事上的工智能安全问题提供支撑。

本文运用可视化科学计量工具以及 Web of Science 核心论文数据集,对人工智能安全研究的文献情况进行计量分析,针对潜在的人工智能安全议题,构建定性分析框架,对人工智能安全进行定性定量结合的分析,在此基础上通过风险与技术映射,初步对若干人工智能安全前沿技术进行了梳理和评估,为相关研究尤其是构建未来智能化发展的安全技术体系提供思路和参考。

参考文献(References)

[1] CHEN C. CiteSpace Ⅱ: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359 - 377.

[2] 任利强, 郭强, 王海鹏, 等. 基于 CiteSpace 的人工智能文献大数据可视化分析[J]. 计算机系统应用, 2018, 27(6): 18 - 26.
REN Liqiang, GUO Qiang, WANG Haipeng, et al. CiteSpace-based visualization analysis of literature big data on artificial intelligence[J]. Computer Systems and Applications, 2018, 27(6): 18 - 26. (in Chinese)

[3] 中华人民共和国国务院. 新一代人工智能发展规划[R]. 北京: 中华人民共和国国务院, 2017.
State Council of the People's Republic of China. New generation artificial intelligence development plan [R]. Beijing: State Council of the People's Republic of China, 2017. (in Chinese)

[4] The Future of Life Institute. Asilomar AI principles[EB/OL]. (2018 - 03 - 20)[2019 - 10 - 22]. https://futureoflife.org/ai-principles/.

[5] 黄辛. 上海发布《人工智能安全发展上海倡议》[N]. 中国科学报, 2019 - 07 - 11(7).
HUANG Xin. Shanghai propose “Shanghai initiative for artificial intelligence security development”[N]. China Science Daily, 2019 - 07 - 11(7). (in Chinese)

[6] 方莹馨. 欧盟发布人工智能伦理准则[N]. 人民日报, 2019 - 04 - 11(17).
FANG Yingxin. EU issue artificial intelligence ethics guidelines[N]. People's Daily, 2019 - 04 - 11(17). (in Chinese)

[7] 科技部. 发展负责任的人工智能: 新一代人工智能治理原则发布[EB/OL]. (2019 - 06 - 17)[2019 - 10 - 03]. http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm.
MOST. Developing responsible artificial intelligence: new generation of artificial intelligence governance principles released[EB/OL]. (2019 - 06 - 17)[2019 - 10 - 03]. http://www.most.gov.cn/kjbgz/201906/t20190617_147107.htm. (in Chinese)

[8] 谭良, 佘堃, 周明天. 信息安全评估标准研究[J]. 2006, 27(04): 634 - 637.
TAN Liang, SHE Kun, ZHOU Mingtian. Search of security evaluation criteria [J]. Journal of Chinese Mini-Micro Computer Systems, 2006, 27(04): 634 - 637. (in Chinese)

[9] MANKINS J C. Technology readiness levels: a white paper[R]. Washington DC: NASA, 1995.