软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(9):3941−3965 [doi: 10.13328/j.cnki.jos.006875]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

可信人工智能系统的质量属性与实现: 三级研究*

李功源[1,2], 刘博涵[1,2], 杨雨豪[1,2], 邵 栋[1,2]

[1](南京大学 软件学院, 江苏 南京 210023)

[2](计算机软件新技术国家重点实验室 (南京大学), 江苏 南京 210023)

通信作者: 刘博涵, E-mail: bohanliu@nju.edu.cn

摘 要: 人工智能系统以一种前所未有的方式, 被广泛地用于解决现实世界的各种挑战, 其已然成为推动人类社会发展的核心驱动力. 随着人工智能系统在各行各业的迅速普及, 人们对人工智能系统的可信性愈发感到担忧, 其主要原因在于, 传统软件系统的可信性已不足以完全描述人工智能系统的可信性. 对于人工智能系统的可信性的研究, 具有迫切需要. 目前已有大量相关研究, 且各有侧重, 但缺乏一个整体性、系统性的认识. 研究是一项以现有二级研究为研究对象的三级研究, 旨在揭示人工智能系统的可信性相关的质量属性和实践的研究现状, 建立一个更加全面的可信人工智能系统质量属性框架. 收集、整理和分析 2022 年 3 月前发表的 34 项二级研究, 识别 21 种与可信性相关的质量属性及可信性的度量方法和保障实践. 研究发现, 现有研究主要关注在安全性和隐私性上, 对于其他质量属性缺乏广泛且深入的研究. 对于需要跨学科协作的两个研究方向, 需要在未来的研究中引起重视, 一方面是人工智能系统本质上还是一个软件系统, 其作为一个软件系统的可信值得人工智能和软件工程专家合作研究; 另一方面, 人工智能是人类对于机器拟人化的探索, 如何从系统层面保障机器在社会环境下的可信, 如怎样满足人本主义, 值得人工智能和社会科学专家合作研究.

关键词: 人工智能系统; 可信; 质量属性; 实践

中图法分类号: TP18

中文引用格式: 李功源, 刘博涵, 杨雨豪, 邵栋. 可信人工智能系统的质量属性与实现: 三级研究. 软件学报, 2023, 34(9): 3941–3965. http://www.jos.org.cn/1000-9825/6875.htm

英文引用格式: Li GY, Liu BH, Yang YH, Shao D. Quality Attributes and Practices of Trustworthy Artificial Intelligence Systems: A Tertiary Study. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 3941–3965 (in Chinese). http://www.jos.org.cn/1000-9825/6875.htm

Quality Attributes and Practices of Trustworthy Artificial Intelligence Systems: A Tertiary Study

LI Gong-Yuan[1,2], LIU Bo-Han[1,2], YANG Yu-Hao[1,2], SHAO Dong[1,2]

[1](Software Institute, Nanjing University, Nanjing 210023, China)

[2](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: Artificial intelligence systems are widely used to solve various challenges in the real world in an unprecedented way, and they have become the core driving force for the development of human society. With the rapid popularization of artificial intelligence systems in all walks of life, the trustworthiness of artificial intelligence systems is becoming more and more worrying. The main reason is that the trustworthiness of traditional software systems is not enough to fully describe that of artificial intelligence systems. Therefore, research on the trustworthiness of artificial intelligence systems is urgently needed. At present, there have been a large number of relevant studies, which focus on different aspects. However, these studies lack a holistic and systematic understanding. This study is a tertiary study with

* 基金项目: 国家自然科学基金 (62072227, 62202219); 国家重点研发计划 (2019YFE0105500); 江苏省重点研发计划 (BE2021002-2); 南京大学计算机软件新技术国家重点实验室创新项目 (ZZKT2022A25); 海外开放课题 (KFKT2022A09)
本文由“AI 软件系统工程化技术与规范”专题特约编辑张贺教授、夏鑫博士、蒋振鸣副教授、祝立明教授和李宣东教授推荐.
收稿时间: 2022-09-04; 修改时间: 2022-10-13; 采用时间: 2022-12-14; jos 在线出版时间: 2023-01-13
CNKI 网络首发时间: 2023-07-05

the existing secondary study as the research object. It aims to reveal the research status of quality attributes and practices related to the trustworthiness of artificial intelligence systems and establish a more comprehensive quality attribute framework for trustworthy artificial intelligence systems. This study collects, sorts out, and analyzes 34 secondary studies published until March 2022. In addition, it identifies 21 quality attributes related to trustworthiness, as well as measurement methods and assurance practices of trustworthiness. The study finds that existing research mainly focuses on security and privacy, and extensive and in-depth research on other quality attributes is fewer. Furthermore, two research directions requiring interdisciplinary collaboration need more attention in future research. On the one hand, the artificial intelligence system is essentially a software system, and its trustworthiness as a software system is worthy of collaborative research by artificial intelligence and software engineering experts. On the other hand, artificial intelligence belongs to human's exploration of machine anthropomorphism, and research on how to ensure the trustworthiness of machines in the social environment from the system level, such as how to satisfy humanism, is worthy of collaborative research by artificial intelligence and social science experts.

Key words: artificial intelligence system; trustworthy; quality attribute; practice

在过去的 10 年间, 计算机处理能力的提高、数据集的扩大和算法准确性的提升推动了人工智能 (artificial intelligence, AI) 技术的进步[1]. 随着越来越多的 AI 技术从实验室走向产业界, 基于 AI 技术构建的软件系统也被广泛应用于医疗、工业、教育等各领域[2]. AI 系统在这些应用领域中展现出新颖且出色的性能, 其适用的范围和场景都在以革命性的方式迅速扩大. AI 系统提供的许多创新功能, 如自动驾驶, 是前所未见的; AI 系统利用的大量数据是过去不曾被使用的. 然而, 新兴的事物总易遭受质疑, 强大的能力需要受到更严格的约束. AI 所体现出的优秀特质也导致了社会对于 AI 系统可信性的广泛讨论, 这些讨论涉及人工智能系统的隐私性、公平性、安全性等多个方面. AI 系统基于大量数据建立影响社会的规则和行为, 如果在数据中存在偏见, 则 AI 系统会存在公平性的问题. 例如, 美国法院基于 AI 预测再犯概率的系统被证实对黑人存在偏见[3]; 极端光照条件会影响视觉识别系统, 进而影响自动驾驶系统的安全[4]; 对人脸识别系统的恶意攻击可能会导致用于训练的个人图像的泄漏[5]. 类似问题正在"人本主义的人工智能 (human-centered AI, HCAI)""人工智能伦理 (AI ethics)""可信人工智能 (trustworthy artificial intelligence)"等主题下被广泛热议[6−8].

为了有效解决 AI 系统在实际应用中存在的问题, 学术界、产业界以及政府组织都在密切关注 AI 系统的可信性, 构造可信的 AI 系统应作为其取得更加广泛应用的前提. 近年来, 可信 AI 领域涌现了大量标准、指南等规范性文件. 国际标准化组织 (International Organization for Standardization, ISO) 的技术报告 ISO TR 24028[9]分析了影响 AI 系统可信性的因素并从公平性、透明性、问责性和可控性等方面讨论了提高 AI 系统可信性的方法. 欧盟提出了《可信 AI 伦理指南 (ethics guidelines for trustworthy AI)》[10], 建议通过尊重人类自治, 伤害预防, 公平性, 可解释性等 4 项原则监管 AI 系统, 并提出了《可信 AI 评估指南 (assessment list for trustworthy artificial intelligence, ALTAI) 》[11]用于帮助企业或其他组织评估 AI 系统的开发、部署、采购和使用是否符合《可信 AI 伦理指南》的相关原则. 中国信息通信研究院和京东探索研究院于 2021 年发布的《可信人工智能白皮书》[12]从落实 AI 治理共识的角度出发, 围绕可靠可控、透明可释、隐私保护以及明确责任等方面, 将各项要求引入 AI 系统研发的全流程.

可信 AI 同样引起了学术界的广泛关注. 可信是一个抽象的概念, 要理解并进一步实现可信, 需要从具体的与可信相关的质量属性入手. 自 2017 年以来, 涌现了大量关于可信 AI 系统质量属性及实践的二级研究 (secondary study). 二级研究是指评价与特定研究问题相关的所有一级研究 (primary study) 的研究, 目的是综合与特定研究问题相关的证据; 其中一级研究是指调查特定研究问题的经验研究 (empirical study)[13]. 但是这些二级研究都有着一定的局限性, 它们讨论了可信 AI 系统的一个或多个质量属性, 但覆盖范围不够全面. 例如, 文献 [14] 仅在欧盟的 4 项原则下[10]讨论了可信 AI 系统的质量属性, 缺少对质量属性的辨析以及对相关实践的讨论. 文献 [15] 仅关注安全性和鲁棒性两个可信 AI 系统的质量属性及相关实践, 缺乏对可信性全面的讨论. 文献 [16] 仅讨论了公平性、隐私性、可解释性、可问责性和可接受性 5 项要求及其相关的保障实践, 缺乏对更多质量属性的研究, 同时, 也未讨论这些质量属性的度量评估方法. 更多的二级研究, 如文献 [17−19] 仅聚焦于单个质量属性.

为了对 AI 的可信建立一个更加系统且全面的认识, 本文开展了一项三级研究 (tertiary study) 来收集并分析现有与可信 AI 相关的二级研究, 具体从可信 AI 系统的二级研究现状、质量属性、评估度量方法、保障改进实践等 4 个方面开展了研究. 三级研究是将二级研究的输出作为输入的综述性研究[20,21]. 当一个研究主题上的二级研

究足够多时, 三级研究可以覆盖更广泛的证据并从更高的抽象层次开展研究, 为研究者建立更系统性地理解和定位相关的二级研究提供帮助, 弥补单项二级研究对问题认识不全面及弱化的单项研究中可能引入的偏见[22,23].

本文检索、挑选和分析了 34 项相关的二级研究[14−19,24−51], 具体而言, 主要贡献包括: (1) 本文揭示了可信 AI 系统的质量属性及相关实践的二级研究的研究现状. (2) 总结了现有二级研究中讨论的质量属性, 按照可信 AI 的基本原则、质量属性及子属性的层次结构建立了可信人工智能系统的质量属性框架. (3) 梳理了现有的可信性评估度量方法. (4) 梳理了现有的可信性保障改进实践.

本文第 1 节阐明本三级研究的研究问题及所采用的研究方法. 第 2 节至第 5 节对本研究的 4 个研究问题分别进行回答. 第 6 节基于研究结果进行讨论. 第 7 节讨论本文的效度威胁. 第 8 节对本文进行总结.

1 研究方法

本研究开展的是以二级研究为研究对象的三级研究. 二级研究通常指以一级研究为对象开展的研究, 包括采用系统性文献综述 (systematic literature review, SLR)、映射研究 (mapping study, MS)、一般综述 (review) 等方法的综述性研究[13]. Kitchenham 等人[20,21]将三级研究定义为“使用一个学科内的二级研究的输出作为输入的研究”. 本研究的开展遵循 Kitchenham 等人关于软件工程领域系统性文献综述的指南[13]. 目前, 该指南[13]描述的系统性文献综述方法被广泛应用于软件工程领域的二级研究和三级研究中, 是一种较映射研究和一般综述更加系统且严谨的研究方法. 两名硕士研究生和他们的两名导师作为研究人员参与了这项研究.

本节主要阐述研究过程中所采用的研究方法. 首先提出本三级研究的研究问题 (第 1.1 节), 其次描述文献的检索和筛选过程 (第 1.2 节), 最后描述本研究的数据抽取与分析过程 (第 1.3 节).

1.1 研究问题

为了分析可信 AI 领域内的相关二级研究的研究现状, 对可信 AI 的质量属性与相关实践建立系统性的理解, 本文提出了以下 4 个研究问题 (research questions, RQs).

RQ1: 现有可信人工智能系统相关二级研究的研究现状如何?

RQ2: 现有二级研究中讨论了哪些与人工智能系统的可信性相关的质量属性?

RQ3: 现有二级研究中讨论了哪些人工智能系统的可信性 (即各质量属性) 的评估度量方法?

RQ4: 现有二级研究中讨论了哪些人工智能系统的可信性 (即各质量属性) 的保障改进实践?

其中, RQ1 旨在对可信 AI 相关的二级研究的发表趋势, 发表渠道以及研究范畴等建立系统的理解. RQ2 旨在调研当前可信 AI 相关的二级研究中讨论的质量属性间的差异、联系及层次关系, 从而建立一个全面的 AI 系统可信性的质量属性框架. RQ3 和 RQ4 分别旨在对 AI 系统的可信性 (即各质量属性) 的评估度量方法和保障改进实践建立较为全面的认识, 识别出研究的热点和空白, 进而挖掘出未来可能的研究方向.

1.2 相关文献收集

本研究对现有二级研究文献的收集过程如后文图 1 所示, 包含手动检索、自动检索、文献筛选和滚雪球检索等 4 个步骤. 为了尽可能降低文献检索过程中遗漏相关文献的风险, 本文采用了基于准黄金标准 (quasi golden standard, QGS) 的检索策略[52], 具体的文献检索过程包括手动检索、自动检索以及滚雪球检索 3 个步骤, 其中滚雪球检索过程中包含了对文献的筛选. 文献的筛选由两名硕士研究生分别独立完成, 他们之间的任何差异都会和导师进行讨论, 所有被纳入的文献都由导师进行最终确认. 本研究最终纳入了 34 篇相关文献. 需要说明的是, 文献收集过程中没有对文献属于二级研究或三级研究加以限制, 最终所有识别到的文献均为二级研究, 即目前还没有一篇已发表的关于可信 AI 的三级研究. 以下将展开介绍文献收集过程的各具体步骤.

1.2.1 手动检索

手动检索基于领域内综述类文献的顶级期刊 ACM Computing Surveys 进行. 手动检索的主要目的是为后续的自动检索设计更加全面且准确的检索字符串提供基础. 同时, 使本研究的所有参与者对研究问题和研究范畴建立清晰且一致的理解, 并进一步确立文献选择标准. 我们逐一浏览了出版时间在 2018 年至 2022 年 3 月之间的文献并进行筛选, 共收集到了 6 篇关于可信 AI 质量属性的二级研究.

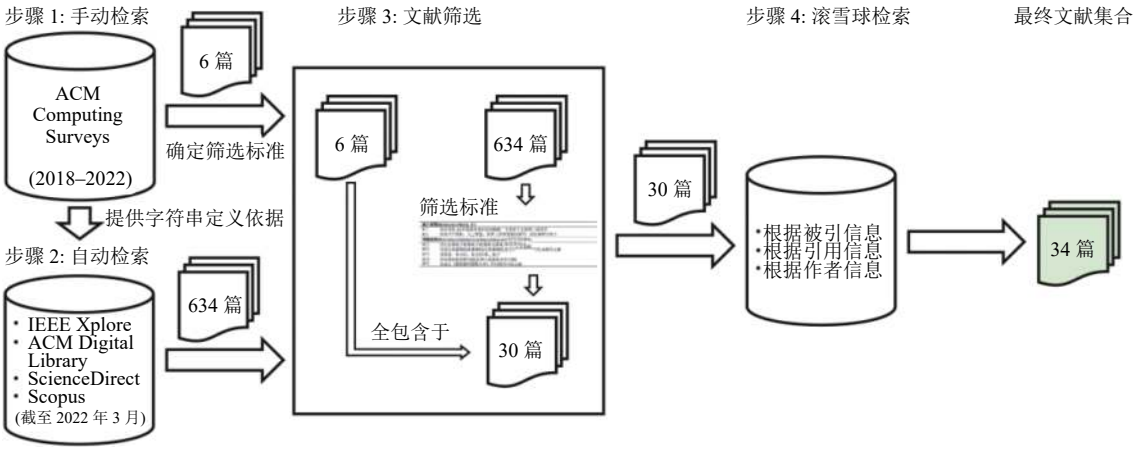

图 1　相关文献收集过程

1.2.2　自动检索

在确定自动检索的字符串时, 主要参考了手动检索得到的 6 篇文献中的关键词. 首先确定的是关键词“可信”和“人工智能”, 即“Trustworthy”和“AI”及它们的近义词. 进一步地, 收集了这 6 篇文献以及现有的与可信 AI 相关的 7 项规范性文件中涉及的质量属性作为关键词, 这些规范性文件包括 ISO TR 24027[53], ISO TR 24028[9], ISO TR 24029[54], IEEE 3652.1[55], IEEE P7001[56], ALTAL[11]. 由于本文的研究对象是二级研究, 所以增加了“Review”“SLR”等关键词. 最终确定的检索字符串如下所示.

(“Trustworthy” OR “Trustworthiness” OR “Quality Attribute” OR “QA” OR “Reliabl*” OR “Transparency” OR “Privacy” OR “Secur*” OR “Safe*” OR “Fair*” OR “Robust*” OR “Account*”) AND

(“AI” OR “Artificial Intelligence” OR “Machine Learning” OR “Deep Learning”) AND

(“Review” OR “Mapping” OR “SLR” OR “SMS”)

本文检索了 4 个文献数据库, 包括 IEEE Xplore, ACM Digital Library, Science Direct, Scopus. 其中 Scopus 是最大的同行评议文献数据库, 且收录中文文献, 能够对前 3 个数据库进行有效的补充, 该阶段在每个检索源中获得的检索结果如表 1 所示, 去重后共检索到 634 篇文献. 在手动检索阶段得到的 6 篇文献均包含在这 634 篇文献中, 说明检索字符串具有较高的有效性.

表 1　自动检索的中间结果

在线图书馆	文献数量
IEEE Xplore	242
ACM Digital Library	29
ScienceDirect	110
Scopus	253
总计 (未去重)	634

1.2.3　文献筛选

为了排除已获得文献集合中与研究问题不相关和关联度较小的文献, 本文定义了如表 2 所示的纳入和排除标准. 只有满足所有纳入标准且没有符合任意一条排除标准的文献才会被纳入最终的文献集合.

根据上述纳入/排除标准, 本文进行了文献的筛选, 具体筛选过程包括概要浏览和全文通读两个阶段. 概要浏览阶段, 通过浏览文献的标题、关键词和摘要以判断它们是否符合纳入的标准. 在这之后, 我们通读了满足纳入标准的文献, 并排除了那些符合任意排除标准的文献. 在进行文献筛选时, 每篇文献的筛选都由两位研究人员独立完成, 每一篇存在争议的文献都由他们的指导者评审并进行讨论以达成共识. 需要特别说明的是, 根

据中国科学院发布的《国际期刊预警名单》, 排除了 2 篇来自 IEEE Access 的二级研究. 通过文献筛选过程, 共保留了 30 篇二级研究.

表 2 纳入/排除标准

分类	编号	内容
纳入标准 (inclusion criteria, IC)	IC1	关注可信AI质量属性或相关实践的一个或多个方面的二级研究
	IC2	发表于计算机、人工智能、软件工程等领域的期刊、会议和研讨会上
排除标准 (exclusion criteria, EC)	EC1	无法获得电子版全文的文献
	EC2	仅包含质量属性面临的挑战、问题而不包括质量属性描述或实践的文献
	EC3	用英语、中文以外语言撰写的文献
	EC4	具有更新版本的文献 (仅纳入最新版本的文献)
	EC5	发表在《国际期刊预警名单》所列期刊中的文献

1.2.4 滚雪球检索.

在识别到的 30 篇文献的基础上, 本研究通过谷歌学术进行了滚雪球 (snowballing) 来进一步检索相关文献, 包括前向滚雪球, 后向滚雪球和根据作者信息滚雪球 3 个阶段. 其中, 前向滚雪球是检索引用了已识别的 30 篇文献的文献, 后向滚雪球是检索已识别文献的参考文献, 根据作者信息滚雪球是检索文献第一作者的其他相关文献. 在该阶段, 本研究遵循 Wohlin[57]的建议, 反复迭代滚雪球的过程, 直到不再发现新的文献为止. 此外, 除了检索以外, 本阶段还包含了前述的文献筛选. 滚雪球阶段新发现了 4 篇文献. 最终, 共纳入了 34 篇二级研究[14−19,24−51]作为本研究的对象.

1.3 数据抽取与分析

根据研究问题, 本文定义了数据抽取表格, 如表 3 所示, 从而标准化地获取每篇文献中所报告的本研究所需的信息. 为了减少数据抽取过程中个人偏见对研究结果的影响, 在数据抽取过程中, 一位研究人员抽取每一项研究的数据项, 另一位研究人员对抽取结果进行验证, 同时定期与他们的导师进行讨论.

表 3 数据抽取项

抽取项ID	抽取项	描述	对应研究问题
D1	标题	文献的标题	RQ1
D2	作者	文献的作者	
D3	年份	文献发表的年份	
D4	发表源	文献发表的渠道, 期刊、会议或研讨会	
D5	研究问题/目标	文献的研究问题或能体现具体研究目标的相关描述	
D6	质量属性	可信相关质量属性及其具体定义和相关描述	RQ2
D7	评估度量方法	可信相关质量属性的评估度量方法的描述和示例	RQ3
D8	保障改进实践	可信相关质量属性的保障改进实践的描述和示例	RQ4

在对抽取出的数据项进行分析时, 采用了定量分析和定性分析的方法. 定量分析方法用于获取文献集合中二级研究的趋势及分布等信息. 对于抽取项 D1–D4, 采用统计方法来获得文献分布和发表趋势的定量结果. 对于数据抽取项 D5–D8, 采用了定性的主题分析方法[58], 以获得二级研究中涉及的质量属性以及相关的实践. 为了支持主题分析, 本文采用了数据编码方法, 即一种通过标记和组织定性数据以识别主题以及主题之间的关系的方法[59], 使用开放编码和轴向编码从提取出的数据项中聚类得到高抽象层次的概念. 在进行数据合成时, 由两位研究人员分别对提取出的原始文本进行开放编码. 在所有数据完成初始编码后, 本文对获得的编码进行分析和比较, 进而将它们组合成主题. 最终获得的主题都经过了两位研究人员的交叉评审, 任何分歧都与他们的导师进行讨论, 直到达成共识.

2 研究现状 (RQ1)

本节中简要概述了本三级研究所纳入的二级研究的总体情况和研究范畴, 以给出 RQ1 的回答.

2.1 总体情况

本文纳入的 34 篇二级研究中, 包含 27 篇英文文献和 7 篇中文文献. 以下将从年份分布、地理分布、渠道分布 3 个方面进行统计分析.

● 年份分布. 图 2 展示了纳入文献的发表年份分布情况. 虽然自动检索时的检索范围截至 2022 年 3 月, 但从图 2 中可以看出, 在 2016 年之前还没有相关二级研究被发表. 可信 AI 相关的二级研究始于 2017 年, 且发表数量在 2017 至 2021 年呈逐年增长的趋势, 且 2022 年截止到 3 月发表的文献数量已接近 2021 年全年一半的数量. 这一方面表明 AI 的可信的问题已愈发受到研究人员的关注; 另一方面也说明可信 AI 是一个问题繁多的研究方向. 纳入的文献中, 发表最早的中文二级研究始于 2019 年, 且在此后同样保持稳定的发表趋势.

● 地理分布. 图 3 展示了纳入文献的第一作者所属研究机构的地理分布情况. 全球除了南美洲和南极洲 (无国家) 均有研究者从事可信 AI 相关的二级研究. 从占比来看, 来自亚洲的研究者占比最高 (50%), 这得益于中国研究者发表的二级研究占总体的 32%. 这应该与中国的学术界和工业界近年来在 AI 领域两面开花、蓬勃发展的现状密切相关.

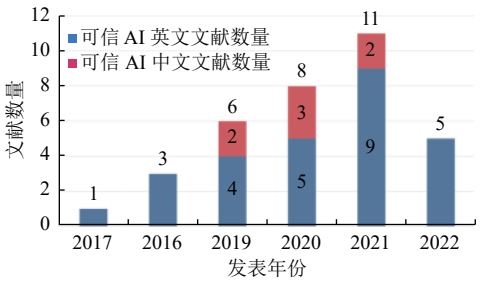

图 2 发表年份分布情况

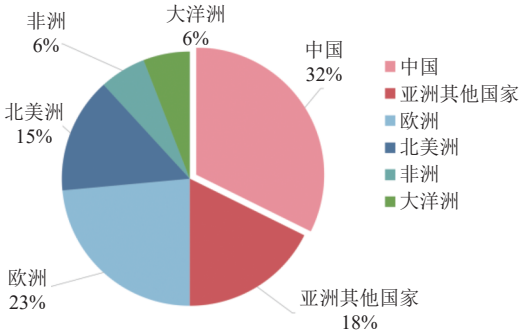

图 3 第一作者所属研究机构的地理分布情况

● 渠道分布. 图 4 展示了纳入文献在期刊 (Journal), 会议 (Conference) 和研讨会 (Workshop) 这 3 种发表渠道的分布情况. 其中半数以上 (62%) 为期刊论文, 29% 为会议论文, 只有 3 篇为研讨会论文. 对于具体的期刊或会议, 总体发表渠道较为分散. 但统计发现, 发表在国内外高水平刊物上的二级研究占比更高. 其中 6 篇英文论文发表在了综述类论文的国际顶级期刊《ACM Computing Surveys》上. 所有纳入的中文文献均发表在国内顶级期刊上, 其中 4 篇发表在《软件学报》, 3 篇发表在《计算机研究与发展》.

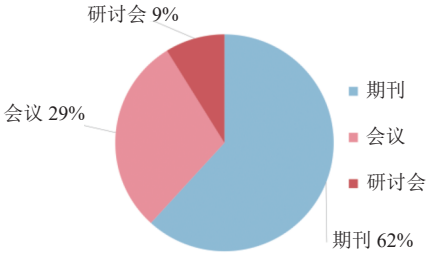

图 4 发表渠道分布

2.2 研究范畴

本文识别并综合了现有二级研究中关于可信的研究对象 (关注点), 建立了可信人工智能研究范畴元模型, 如

图 5 所示. 元模型以软件系统及可信为核心, 围绕 AI 系统的可信, 即一种特定的软件系统的可信展开. 可信作为软件系统的一般需求通常包含多个可以由质量属性表示的特征组成, 相关的实践可以提升或保障相应的质量属性, 质量属性的评估需要使用一定的度量, 每一类的实践和度量通常包含具体的方法, 同时也需要使用一定的工具. 对软件系统而言, 主动攻击技术和软件所处的外部环境会导致特定的威胁和挑战, 进而对系统的质量属性造成负面影响. AI 系统作为特定的软件系统, 其可信性存在着一定的特殊性: 首先, 不同于传统软件, AI 模型是 AI 系统的一个重要组成部分; 其次, 由于 AI 系统的外部环境的特殊性, 其需要考虑以人本主义为主的伦理道德对环境的约束. 威胁与挑战除了源于外部环境, 还可能来源于一些黑客的主动攻击. 可信性、威胁及挑战等最终都指向能够描述可信性的具体质量属性. 更进一步的研究重点是, 采用怎样的度量方法或使用怎样的工具能够评估具体质量属性对于可信的满足程度; 以及采用何种实践能够保障甚至改进具体质量属性对于可信的满足程度. 实践可能基于具体的工具或方法, 具体评估或保障方法可能会应用一些 AI 模型或算法.

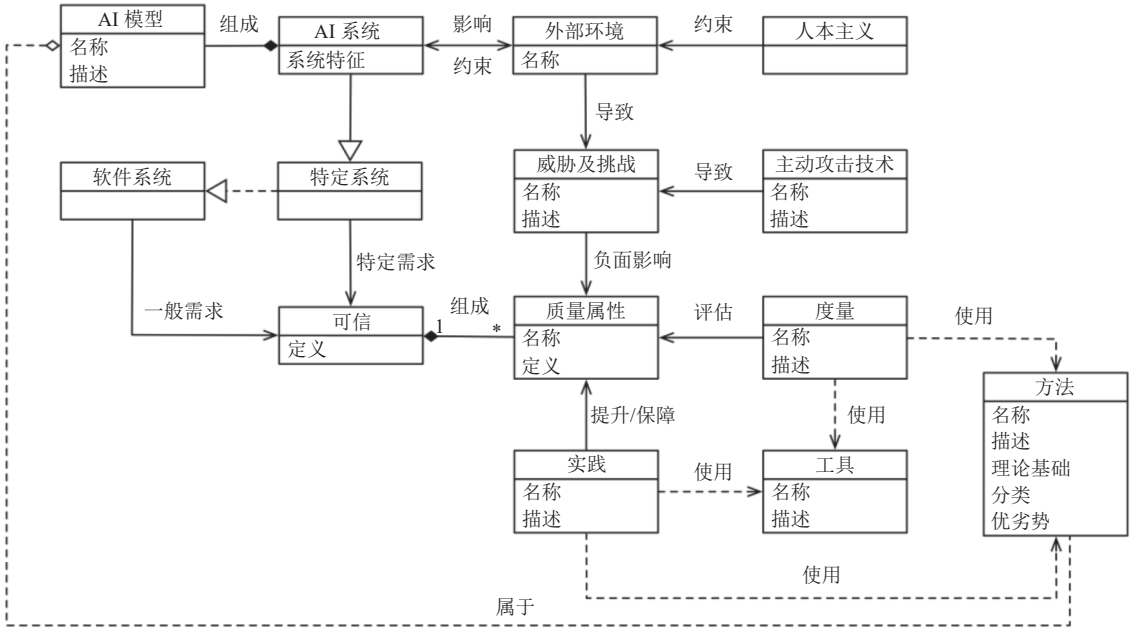

图 5 可信人工智能研究范畴元模型

对于同样的研究对象, 不同研究可能从多种不同的角度开展研究, 所以本文借鉴了文献 [60] 中的方法, 即采用 5W+1H 模型进一步分析了现有二级研究对图 5 中各研究对象的研究情况, 如表 4 所示. 在表 4 中, Wt, Wy, Wo, Wn, Wr, H 分别表示 What, Why, Who, When, Where, How. 本文参考了文献 [61,62] 中的定义, 并基于上下文, 定义如下.

What (Wt): 有哪些研究对象? 它们是什么? 例如, 文献 [26] 的一个研究问题是“有哪些用于开发和评估 AI 系统的工具”.

Why (Wy): 为什么实现这些研究对象? 例如, 文献 [16] 在研究质量属性时, 研究了“为什么可信 AI 需要这些质量属性”.

Who (Wo): 谁提出或实现了这些研究对象? (只是引用了作者不属于该类别, 应有如机构分布等类似的对于作者的分析). 例如, 文献 [37] 中的一个研究问题是“谁在领导基于 AI 的软件的质量的研究 (学术界, 工业界, 及两者的协作)”.

When (Wn): 在什么情况下考虑或使用这些研究对象? 例如, 文献 [27] 对于公平性的保障改进方法, 研究了“不同 AI 模型应该采用何种方法”.

Where (Wr): 在什么 AI 系统开发阶段考虑或使用这些研究对象? 例如, 文献 [16] 在研究多种质量属性的保障改进实践时, 研究了“各实践应该映射到 AI 系统开发的哪些阶段 (如, 建模前、建模过程中和建模后) ”.

How (H): 如何实现或如何应用这些研究对象? 例如, 文献 [24] 的一个研究目标是“如何通过实践来保障 AI 系统可信性”.

表 4　基于 5W+1H 模型的二级研究的研究关注点

文献	可信	质量属性	度量/评估	实践	方法	工具	威胁及挑战	AI模型	主动攻击技术	人本主义	特定系统
[14]	Wy, H	Wt	—	—	Wt	—	Wt	—	—	—	—
[15]	—	—	Wt	—	Wt	—	Wt	—	Wt	—	—
[16]	Wy, H	Wt, Wy, H	Wt	Wt, Wr	Wt, Wr	Wt	—	—	—	Wt	—
[17]	—	—	—	—	Wt	—	Wt	Wt, H	Wt	—	—
[18]	—	—	—	—	Wt, Wr	—	—	Wt	—	—	—
[19]	—	—	Wt	—	Wt, Wy	—	—	H	—	—	—
[24]	H	—	—	Wt, Wr	—	—	—	—	—	—	—
[25]	—	—	—	—	Wt	Wt	Wt, H	—	Wt, Wr	—	—
[26]	—	Wt	—	—	—	Wt	Wt	—	—	Wt, Wy	—
[27]	—	Wt	—	—	Wt, Wr, Wn	Wt	Wt, Wr	Wt	—	—	—
[28]	—	—	—	—	Wt, Wr	—	Wt, Wr	—	Wt, Wr	—	—
[29]	—	—	—	Wt	Wt, Wr	—	Wt	—	Wt	—	—
[30]	—	—	—	—	Wt, Wr	—	Wt	—	Wt	—	—
[31]	—	—	—	—	Wt	—	Wt	—	—	—	Wt
[32]	—	—	—	—	Wt, Wr	—	H	—	—	—	—
[33]	—	—	—	—	Wt, Wr	—	—	—	—	—	—
[34]	—	—	—	—	Wt, Wn	—	—	Wt, H	—	—	—
[35]	—	—	—	—	Wt, Wn	—	Wt	—	—	—	Wt
[36]	—	—	—	—	Wt, Wn	—	Wt	—	Wt	—	—
[37]	—	Wt, Wo	—	—	Wt	—	Wt	Wt	—	—	—
[38]	—	—	Wt, Wn	—	—	—	—	—	—	—	—
[39]	—	—	—	—	Wt	—	Wt	—	Wt	—	—
[40]	—	—	—	—	Wt, Wn	—	Wt	—	—	—	Wt
[41]	—	—	—	—	Wt, Wr	—	—	Wt, Wr	—	—	—
[42]	—	—	—	—	Wt, Wn	—	Wt	—	Wt	—	—
[43]	—	—	—	—	Wt, Wn	—	Wt	Wt, Wn, H	Wt, Wn	—	—
[44]	—	—	—	—	Wt, Wn, H	—	Wt	Wt, H	—	—	—
[45]	—	—	—	—	Wt, Wn	—	—	—	Wt	—	—
[46]	—	—	—	—	Wt, Wr, H	—	Wt	—	Wt	—	—
[47]	—	—	—	—	Wt, Wn	—	Wt, Wn	Wt	Wt	—	—
[48]	—	Wt	—	Wt, Wr	Wt	—	Wt	Wt, H	Wt	—	—
[49]	—	—	Wt	Wt, Wr, H	Wt	—	—	Wt, Wn, H	—	—	—
[50]	—	—	—	—	Wt, Wn	—	—	—	Wt, Wn	—	—
[51]	—	Wt, Wy	—	—	Wt, Wn	—	—	H	—	—	—

注: Wt, Wr, Wn, Wy, H分别表示What, Where, When, Why, How, 代表具体研究问题的不同角度

例如文献 [27] 在方法这一列为“Wt, Wr, Wn”, 表示针对方法研究了现有一级研究采用了哪些方法 (Wt), 这些方法适用的场景和使用的条件是什么 (Wn), 以及这些方法应用于 AI 系统开发的哪个阶段 (Wr). 例如文献 [14] 在可信这一列为“Wy, H”, 表示针对可信研究了为什么需要可信 (Wy) 和如何实现可信 (H). 通常, 是什么 (Wt) 是最基本的研究问题, 绝大部分研究都会首先从是什么的角度讨论一个问题. 对于为什么 (Wy) 的问题更多是在研究动机中已经阐明的, 尤其是对于实践、方法等研究对象, 一般不会从为什么的角度进行研究. 对于如何实现或如何应

用 (H) 这类较为细节的问题, 在二级研究中同样少有研究, 因为通常, 这类问题会被抽象为在什么情况下考虑或使用 (Wn) 以及在什么阶段考虑或使用 (Wr) 的问题. 仅有一篇文献[37]专门设立研究问题以分析作者信息 (Wo), 然而, 这虽然是一个几乎没有研究难度的问题, 但通常又是读者想了解的信息. 对于具体的研究对象, 可以看到, 更多的研究关注的是威胁及挑战和应对威胁及挑战的方法, 这是由于大部分研究关注的都是安全性和隐私性的主题, 而这两个质量属性涉及众多威胁及基于不同技术路线的应对措施, 将在后续章节具体讨论.

RQ1 要点总结: 可信 AI 的二级研究始于 2017 年并呈稳定增长趋势; 全球除了南美洲和南极洲 (无国家) 均有研究者从事可信 AI 相关的二级研究. 已有可信 AI 的二级研究主要关注点在于质量属性的保障改进方法及威胁与挑战, 然而关注质量评估方法的研究较少, 从为什么 (Why) 的角度讨论问题以及对一级研究的作者信息 (Who) 进行分析的二级研究也相对较少.

3 可信人工智能系统的质量属性框架 (RQ2)

本文从现有二级研究中共识别到了 21 种可信相关的质量属性, 其被研究的频率分布如图 6 所示, 其中隐私性、安全性和可解释性等与 AI 系统区别于一般软件系统的特征相关的质量属性是被研究最多的, 而与一般软件系统的可信相关的可维护性、可移植性等质量属性仅在一项二级研究[37]中进行了讨论. 欧盟发布的可信 AI 伦理道德指南[10]中建议可信 AI 系统需满足 4 项基本原则, 分别是尊重人类自治原则 (the principle of respect for human autonomy), 伤害预防原则 (the principle of prevention of harm), 公平性原则 (the principle of fairness) 以及可解释性原则 (the principle of explicability). 在识别到的 21 种质量属性中, 有 15 种可以分类到这 4 项原则中, 其中安全性包含了 3 个子属性, 可解释性包含了 1 个子属性. 而其余 6 种质量属性属于一般软件系统的可信性需求, 与 ISO/IEC 25010:2011[63]中提出的质量模型相符. 参考可信 AI 伦理道德指南[10]和 ISO/IEC 25010:2011[63], 以及二级研究中的描述, 本文构建了一个可信 AI 质量属性框架, 如图 7 所示. 需要说明的是, 在 ISO/IEC 25010:2011[63]中提出了 8 种软件质量属性, 分别是兼容性, 可维护性, 功能适用性, 可移植性, 性能效率, 可靠性, 易用性和安全性. 其中可靠性和安全性均属于伤害预防原则的范畴, 但实际上本研究只识别到了 AI 系统在安全性有特定的需求, 而在可靠性上的需求与一般软件一致. 此外, 本文将不属于 4 项原则的其他质量属性分类到了一般软件质量属性类别中. 本文所提出的可信 AI 质量属性框架与 ISO/IEC 25010:2011[63]中的质量属性框架之间的差异, 可以反映出相较于一般软件系统, 对 AI 系统的可信提出的新的要求. 以下将对各质量属性按框架中的 5 种类别进行展开介绍.

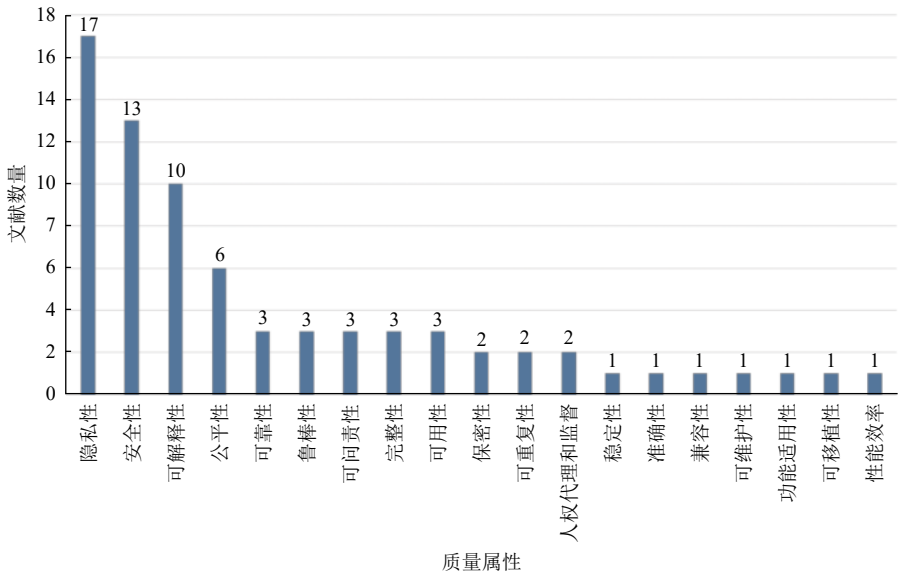

图 6 质量属性相关研究数量分布

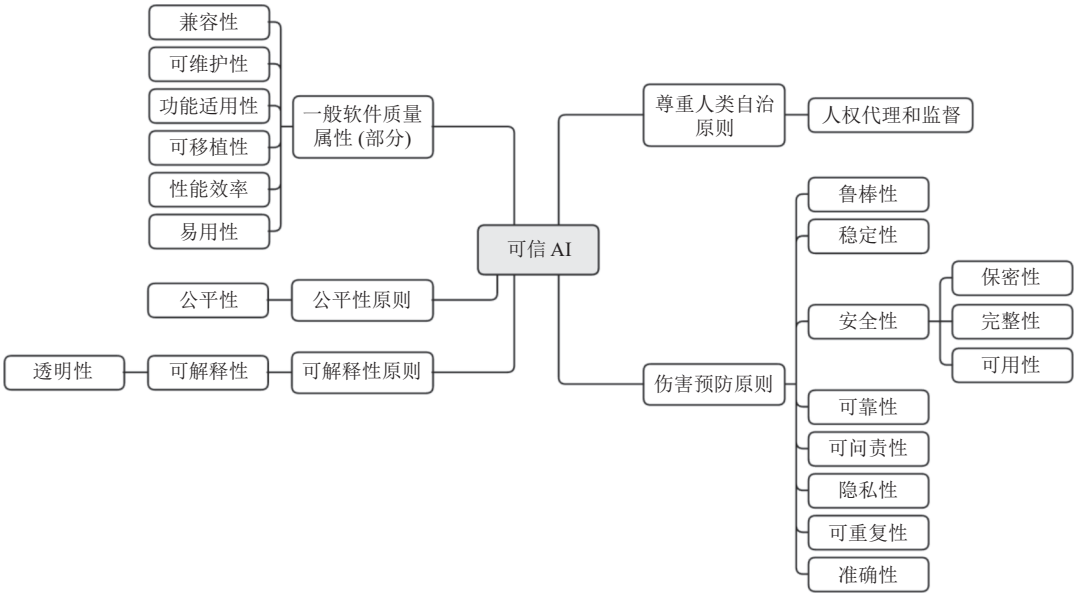

图 7 可信人工智能系统的质量框架

3.1 尊重人类自治原则

人工智能系统的用户应能够保持充分和有效的自我决策, AI 系统的设计应该以补充和增强人的能力为目的[10]. 在获取到的质量属性中, 只有一种质量属性, 即人类代理和监督可以映射到该原则下.

● 人类代理和监督 (human agency and oversight). 有 2 篇文献[14,16]讨论了这一质量属性, 其中文献 [14] 将尊重人类自治原则视为设计可信人工智能系统最重要的原则. 文献 [14,16] 均指出人工智能系统应该赋予人类新的权利而不是取代人类. 综合两项研究[14,16]中的定义, 本文将人类代理和监督定义为人工智能系统应该始终处于人类的控制之下以避免系统对人的生命健康安全或其他基本权利造成危害. 人类代理和监督质量属性的保障和改进, 在于解决人工智能系统对人类行为造成影响的问题. 应基于风险以及社会和环境影响因素, 保障和尊重人类自治的基本原则, 保证人类对于系统决策过程的参与达到一定程度.

3.2 伤害预防原则

伤害预防原则涉及 3 方面的要求[10], 包括: (1) AI 系统不应对用户或环境造成伤害或加剧伤害; (2) AI 系统及其运行环境必须是可靠的; (3) AI 系统在技术上必须是健壮的. 基于上述要求, 本文将鲁棒性, 稳定性, 安全性 (含保密性、完整性、可用性等 3 个子属性), 可靠性, 可问责性, 隐私性, 可重复性和准确性共 11 个质量属性映射到该原则下.

● 安全性 (security). 有 13 篇文献[14,15,17,25,26,28,31,36,37,40,43,45,50]讨论了安全性, 其中 4 项研究[14,25,26,37]对安全性提出了定义性描述. 文献 [25] 将保密性 (confidentiality), 完整性 (integrity) 和可用性 (availability) 视作安全性的主要目标, 据此本文将这 3 个目标定义为安全性的子属性. 文献 [26] 中指出人工智能系统的安全性问题主要围绕人工智能系统安全的验证与确认问题、在易受攻击环境下的人工智能模型的自我保护问题等方面. 文献 [14] 对安全性的描述侧重于 AI 模型而不是系统的安全, 即安全性确保外部攻击不会改变或影响 AI 模型的决策. 文献 [37] 采纳了 ISO/IEC 25010:2011[63]中对于安全性的描述, 即产品或系统保护信息和数据的程度, 及便于个人或其他产品或系统具有与其授权类型和级别相适应的数据访问程度. 此外, 部分研究如文献 [40,43], 虽然没有明确定义安全性, 但在阐述安全性相关实践时, 讨论了人工智能系统的数据安全性.

综上, 本文对人工智能系统的安全性给出如下定义: 人工智能系统抵御非法操纵或攻击以保护数据、模型和其他组件免受外部攻击影响的程度, 其子属性按目标划分包括保密性, 完整性和可用性. 其中, 保密性指系统保护

训练数据中的机密信息以及模型参数等信息以避免误用和未经授权的访问的程度[25,43]; 完整性指系统防止未经授权的修改以确保系统数据一致性和准确性的程度[14,25,43]; 可用性则指人工智能系统在面对异常输入、恶意输入等异常情况仍能提供正常服务的能力[25,43].

● 隐私性 (privacy). 有 17 篇文献[14−18,25,28−30,36,39,40,42,43,45−47]讨论了隐私性, 这是受关注度最高的质量属性. 6 项研究[14,16,18,30,45,46]对隐私性提出了定义性描述. 文献 [14] 从 AI 模型的角度考虑隐私性, 指出模型的隐私性意味着 AI 模型应该保护它所训练的数据和使用它的用户的身份. 文献 [16] 关注 AI 系统的数据隐私, 指出隐私性确保由个人分享或 AI 系统收集的敏感数据受到保护. 文献 [18] 指出 AI 模型 (特指深度学习模型) 的隐私性问题与敏感的输入数据以及共享训练等因素有关. 文献 [30,45] 指出机器学习的隐私性与训练数据以及模型和模型参数等资产不受推断、窃取和干预等影响的能力有关. 文献 [46] 提供了一个更加全面的理解, 将 AI 模型的隐私分为训练数据隐私、模型隐私与预测结果隐私.

综上, 本文将隐私性定义为: 人工智能系统保护用户信息, 训练数据, 模型相关信息以及决策结果等不受窃取、推断和干扰的能力. 从给出的定义可以看出, 人工智能系统的隐私性与信息安全性密切相关, 因此超过一半的二级研究[14,15,17,25,36,40,43,45]在讨论隐私性的同时会讨论安全性.

● 可靠性 (reliability). 有 3 篇文献[16,37,40]讨论了可靠性. 文献 [16] 将可靠性描述为系统按预期运行的程度, 即在规定的限度内, 系统没有任何故障, 对相同的输入持续产生相同的输出的程度. 这一定义与文献 [37] 中采纳的 ISO SQuaRE[63]中的定义类似, 后者是系统、产品或组件在规定条件下在规定时间内执行规定功能的程度. AI 系统作为一种特定的软件系统, 也需要满足一般软件系统的质量属性. 由于没有在文献中识别到对于 AI 系统的可靠性所特有的需求, 本文同样建议采用 ISO SQuaRE[63]中的定义.

● 鲁棒性 (robustness). 有 3 篇文献[15,37,38]讨论了鲁棒性. 其中文献 [37] 没有对鲁棒性提供明确的定义. 文献 [15] 中将鲁棒性定义为 AI 系统面对恶意攻击的弹性程度. 文献 [38] 中认为鲁棒性是指 AI 模型不受攻击者影响或不对异常值进行错误分类的能力. 与可靠性相比, 鲁棒性强调面对不正常输入的稳定性, 而可靠性强调行为的正确性. 与安全性相比, 鲁棒性强调 AI 系统在面对扰动时保持正常运行的能力, 而安全性强调系统避免系统扰动而产生危害.

综上, 本文将鲁棒性定义为: AI 系统面对异常情况时, 在软件系统层面和 AI 模型层面均保持正常运行的能力.

● 稳定性 (stability). 仅有 1 篇文献[19]讨论了稳定性. 文献 [19] 认为如果微小的扰动 (如数据源本身的噪声) 不会对模型的决策造成误导, 则认为该模型是稳定的. 将其扩展到更广泛的系统层面, 当系统面对扰动时模型不受误导则表明系统仍处于稳定状态.

借鉴文献 [19] 的观点, 本文将稳定性定义为: 当 AI 系统面对微小扰动时, 系统仍能保持稳定状态的能力.

● 准确性 (accuracy) 仅有 1 篇文献[14]讨论了准确性. 文献 [14] 中将模型的准确性定义为模型正确预测结果的能力, 同时文中指出 AI 系统的准确性应保持在一定阈值之上以保证系统能够做出可靠的决策.

综上, 本文将准确性定义为: AI 系统提供正确决策的能力, 其准确性应保持在一定阈值之上.

● 可问责性 (accountability). 有 3 篇文献[14,16,19]讨论了可问责性. 文献 [14] 关注 AI 模型的可问责性, 其指出模型的可问责性指模型向系统用户证明其所做决策的能力. 文献 [16,19] 从 AI 系统的层面提供了类似的定义, 即向与系统交互的不同用户解释和证明系统的决策和行为的能力. AI 系统的决策和行为是由其模型推导出来的, 因此从模型角度或从系统角度进行定义, 其意思是相同的. 本文采纳文献 [16,19] 中的定义.

● 可重复性 (reproducibility). 有 2 篇文献[14,19]讨论了可重复性. 这 2 篇文献对可重复性的定义是相似的, 其中文献 [14] 的定义更加全面, 其从系统的角度将可重复性描述为, 如果向系统提供相同的输入参数和条件, 系统做出的所有决定都可重复. 而文献 [19] 中从模型的角度将可重复性描述为, 如果一个模型在相同的数据集上运行数次, 能重复得到相同的结果. 文献 [14,19] 中对于可重复性的描述是一种完全可重复和完全不可重复的二元判断.

从具有统计意义的度量的角度, 本文建议将可重复性定义为: AI 系统在相同的数据集、参数和条件上运行多次能够得到相同的运行结果的程度 (概率).

3.3 公平性原则

公平性原则要求人工智能系统的开发、部署和使用必须公平. 尽管公平性存在不同的定义, 但公平性的实质是要求利益和成本的平等, 通过公平分配以确保个体或群体不受不公平的偏见和歧视[10]. 在识别到的所有质量属性中, 只有一个质量属性, 即公平性可以映射到该项原则下.

● 公平性 (fairness). 有 6 篇文献[14,16,26,27,48,49]讨论了公平性或相似概念. 文献 [26] 将公平性 (fairness), 平等性 (equality) 和无偏见性 (unbiasedness) 视为相同含义并给出如下描述, 如果对于两个具有相同的特征但具有不同的敏感属性 (如种族, 宗教或性别等) 的不同个体, 模型应做出相同的决定. 文献 [14] 中讨论了非歧视性 (non-discrimination), 即能够平等地对待所有用户, 而不因社会阶层等信息区别对待, 事实上, 这与文献 [26] 中的定义是类似的. 文献 [16] 中指出系统的公平性应确保不存在基于任何固有的或获得的与决策无关的特征而对任何个人或群体产生偏见和歧视. 文献 [27] 指出公平性的要求包括对相似个体给出相似预测的个体公平性, 平等对待不同群体的团体公平性以及平等对待子群的子群公平性. 文献 [48] 总结了公平机器学习算法中的 3 类公平性定义, 包括感知公平性, 统计公平性和因果公平性, 这更多是从更细节的算法实现层面而不是从基于伦理道德的外部评价层面提供的定义.

综上, 本文将公平性定义为: AI 系统平等对待任何个体、团队或子群, 不因与决策无关的社会学中常见的敏感属性 (如种族, 宗教或性别等) 导致任何歧视或偏见的能力. 与公平性同义的术语包括平等性、无偏见性和非歧视性.

3.4 可解释性原则

可解释性对于建立和维护用户对 AI 系统的信任至关重要, 可解释性原则要求 AI 系统的能力和目的需要被公开, 且需要尽可能使直接和间接受影响的人能够理解系统的决策过程[10]. 在获取到的质量属性中, 只有一个质量属性, 即可解释性可以映射到该原则下. 事实上, 本文识别到了可解释性 (interpretability), 可解释性 (explainability), 透明性 (transparency). 但这三者是近义的, 其中文献 [19,33] 指出“interpretability”和“explainability”在一级研究中经常被互换使用, 而对“transparency”实际的研究内容与前两者也没有本质差异. 所以本文将三者统一归类为了可解释性. 根据这三者在不同研究中的定义, 它们确实存在一些细微的差异, 但这体现的是可解释的不同方面或不同程度, 以及不同研究者描述的角度. 而本文将综合现有的定义, 对可解释性给出一个更广义的定义. 此外, 可问责性与可解释性的定义类似, 文献 [16] 指出可问责性要求监控 AI 系统的运行从而避免其造成伤害, 而实现这一点的关键是要发现为算法负责的责任方, 因此, 本文认为可问责性强调面对伤害时的责任问题, 厘清责任所需的信息与实现可解释所需的信息存在重叠但并不完全相同. 所以本文将可问责性归为伤害预防原则, 而可解释性归为可解释性原则.

● 可解释性 (interpretability/explainability). 有 10 篇文献[14,16,19,26,33−35,41,44,51]讨论了可解释性或相似概念. 文献 [26,41] 中将“interpretability”描述为使人类理解一个模型如何工作的能力. 文献 [26] 中将“explainability”描述为 AI 系统的内部结构能够以人类可以理解的方法进行解释的程度. 文献 [41] 同样认为模型的“explainability”与 AI 系统的内部逻辑有关. 但文献 [34] 仅使用了“interpretability”一词, 其认为“interpretability”包含模型透明性和模型功能性两个方面, 其中模型透明性指能否使用输入数据和参数模拟系统决策过程以及模型参数和算法是否可以被用户理解; 模型功能性则侧重模型的可视化和能否给出局部解释. 事实上, 在文献 [34] 的定义中, 模型透明性与文献 [26,41] 对于“explainability”的理解基本一致, 而模型功能性与文献 [26,41] 对于“interpretability” 的理解基本一致. 文献 [19] 也使用了“interpretability”一词, 并认为“interpretability”有助于揭示哪些特征促使系统的特定输出的产生以及如何通过模拟系统决策过程以获得决策, 这与文献 [26,41] 中对于“explainability”而不是“interpretability”的解释更加接近. 总体而言, 可以通过现有研究得到一个相对一致的结论, 即可解释性包含两个方面, 一方面是决策行为的可解释性, 即系统总体和各步骤的决策目的、输入、输出能够被人类理解; 另一方面是内部结构的可解释性, 即系统进行具体决策行为时, 其决策过程能够被人类理解. 内部结构的可解释性实际与透明性是近义的. 有 4 篇文献[14,19,26,35]使用了“transparency”一词. 文献 [26] 指出由于缺乏明确的声明性表示, 大多数 AI 系统的内部状态过于不透明导致无法产生具有内在可解释性的元数据, 这里透明性和可解释性之间的差异在于, 前者只需要结构能够被人类观测到, 而后者在要求可观测的基础上还能够被人类理解. 文献 [14] 认为保障 AI 系统的透明性可

以向用户提供一个清晰的蓝图以帮助用户清楚地理解模型所做的决策以及了解模型的内部结构; 文献 [35] 将模型的透明性定义为能够观测到模型的参数和预测是如何生成的. 这两项研究中定义的透明性与可解释性没有本质的差异, 只是视角的不同. 所以最终本文将“interpretability”“explainability”和“transparency”统一归类为了可解释性, 其包含了子属性透明性.

综上, 本文将可解释性定义为: 可解释性包含两个方面, 一方面是决策行为的可解释性, 即系统总体和各步骤的决策目的、输入、输出能够被人类理解; 另一方面是内部结构的可解释性, 即系统进行具体决策行为时, 其决策过程能够被人类理解. 其中内部结构可解释性的前提是透明性, 且通常情况下二者的同义的.

3.5 其他的一般软件质量属性

本文将无法映射到可信 AI 的 4 项基本原则之下的 6 种质量属性分类到了一般软件质量属性类别中, 以保证层次结构的完整性. 这 6 种质量属性都只在文献 [37] 中得到了讨论, 包括兼容性, 可维护性, 功能适用性, 易用性, 可移植性和性能效率.

文献 [37] 在描述上述质量属性时, 采用了 ISO/IEC 25010:2011[63]中的定义. 本文仅作简要转述.

● 兼容性 (compatibility) 指产品、系统或组件与其他产品、系统或组件在交换信息且共享相同硬件和软件环境的情况下, 执行其所需功能的程度.

● 可维护性 (maintainability) 指产品或系统可通过修改以改进、纠正或适应环境和要求变化的有效性和效率.

● 功能适用性 (functional suitability) 指在特定条件下使用时, 产品或系统提供满足规定和隐含需求的功能的程度.

● 易用性 (usability) 指在指定的使用环境中, 指定用户可以使用产品或系统来实现指定目标的有效性、效率和满意度.

● 可移植性 (portability) 指系统、产品或组件可从一个硬件、软件或其他操作和使用环境转移到另一个环境的有效性和效率.

● 性能效率 (performance efficiency) 指在规定条件下与使用的资源数量相关的性能效率.

RQ2 要点总结: 通过数据抽取与合并, 本文获得了 21 个质量属性, 并整理形成了如图 7 所示的可信人工智能系统的质量框架. 综合二级研究中的描述和定义, 本文统一了每个质量属性的定义性描述.

4 可信性的评估度量方法 (RQ3)

在研究过程中, 本文仅获取到了可解释性, 鲁棒性, 公平性和隐私性的评估度量方法, 如表 5 所示. 其他质量属性的评估度量方法均无二级研究进行过讨论. 后文中所有“A”开头的文献研究均来自二级研究, 本文整理一级研究文献内容已在线公开, 可访问网址 https://docs.google.com/document/d/10N5pt0UO-BxPgi7S7oMiLHCEpHRb82zpyaBJftGShPQ 获取.

表 5 可信性的评估度量方法

质量属性	评估度量方法	文献
可解释性	计算度量[A1,A2]	[19]
	认知度量[A3–A7]	[16,19]
	其他可解释性评估方法[9, A8, A9]	[16]
鲁棒性	常见机器学习算法预测性能评估度量[A10–A12]	[15]
	基于攻防的鲁棒性评估方法[A12–A25]	[15,38]
公平性	差异影响度量[A26]	[49]
	人口均等度量[A27]	
	均等几率度量[A28]	
	机会均等度量[A28]	
	个体公平度量[A29, A30]	
隐私性	差分隐私[A31–A33]	[45]

4.1 可解释性评估方法和度量

仅有 2 篇文献[16,19]讨论了人工智能系统可解释性相关的度量方法. 已有研究提出了不同的可解释性保障方法为人工智能系统生成解释 (参见第 5.2 节), 从而提高人工智能系统的可解释性. 具体而言, 文献 [16,19] 讨论的都是用于评估模型可解释性保障方法生成的解释的方法和度量. 依据文献 [19], 本文将这些评估可解释性保障方法生成的解释的方法分为计算度量、认知度量和其他解释评估方法 3 类.

● 计算度量 (computational metrics) 通常使用方程和现有数据进行, 可以在没有人类参与的情况下作为构建具备可解释性的技术的指导方针[A1], 该度量也被称为是可解释 AI 系统生成解释质量的数字度量 (numeric indicator)[A2]. 文献 [19] 并没有详细讨论该度量, 具体定量方法可参考文献 [A1,A2].

● 认知度量 (cognitive metrics) 则是由人类受试者进行评估确定, 要求受试者从自己的角度确定最佳的解释. 例如, Hoffman 等人[A3] 从优良性 (goodness)、满意度 (satisfaction)、易理解性 (understanding) 和效率度量 (efficiency) 等 4 方面讨论了常见的度量方法, 这些方法可以帮助受众从自己的角度对生成的解释进行评估, 其中优良性包括清晰度和精度等影响解释好坏的因素, 满意度指用户认为可解释性方法生成的解释帮助他们理解人工智能系统或过程的程度, 易理解性与可解释性方法对人工智能系统的解释的精神模型、兴趣和信心有关, 效率度量的目标是确定系统在有效执行技术设计任务产生解释方面的成功程度. 更多具体方法可详细参考原文献[A4–A7].

● 其他可解释性评估方法. ISO TR 24028[9]中提出了 3 种对生成解释的度量, 包括一致性, 连续性和选择性. 一致性要求对相似的预测产生相同或相似的解释; 连续性用来衡量输入变量重要性的变化在解释方法的特征分数的反应程度; 选择性则意味着可解释性方法应该从特征空间中选择影响最大的特征. Arya 等人[A8] 根据解释的内容、级别以及生成方式, 将这些方法组织在决策树中. Sokol 等人[A9] 提出了一份用于系统评估可解释方法的说明书, 根据功能、操作、可用性、安全性和验证维度评估可解释模型.

4.2 鲁棒性评估方法

仅有 2 篇文献[15,38]讨论了鲁棒性的评估方法. 鲁棒性的评估通常是定量的, 例如通常理解的模型预测性能的评估属于鲁棒性评估.

● 常见机器学习算法预测性能评估度量. 文献 [15] 指出常见准确性, 精确度, 召回率和 $F1$ 得分等性能指标已被研究人员广泛用于各种机器学习算法和应用中, 以对系统应对攻击时的性能进行评价. 例如, Dunn 等人[A10] 在物联网环境的应用中, 使用准确性, 精准度, 假阳性率和真阳性率 4 种性能指标来评价投毒攻击对梯度增强, 随机森林, 朴素贝叶斯和前馈深度学习 4 种模型完整性的负面影响; 相应地, 这些指标也可用于评估系统的防御机制, 从而度量系统的鲁棒性. 更多关于这些性能评估度量在评估人工智能系统鲁棒性中的应用方法参考原文献[A11, A12].

● 基于攻防的鲁棒性评估方法. 除常见的预测性能评估以外, 文献 [15] 中介绍了已有的鲁棒性评估方法, 例如 Biggio 等人[A12] 提出了一个经验安全性评估框架 (empirical security evaluation framework), 该框架由一个可以定义任何攻击场景的对抗模型、相应的数据分布模型以及用于经验性能评估的训练集和测试集生成方法组成, 为不同分类器、学习算法和分类任务的安全性评估提供了定量和通用的基础. Katzir 等人[A13] 提出了一种称为模型鲁棒性评分 (model robustness score, MRB score) 的度量方法, 该方法基于攻击成本和特征变换成本对攻击者的能力进行建模, 量化应用于网络安全的各种机器学习分类器的弹性以评估机器学习模型的相对弹性. 文献 [38] 指出模糊测试、故障注入等传统软件的测试技术也被应用于神经网络的鲁棒性评估, 例如, Guo 等人[A14] 提出了一个名为 DLFuzz 的深度学习系统的差分模糊测试框架, 该框架不断对输入进行变化, 以最大化神经元覆盖率和原始输入与变异输入之间的预测差异. 更多具体方法请参考原文献[A15–A25].

4.3 公平性相关度量

文献 [49] 中讨论了影响 AI 系统公平性的度量. 具体包括差异影响、人口均等、均等几率、机会均等和个体公平.

● 差异影响度量 (disparate impact)[A26] 要求对两个群体产生的预测结果的阳性预测率之间有着较高的比值, 这意味着在不同群体之间的阳性预测比例是相近的. 该指标的数值越高, 则不同群体之间的比率越接近. 根据文献[49], 该度量的计算公式如下.

$$\frac{P[\hat{Y}=1 \mid S \neq 1]}{P[\hat{Y}=1 \mid S=1]} \geqslant 1-\varepsilon \tag{1}$$

其中, S 表示种族、性别等受保护的属性, $S=1$ 为特权 (预测结果的阳性预测率较高) 群体, $S \neq 1$ 则表示非特权群体, $\hat{Y}=1$ 表示预测结果为阳性.

一种典型的差异影响度量是 Feldman 等人[A26] 提出的“80% 规则 (rule)”, 该规则要求任何种族、性别或群体 (即敏感属性) 的接受率至少为接受率最高群体的 80% 以上.

● 人口均等度量 (demographic parity) 的度量方法与差异影响类似, 但在度量时采用的是差异而不是比值[A27]. 该度量的计算公式为:

$$|P[\hat{Y}=1 \mid S=1]-P[\hat{Y}=1 \mid S \neq 1]| \leqslant \varepsilon \tag{2}$$

文献 [49] 指出差异影响度量和人口均等度量这两种度量方法存在一定的缺陷, 因为当不同群体之间的实际阳性结果显著不同时, 一个完全准确的分类器可能会被度量为不公平.

● 均等几率度量 (equalized odds)[A28] 计算假阳性率 (false-positive rates, FPRs) 以及两个群体之间的真阳性率 (true-positive rates, TPRs) 的差异以克服上述差异影响度量和人口均等度量的缺陷. 该度量可以通过如下公式计算.

$$|P[\hat{Y}=1 \mid S=1,\ Y=0]-P[\hat{Y}=1 \mid S \neq 1,\ Y=0]| \leqslant \varepsilon \tag{3}$$

$$|P[\hat{Y}=1 \mid S=1,\ Y=1]-P[\hat{Y}=1 \mid S \neq 1,\ Y=1]| \leqslant \varepsilon \tag{4}$$

其中, 公式 (3) 和公式 (4) 分别要求两个群体之间假阳性率和真阳性率的差的绝对值以 ε 为界.

● 机会均等度量 (equal opportunity) 要求不同群体之间的真阳性率相似[A28], 该度量与均等几率度量类似, 但只关注于真阳性率. 该度量的计算公式如公式 (5) 所示.

$$|P[\hat{Y}=1 \mid S \neq 1,\ Y=1]-P[\hat{Y}=1 \mid S=1,\ Y=1]| \leqslant \varepsilon \tag{5}$$

● 个体公平度量 (individual fairness) 要求相似的个体受到相似的对待. 该度量在定义公平性时不仅考虑敏感属性, 还考虑个体的其他属性[A29]. 个体公平度量可以用公式 (6) 描述.

$$\left|P\left(\hat{Y}^{(i)}=y \mid X^{(i)},\ S^{(i)}\right)-P\left(\hat{Y}^{(j)}=y \mid X^{(j)},\ S^{(j)}\right)\right| \leqslant \varepsilon;\quad \text{if}\quad d(i,j) \approx 0 \tag{6}$$

其中, i 和 j 分别表示两个个体, $S^{(\cdot)}$ 表示个体的敏感属性, $X^{(\cdot)}$ 表示与敏感属性相关的属性, $d(i,j)$ 为根据预期用途定义的相似度的距离度量.

文献 [A30] 指出要对个体之间的相似性进行度量, 还需要考虑特征和标签之间关系的假设.

4.4 隐私性评估方法

仅有一项研究[45]讨论了隐私性的度量方法. 文献 [45] 讨论了一级研究中使用差分隐私来评估隐私性风险的方法. 例如, McSherry[A31] 基于差分隐私的可组合性, 通过计算隐私消耗的总和来度量隐私性风险. Abadi 等人[A32] 采用标准 Markov 不等式来跟踪隐私损失, 在经验上获得更加严格的隐私损失约束. 文献 [45] 指出, 上述方法仅限于使用差分隐私框架的系统. Long 等人[A33] 提出了差分隐私训练 (DPT) 的方法, 该方法可以用于评估未使用差分隐私方法的分类器的隐私风险. 差分隐私方法更多地被应用于 AI 系统的隐私性保障, 本文将在后续章节对这一用途进行讨论.

RQ3 要点总结: 本文识别到了可解释性, 鲁棒性和公平性的相关评估度量以及鲁棒性和隐私性的评估方法. 对于其他质量属性的度量和评估方法, 现有二级研究鲜有讨论. 对于识别到的度量方法, 现有二级研究中的讨论也较为粗略, 读者仍需以本文或相关二级研究为入口, 从相关一级研究中获得更详尽的信息.

5 可信性的保障改进实践 (RQ4)

现有二级研究中, 仅有 7 种质量属性的保障改进实践或方法被讨论了, 分别是安全性, 鲁棒性, 可解释性, 隐私性, 公平性, 可问责性和人类代理和监督. 其中安全性和鲁棒性难以作出明确的区分, 因为同样的实践或方法可能在有些文献中被认为是保障安全性的, 而在另一些文献中被认为是提升鲁棒性的, 而它们所表达的含义是相同的, 所以本文将安全性及鲁棒性的实践一起讨论. 本文将实践定义为一个比方法更抽象的概念, 即一类实践可能包含

多种实现方法. 本节将分别对每种质量属性的保障改进实践及相关方法展开介绍.

5.1 安全性及鲁棒性相关实践

共有 12 篇文献[14,15,17,24,25,28,31,36,40,43,45,50]讨论了安全性及鲁棒性相关的保障实践或方法. 对于 AI 系统, 常见的且被现有研究讨论的攻击方法包括对抗 (逃避) 攻击、投毒攻击和模型窃取[15,17,25,43,45], 按照所应对的攻击种类的不同, 本文将所有的方法分类为投毒攻击防御、对抗攻击防御、模型窃取防御和与攻击无关的模型优化实践共 4 类, 如表 6 所示.

表 6 安全性及鲁棒性相关实践和方法

保障改进实践	保障改进方法	文献
投毒攻击防御	数据标准化[A34–A45]	[28,40,43,50]
	后门攻击检测[A46,A47]	[43,50]
对抗攻击防御	对抗样本检测[A48–A54]	[17,36,50]
	对抗训练[A55–A69]	[15,17,36,43,45,50]
	防御蒸馏[A57,A70–A72]	[15,17,28,45,50]
	数据处理[A73–A75]	[43,45,50]
	梯度正则化[A57,A59,A76–A80]	[15,45,50]
	防御网络[A50,A81,A82]	[25,45]
模型窃取防御	隐藏式安全[A75,A83–A92]	[43]
模型优化	模型可解释性	[31]
	AutoML	[31]

● 投毒攻击防御. 投毒攻击是基于一种利用受污染数据进行学习的思想发展起来的攻击方式[15]. 攻击者将受污染的数据或样本植入到训练集中以控制训练数据的分布, 从而破坏模型或使模型得出错误的结果[43]. 文献 [50] 指出投毒攻击主要有两种攻击方式: 一种是通过植入污染数据影响系统的边界, 从而破坏系统的可用性; 另一种是生成后门从而破坏系统的完整性. 相应地, 投毒攻击防御实践也包含 2 种不同类型的应对方法, 即数据标准化和后门攻击检测. 数据标准化的思想是将敌对样本从正常样本中分离出来, 然后去除这些恶意样本以保证训练数据的纯度[40]. 后门攻击检测的思想是检测攻击者植入的后门触发器. 然而, 由于只有当存在后门触发器时才会触发恶意行为, 且后门触发器在触发前只有攻击者知道, 因此后门攻击检测极具挑战性[43].

● 对抗攻击防御. 对抗攻击发生在 AI 系统的应用阶段, 其基本思想是构造促使 AI 模型产生错误分类的对抗样例[43]. 以一般的分类器为例, 投毒攻击改变的是模型的分类边界, 而对抗攻击则是将输入的样本修改为错误的类别[25]. 对抗攻击防御实践包括对抗样本检测、对抗训练、防御蒸馏等 7 种具体的方法. 对抗样本检测基于二分类思想, 利用检测器将系统的输入分为原始输入和对抗样本两类[36]. 对抗训练旨在提高系统的鲁棒性, 核心思想是将对抗样本加入原始训练数据集中进行训练, 使训练后的模型能够学习对抗样本的特征[36,43,45]. 防御蒸馏是神经网络应对对抗攻击的一种方法, 该方法使用蒸馏的方式降低 DNN 的计算复杂度[43,45]. 数据处理通过改变或转换输入的格式及特征以应对对抗攻击[43,45]. 梯度正则化 (或称梯度掩蔽) 方法通过修改输入数据的梯度、损失函数或激活函数来增强模型的鲁棒性[15]. 防御网络是使用神经网络等工具对对抗样本进行自动对抗的方法[45].

● 模型窃取防御. 模型窃取是一种查询目标模型并利用其功能的攻击方式[43]. 文献 [A57,A58] 提出了一种隐藏式安全方法, 即一种通过向攻击者隐藏信息以防御攻击者查询目标模型的方法.

● 模型优化. 文献 [31] 中指出提高模型的可解释性和采用 AutoML 也可以作为防御策略. 提高模型的可解释性有助于开发人员理解模型决策背后的原因, 从而识别系统的弱点和不足以进行防御. AutoML 旨在帮助模型选择最优的值和权重来进行训练, 从而提升模型防御各种攻击的能力.

5.2 可解释性相关实践

共有 10 篇文献[14,16,19,24,33−35,41,44,51]讨论了可解释性相关的保障实践或方法. 如表 7 所示, 本文将相关实践划分为 3 类, 包括预建模、建模时、建模后的可解释性保障实践.

表 7 可解释性相关实践和方法

保障改进实践	保障改进方法	文献
预建模可解释性保障	解释训练数据[A43,A93–A97]	[16]
建模时可解释性保障	创建白盒模型[A98–A117]	[16,19,33,35,41,44,51]
建模后可解释性保障	基于实例的方法[A118–A121]	[16,19,51]
	基于规则的方法[A122–A130]	[16,44,51]
	基于可视化的方法[A131–A141]	[14,16,19,34,51]
	基于代理的方法[A142–A146]	[14,19]
	基于传播的方法[A147–A153]	[19,34,44]

● 预建模可解释性保障. 文献 [16] 指出为 AI 系统提供可解释性的一种方法是解释系统的训练数据集[16], 其要求在开发模型前, 探索和理解训练系统所需的数据集从而为系统提供事前解释. 例如, 可以在使用数据前采用可视化技术辅助理解数据, 这类技术能够帮助系统的设计和开发人员更好地理解数据集中各种属性的分布, 从而更好地理解系统的模型. 通过预建模可解释性保障实践提供的可解释性也被称为事前可解释性 (ex-ante explainability)[16].

● 建模时可解释性保障. 在建模时, 可以专门创建可解释的或容易理解的模型, 即创建白盒模型的方法[41]. 一个透明的可解释的模型应该能够从整体上被完全理解, 同时也应使人能模拟出模型的计算过程[44]. 决策树、线性模型和基于规则的模型等模型族均能提供较好可解释性. 不过显然, 只有部分复杂度不高的模型能够满足可解释性的要求, 即建模时可解释性保障具有明显的局限性[16].

● 建模后可解释性保障. 这类实践保障的是系统的事后可解释性 (post-ante explainability), 即解释训练后的模型[44]. 现有研究讨论到的可解释性实践大部分属于这类实践, 有多种实现方法, 包括基于实例的方法, 基于规则的方法, 基于可视化的方法, 基于代理的方法以及基于传播的方法等. 基于实例的方法通过解释系统的具体实例, 即从输入数据集中选择部分实例并监控它们的输出以提供系统的可解释性[16,19]. 基于规则的方法通常应用于神经网络, 这类方法通过从已训练的模型中提取解释规则等信息来提供可解释性[16,44]. 基于可视化的方法通过将原本不透明的 AI 系统的内部工作进行可视化来提供可解释性[16,19]. 基于代理的方法通过创建与黑盒模型类似的简单模型以保障可解释性[19]. 基于传播的方法适用于基于神经网络构建的系统, 其基于反向传播和正向传播, 通过对扰动特征后的输出差异量化系统的特征, 从而达到提升可解释性的目的[19].

5.3 隐私性相关实践

有 16 篇文献[15,16,18,24,25,28−30,36,39,40,42,43,45−47]讨论了隐私性的保障实践或方法. 具体可以划分为差分隐私, 同态加密, 安全多方计算, 去识别技术, 联邦学习, 数据降维, 函数加密共 7 类实践, 如表 8 所示.

表 8 隐私性相关实践和方法

保障改进实践	保障改进方法	文献
差分隐私	输入扰动[A154–A161]	[29,43,46,47]
	参数扰动[A32,A162–A169]	[29,36,39,43,45−47]
	目标函数扰动[A170–A177]	[29,36,39,45−47]
	输出扰动[A174,A178–A183]	[29,39,45−47]
同态加密	训练数据加密[A184–A192]	[39,45]
	模型加密[A186,A193–A195]	[39,40,43]
安全多方计算	共20种未分类方法[A196–A215]	[15,18,30,39,46,47]
去识别技术	共6种未分类方法[A216–A221]	[16]
联邦学习	共4种未分类方法[A222–A225]	[16,18,39,42]
数据降维	Hamn[A226]	[15]
函数加密	Sans等人[A227], Marc等人[A228]	[15]

● 差分隐私是一类典型的基于扰动的方法, 其通过向数据中加入噪声以减少个人信息泄露的隐私风险[15,29]. 根据添加噪声的阶段不同, 差分隐私可以更具体地分为输入扰动方法, 参数扰动方法, 目标函数扰动方法和输出扰动方法 4 类. 输入扰动方法是对数据集进行预处理, 通过向原始数据添加噪声以避免模型接触到用户的真实数据[46]. 参数扰动是梯度级的扰动方法, 通过向参数中添加噪声避免隐私信息泄漏[29,36]. 目标函数扰动也被简称为函数扰动, 主要方法是向目标函数或目标函数的展开式系数中添加噪声[29]. 输出扰动是对训练后的模型参数添加噪声或模型预测后的输出结果添加噪声以避免模型受到提取攻击或推断攻击的方法[29,46].

● 同态加密是一种利用加密算法的实践, 加密的目标是允许对密文进行任意运算, 且解密后的结果与明文运算一致[15,40]. 具体依据加密主体的不同, 同态加密可以分为训练数据加密和模型加密两种. 文献 [39] 中指出, 在训练过程中加入同态加密会使得训练至少慢上一个数量级, 因此其往往用于训练简单分类器时. 模型加密则是对模型的梯度、参数等进行加密以保护隐私的方法, 例如 Gilad-Bachrach 等人[A186] 将实数编码到多项式中并进行非线性函数的低次多项式逼近, 以保护神经网络的隐私性.

● 安全多方计算与同态加密同属基于加密的隐私保护实践[39]. 安全多方计算是加密技术在多方场景下的扩展, 主要用于在不透露两个或多个参与方的个人信息的前提下, 计算得出各参与方的联合函数[15]. 文献 [39] 指出多方安全计算一般组合使用加密和遗忘传输, 在单个组件不可见的情况下由各方私下完成计算, 这可以保证计算模型的更新而不需要访问数据和模型[39]. 例如, Mohassel 等人[A196] 利用随机梯度下降方法和多方计算的秘密共享设置, 引入了 SecureML 以实现两方计算场景下的逻辑回归和神经网络的私人训练. Mehnaz 等人[A197] 提出了一个基于安全和计算的通用框架, 该框架使得多方能够以隐私保护的方式对分割数据进行训练. 更多多方安全计算的具体方法详细参见原文献[A198–A215].

● 去识别技术通过去除数据中存在的直接或间接的身份标识及其关联关系以保护个体隐私[16]. 例如, Garfinkel 等人[A216] 提出使用数据抽样和聚合实现基于样本子集或汇总版本来表示数据, 该方法能够避免发布完整数据集带来的风险, 进而实现对数据的去识别. Khail 等人[A217] 提出了一种称为抑制技术的屏蔽技术来屏蔽敏感属性的值, 达到去识别化的目的. 更多具体方法详细参见原文献[A218–A221].

● 联邦学习的核心思想是模型在用户自有的设备上进行训练后再进行共享, 通过这种协作学习的机制能够在一定程度上保护用户原始数据的隐私[16,43]. Geyer 等人[A222] 提出了一种在分布式学习场景下维护用户数据的联邦学习框架, 该框架可以保护用户级别的差分隐私. Hao 等人[A223] 提出了一种名为 PEFL (privacy enhanced federated learning) 的联邦学习方法, 以降低共享参数被对手利用的风险. Konečný 等人[A224] 提出了一种利用聚合进行协作学习的联邦学习框架, 以提高联邦学习场景下通信的效率.

● 数据降维通过将高维数据映射到低维以避免攻击者重建数据或推断数据中的敏感信息[15]. Hamn[A226] 提出了一种基于数据将为的防御方法, 该方法作用于模型的训练阶段, 可以在保留目标任务信息的同时, 删除数据中的敏感属性, 从而使得对手难以从输出中推断用户的隐私性数据.

● 函数加密与同态加密类似, 区别在于数据的加密方可以通过使用与计算结果关联的密钥进行解密, 这样只会泄漏计算的结果而不会泄漏原始数据的任何信息[15]. Sans 等人[A227] 展示了一种用于 MNIST 数字的分类器, 该分类器基于一个具有二级激活函数的隐层神经网络, 通过解密密钥允许服务器学习分类的结果. Marc 等人[A228] 提供了一个开源的函数式加密库, 实现了机器学习中包括内积、基于属性的机密等常用的函数加密原语.

5.4 公平性相关实践

有 5 篇文献[14,16,27,48,49]讨论了公平性相关的实践和方法. 对于公平性的质量保障实践, 本文根据作用于模型处理阶段的不同将这些实践映射为预处理机制, 处理中机制和处理后机制共 3 类, 如表 9 所示.

表 9 公平性相关实践

保障改进实践	保障改进方法	文献
预处理机制	共22种未分类方法[A26,A229–A249]	[14,16,27,48,49]
处理中机制	共30种未分类方法[A27,A250–A278]	[16,27,48,49]
处理后机制	共7种未分类方法[A28,A279–A284]	[16,27,48,49]

● 预处理机制是在模型训练之前对数据集进行预处理以确保数据不会存在偏见或歧视, 进而保障公平性[16]. 文献 [48] 将这类实践的任务称为公平表征任务, 这类任务的目标是使输入的特征仅保留与预测目标有关的信息, 而排除与预测目标无关的个体特征, 从而提取出公平特征以建立公平数据集. 例如, Xu 等人[A229] 提出的 FairGAN 模型能够通过两个特征网络, 在保障公平性的前提下生成特征, 以保证生成的特征与原始数据以及不同群体的属性分布均相似. Kamiran 等人[A230] 提出了一种通过改变一些实例的标签或权重从而使得分类器更加公平的方法. 更多具体方法详细参见原文献[A26,A231–A249].

● 处理中机制通过修改决策算法或模型以防止和减轻偏见对公平性的影响[16]. 文献 [48] 将这类实践的任务视为公平建模任务, 这类任务的目标是改进算法模型, 在保证准确性的同时保障模型的公平性. 例如, Berk 等人[A250] 提出了一种使用个体公平、群体公平和组合公平 3 个正则化项对损失函数进行加权以实现公平回归算法的方法, 该方法在提供个体公平和群体公平的同时, 能够计算出保障公平性对预测准确性的影响. Lepri 等人[A251] 建议由多学科专家团队对算法进行审查, 以保证算法不偏不倚. 更多具体方法详细参见原文献[A27,A252–A278].

● 处理后机制通过对系统的输出进行后置处理以消除决策中存在的偏见以保障决策的公平性[16]. 文献 [48] 将这类实践的任务称为公平决策任务, 这类任务的目标是对决策结果进行调整以确保决策对每个群体是公平的. 例如, Menon 等人[A279] 建议通过对不同群体的目标函数使用不同的阈值以减少决策中可能出现的偏见. 类似的, Dwork 等人[A280] 提出了一种对不同群体使用不同分类器的解耦技术用于减少偏差. 更多具体方法详细参见原文献[A28,A281–A284].

5.5 可问责性相关实践

有 2 篇文献[14,16]讨论了可问责性的相关保障实践/方法. 在进行方法分类时, 本文采用了文献 [16] 中对于可问责性保障实践的分类, 分为预建模 (ex-ante) 可问责性保障, 建模时 (in-ante) 可问责性保障, 建模后 (post-ante) 可问责性保障方法, 如表 10 所示.

表 10 可问责性相关实践

保障改进实践	保障改进方法	文献
预建模可问责性保障	共4种未分类方法[A285–A288]	[16]
建模时可问责性保障	共6种未分类方法[A289–A294]	[16]
建模后可问责性保障	共6种未分类方法[A294–A299]	[14,16]

● 预建模可问责性保障实践是在算法实际开发之前的规划和设计阶段, 为系统所做的决定进行责任分配, 这类方法还可以用于清晰地描述所有直接或间接受系统影响的用户[16]. 这类实践有很多更具体的方法, 例如, 通过优先级排序来解决利益冲突问题并实现更好地治理和问责[A285]. Broeders 等人[A286] 提出了一种时间框架机制 (time frame mechanism), 该框架通过在设计过程中及时重新评估系统, 以确保按照预期的规范和指南工作, 从而避免造成潜在的伤害. 还有研究人员[A287,A288] 指出列明所有的设计规范并清楚地描述系统在不同情况下的行为, 有利于更好的治理和问责.

● 建模时可问责性保障实践是在 AI 系统的开发阶段采取措施保障可问责性[16]. 这类实践管理 AI 系统整个开发阶段 (含测试、评估等), 在确保系统满足其他质量属性的同时保证系统能够向用户证明其所做决策的能力. 文献 [16] 指出具体的建模时实践包含 3 个步骤, 首先确保训练数据没有偏见[A289–A291], 然后根据问题选择合适的算法模型[A292,A293], 最后在系统部署之前进行测试[A294].

● 建模后可问责性保障实践在模型或系统部署之后解决可问责性问题[16]. 这类实践包括一些审计技术或框架. 例如, Kroll 等人[A294] 提出了一个框架用于确保部署的模型能够在指定的边界内工作. Raji 等人[A295] 提出了一种用于算法审计的框架, 该框架作为一个内部审计框架, 支持专家对开发过程的每个步骤进行审计, 有助于在问题发生前进行预防或发生后进行缓解. LeBrie 等人[A296] 提出一个可用于 AI 系统外部验证的道德审计框架, 用于帮助部署的 AI 系统避免偏差和错误. 更多具体方法详细参见原文献[A297–A299].

5.6 人类代理和监督相关实践

仅有 1 篇文献[14]讨论了人类代理和监督的相关保障实践/方法. 根据该文献, 本文将这些保障方法映射到人机协作机制下.

● 人机协作机制旨在通过人类的参与来提高人工智能系统的信任度和准确性[14]. 例如, Veeramachaneni 等人[A300] 引入了一个用于入侵检测的分析在环 (analysis-in-loop) 的 AI 系统, 该系统从安全分析师那里获取反馈, 以减少系统的误判. Kaur 等人[A301] 提出了一种人机协作机制用于控制警察和机器之间的交互, 以提高犯罪热点检测的准确性. 更多详细方法参见原文献[A302−A306].

RQ4 要点总结: 本文获取到了安全性, 鲁棒性, 可解释性, 隐私性, 公平性, 可问责性以及人类代理和监督共 7 个质量属性的质量保障改进实践. 其中, 现有二级研究中对安全性, 隐私性和可解释性的保障改进实践的总结更为系统、全面, 而可问责性与人类代理和监督的保障改进实践则相对单薄.

6 讨 论

本节将基于研究结果, 从研究现状、质量属性与实践两个方面进一步展开讨论, 并识别潜在的研究方向.

6.1 研究现状

● 现有二级研究覆盖较为全面, 但焦点具有非常强的倾向性. 图 5 描绘了可信 AI 的研究范畴元模型, 现有研究从高抽象级的可信到细粒度的实践、方法和工具, 均有所覆盖. 然而, 目前已有的二级研究在可信的关注点以及研究层次上分布并不均衡. 大部分研究关注于可信的威胁与挑战, 以及相应的保障改进实践, 尤其关注的是安全性和隐私性的相关问题. 安全性和隐私性均属于伤害预防原则中的质量属性. 相较于与伦理道德密切相关的尊重人类自治原则和公平性原则, 伤害预防原则的适用范围更广, 因为并不是每一种 AI 系统都会面临伦理道德的考验, 比如软件工程领域中被广泛研究的基于 AI 的缺陷预测系统目前并不涉及尊重人类自治原则和公平性原则中的问题, 但其会涉及伤害预防原则中的安全性、鲁棒性等问题. 表 4 对现有研究的研究问题进行了整理, 看起来每种研究对象都被覆盖了, 但具体到各个质量属性, 还有相当大的空白; 具体到不同粒度的研究问题, 关注于 How 层面的研究非常少, 这一定程度上是二级研究的抽象程度决定的, 但相当多的研究问题都是停留在 What 层面的, 更深入的二级研究是研究者们未来可以关注的方向.

● 可信的范畴太广, 术语的一致性问题是研究的重要阻碍. 由于通常会对系统的预测性能进行评估, 绝大部分 AI 系统的一级研究都会涉及鲁棒性的问题. 预测性能的评估方法也一直是一个被研究者们关注的问题. 但在本文纳入的 34 篇文献中, 仅有 2 篇文献讨论了鲁棒性. 一些关于预测性能评估方法的研究, 例如文献 [64], 是不包含鲁棒性等关键词的. 由于术语的问题, 真正足够全面地开展可信相关的二级研究以及三级研究是困难的. 另一个被研究较多的质量属性是可解释性, 自深度学习技术诞生以来, 可解释性就一直是一个使其备受争议的问题, 所以可解释性获得较高的关注度也是顺理成章的. 然而, 正如第 3.4 节中讨论的, 可解释性的术语也并不统一. 这些都会使研究者想要了解相关信息, 甚至是开展一项二级研究时, 难免有疏漏. 本文的一个重要价值, 正在于梳理与可信相关的研究范畴及其术语概念, 为可信相关的研究提供定位.

● 现有二级研究的质量参差不齐. 在数据抽取的过程中, 本文发现所获取到的二级研究所采用的研究方法大多没有遵循系统性文献综述 (SLR) 或系统性文献映射 (SMS) 等更加严谨的综述方法[13]. 在 Kitchenham 等人[23]将基于证据的软件工程 (evidence based software engineering, EBSE) 引入软件工程研究领域后, SLR 和 SMS 已经成为软件工程领域开展二级研究的主要研究方法. 在获取到的文献中, 仅有 2 篇文献[24,37]基本遵循了 SLR 或 SMS 的规范, 这可能会给二级研究的完整性和有效性带来一定的威胁, 例如, 对于文献检索环节上, 大部分研究没有遵循规范的文献检索策略, 这可能会导致关键文献的遗漏, 进而导致二级研究的研究结果不够完整. 从它们最终纳入分析的一级研究也可见一斑, 例如研究[A33,A46,A60,A61] 未发表到同行评审的会议或期刊上. 这些二级研究的质量, 也会一定程度上对本研究的结果产生一定的威胁.

● 多学科融合的二级研究太少. 人本主义是对 AI 可信提出要求的重要依据. 然而, 由于人本主义不属于计算

机相关学科, 所以仅有两篇文献[16,26]讨论了人本主义与 AI 系统之间的关系, 但文献 [26] 是一篇短文, 内容浅尝辄止, 仅讨论了人本主义和可解释性、公平性等质量属性之间的关系. 文献 [16] 中关于人本主义的讨论主要借鉴了《可信 AI 伦理指南》[10]. 相较于一般软件系统, AI 系统的可信中相当多的问题都源于人类社会对其提出的要求. 对于可信的评估、保障和改进, 最终都需要经受社会环境的验证. 多学科的融合是困难的, 但近年来各学科的融合已然是学术界的一大趋势, 对于 AI 系统可信的研究, 学科融合是必不可少的.

6.2 质量属性与实践

● 模型、AI 系统、一般软件系统之间的研究界限问题. 在第 3 节中可以看到, 对于同样的质量属性, 如安全性, 文献 [25] 系统角度讨论了安全性的 3 个子目标, 即保密性, 完整性和可用性. 文献 [26] 分别从系统角度和模型角度讨论了安全性. 而文献 [37] 仅从一般软件系统的角度讨论了安全性. 从这 3 项研究中可以看出从模型、AI 系统、一般软件系统等不同的角度研究安全性, 侧重是不同的, 例如对于一般软件系统, 如文献 [65] 在研究如何保障保密性, 完整性和可用性 3 个安全性子目标时, 往往从系统整体出发考虑授权, 认证等问题, 而对于 AI 模型而言, 研究的关注点常聚焦于投毒攻击, 对抗攻击等针对 AI 模型的攻击方法上[26]. 然而, 并不是关于所有质量属性的研究, 研究者们都做出了清晰的界定. 例如, 文献 [14] 和文献 [16] 均研究了可问责性, 前者是从模型角度定义可问责性, 后者是从系统角度进行定义, 但两者所描述的内容实质上是相同的. 可重复性、隐私性、公平性等多个质量属性都存在类似的问题. 本文认为, 模型作为系统的一部分, 系统的可信是包含模型的可信的. 在 ISO/IEC 25010:2011[63]中指出的适用于一般软件系统的质量属性, 如安全性, 在 AI 系统会因为 AI 模型而面临新的问题. AI 系统既包含一般软件系统的安全性问题, 也包含 AI 模型的安全性问题, 还可能包含模型和软件间的交互而产生的其他问题, 所以这三者并不等价. 而对于可重复性、隐私性、公平性等 AI 系统所特有的问题, 通常情况下, 研究中指代为模型或系统, 其含义是相同的.

● AI 系统与一般软件系统可信的联系与差异问题. 对于 5 种属于一般软件系统可信的质量属性兼容性, 可维护性, 功能适用性, 可移植性, 性能效率, 仅有一篇文献[33]进行了研究, 且采用了 ISO/IEC 25010:2011[63]中的定义. 由于缺乏其他的旁证, 且文献 [37] 仅研究了 AI 系统中保障上述质量属性开展的实践, 而未与一般软件系统的可信实践进行对比. 所以两者是否有差异, 存在哪些差异, 还有待进一步研究.

● 关于可信的评估度量的二级研究较少. 仅有 6 篇文献[15,16,19,38,45,49]研究了共 4 个质量属性的评估度量问题. 有效的评估各质量属性是评估可信的前提, 而对于可信的评估是获得用户及监管方信任的前提. 在未来的研究中, 这是一个值得探索的方向, 也是一个应该被填补的空白.

7 效度威胁

本三级研究的研究过程中存在的效度威胁主要来自文献收集、数据抽取与合成以及数据分析中环节中潜在的风险.

● 文献收集. 文献收集环节的主要风险来自可信 AI 质量属性相关文献的遗漏. 为了减轻这一风险, 一方面, 我们通过手动检索确定了检索字符串并明确文献的纳入/排除标准, 并基于确定的检索字符串在 4 个在线数字图书馆 (IEEE Xplore, ACM Digital Library, ScienceDirect, Scopus) 中进行了自动检索; 另一方面, 根据 Wohlin[57]的建议, 我们在谷歌学术上进行迭代式的滚雪球过程, 以检查是否有任何遗漏的研究. 此外, 为避免文献筛选过程中由于个人偏见引发的错误, 每一篇文献均由两位研究生独立进行筛选共两次, 然后由他们的导师的进行审核.

● 数据抽取与合成. 这一环节中的风险主要包括遗漏有价值的文本和数据抽取中的个人偏差两方面. 为缓解这一风险, 每一篇文献都由一位研究生进行抽取并由另一位研究生进行验证. 此外, 在数据抽取的过程中, 他们的导师定期提供指导, 并对疑问或有争议的部分进行解答. 在数据合成的过程中, 由两位研究生独立进行, 并由他们的导师审查并讨论冲突或不确定的内容.

● 数据分析. 在数据分析过程中, 人工偏差可能会对结论的有效性产生威胁. 为了减轻这种威胁, 两名研究生及他们的导师组织了几次头脑风暴会议, 讨论主题的分类方法以及结论的组织形式.

尽管在研究工作的各个环节, 我们均采取措施规避有效性风险, 但事实上仍有一些效度威胁未能得到解决. 例

如, 欧盟的《可信 AI 伦理指南》[10]的 7 项基本要求中有多样性 (diversity), 非歧视 (non-discrimination) 和公平 (fairness) 的要求, 但多样性并未在收集到的二级研究中有所讨论. 此外, 所获取二级研究本身也存在着一定的效度威胁, 这同样会对研究结果产生影响. 本文尽可能地基于现有研究提供 AI 可信的全貌, 随着相关领域的发展, 我们可能在未来需要阶段性地更新可信 AI 的质量框架. 研究者们可以针对本文所识别的研究现状中的空白, 开展未来的工作.

8 总 结

本文着眼于定义、评估和保障 AI 系统的可信性, 针对可信所需满足的质量属性及相关的评估度量方法和保障改进实践等方面, 开展了一项三级研究. 通过对 2022 年 3 月前的文献进行检索, 最终收集并综合分析了 34 项二级研究. 本文分析和讨论了相关二级研究的现状, 提出了一个可信人工智能研究范畴元模型供未来的研究者们对可信相关的研究进行定位. 此外, 本文分析和梳理了与可信相关的质量属性及相关实践和方法, 提出了一个可信人工智能系统的质量框架, 并综合现有研究中的描述, 对可信相关的 21 种质量属性提供了更准确的定义. 基于研究的结果, 本文对发现的问题和未来潜在的研究方向进行了讨论. 其中, 本文认为最关键的两方面问题, 都是学科融合的问题. 可信 AI 系统是一个跨学科的研究方向. 可信的要求源于社会科学, 可信 AI 模型的评估和保障需要基于计算机、人工智能、大数据等多学科的合作, 而系统层面的可信又属于软件工程问题. 现有二级研究中存在两处明显研究不足, 一方面 AI 系统作为一种特定软件系统, 与一般软件系统的差异与联系有待深入研究; 另一方面, 如何从系统层面保障机器在社会环境下的可信, 如怎样满足人本主义, 有待深入研究.

References:

[1] Anthes G. Artificial intelligence poised to ride a new wave. Communications of the ACM, 2017, 60(7): 19–21. [doi: 10.1145/3088342]

[2] McGraw G, Bonett R, Figueroa H, Shepardson V. Security engineering for machine learning. Computer, 2019, 52(8): 54–57. [doi: 10.1109/MC.2019.2909955]

[3] Angwin J, Larson J, Mattu S, Kirchner L. Machine bias: Risk assessment in criminal sentencing. ProPublica, 2016.

[4] Duan RJ, Mao XF, Qin AK, Chen YF, Ye SK, He Y, Yang Y. Adversarial laser beam: Effective physical-world attack to DNNs in a blink. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16057–16066. [doi: 10.1109/CVPR46437.2021.01580]

[5] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. Denver: ACM, 2015. 1322–1333. [doi: 10.1145/2810103.2813677]

[6] Shneiderman B. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans. on Interactive Intelligent Systems, 2020, 10(4): 26. [doi: 10.1145/3419764]

[7] Buruk B, Ekmekci PE, Arda B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. Medicine, Health Care and Philosophy, 2020, 23(3): 387–399. [doi: 10.1007/s11019-020-09948-1]

[8] Tolmeijer S, Kneer M, Sarasua C, Christen M, Bernstein A. Implementations in machine ethics: A survey. ACM Computing Surveys, 2021, 53(6): 132. [doi: 10.1145/3419633]

[9] ISO. ISO/IEC TR 24028: 2020 Information technology—Artificial intelligence—Overview of trustworthiness in artificial intelligence. Int'l Organization for Standardization, 2020.

[10] European Commission. Ethics guidelines for trustworthy AI. Publications Office of the European Union, 2019. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[11] European Commission. Assessment list for trustworthy artificial intelligence. Publications Office of the European Union, 2019. https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

[12] China Academy of Information and Communications Technology, JD Explore Academy. White Paper on Trustworthy Artificial Intelligence. 2021 (in Chinese). http://www.caict.ac.cn/kxyj/qwfb/bps/202107/t20210708_380126.htm

[13] Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Keele: Keele University, 2007. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

[14] Kaur D, Uslu S, Durresi A. Requirements for trustworthy artificial intelligence–a review. In: Proc. of the 23rd Int'l Conf. on Network-based Information Systems. Cham: Springer, 2021. 105–115. [doi: 10.1007/978-3-030-57811-4_11]

[15] Xiong PL, Buffett S, Iqbal S, Lamontagne P, Mamun M, Molyneaux H. Towards a robust and trustworthy machine learning system development: An engineering perspective. Journal of Information Security and Applications, 2022, 65: 103121. [doi: 10.1016/j.jisa.2022.103121]

[16] Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: A review. ACM Computing Surveys, 2023, 55(2): 39. [doi: 10.1145/3491209]

[17] Tariq MI, Memon NA, Ahmed S, Tayyaba S, Mushtaq MT, Mian NA, Imran M, Ashraf MW. A review of deep learning security and privacy defensive techniques. Mobile Information Systems, 2020, 2020: 6535834. [doi: 10.1155/2020/6535834]

[18] Boulemtafes A, Derhab A, Challal Y. A review of privacy-preserving techniques for deep learning. Neurocomputing, 2020, 384: 21–45. [doi: 10.1016/j.neucom.2019.11.041]

[19] Hanif A, Zhang XY, Wood S. A survey on explainable artificial intelligence techniques and challenges. In: Proc. of the 25th IEEE Int'l Enterprise Distributed Object Computing Workshop (EDOCW). Gold Coast: IEEE, 2021. 81–89. [doi: 10.1109/EDOCW52865.2021.00036]

[20] Kitchenham BA, Pretorius R, Budgen D, Brereton OP, Turner M, Niazi M, Linkman S. Systematic literature reviews in software engineering—A tertiary study. Information and Software Technology, 2010, 52(8): 792–805. [doi: 10.1016/j.infsof.2010.03.006]

[21] Kitchenham BA, Budgen D, Brereton P. Evidence-based Software Engineering and Systematic Reviews. New York: CRC Press, 2015. [doi: 10.1201/b19467]

[22] Dyba T, Kitchenham BA, Jorgensen M. Evidence-based software engineering for practitioners. IEEE Software, 2005, 22(1): 58–65. [doi: 10.1109/MS.2005.6]

[23] Kitchenham BA, Dyba T, Jorgensen M. Evidence-based software engineering. In: Proc. of the 26th Int'l Conf. on Software Engineering. Edinburgh: IEEE, 2004. 273–281. [doi: 10.1109/ICSE.2004.1317449]

[24] Serban A, van der Blom K, Hoos H, Visser J. Practices for engineering trustworthy machine learning applications. In: Proc. of the 1st IEEE/ACM Workshop on AI Engineering-software Engineering for AI (WAIN). Madrid: IEEE, 2021. 97–100. [doi: 10.1109/WAIN52551.2021.00021]

[25] Dilmaghani S, Brust MR, Danoy G, Cassagnes N, Pecero J, Bouvry P. Privacy and security of big data in AI systems: A research and standards perspective. In: Proc. of the 2019 IEEE Int'l Conf. on Big Data (Big Data). Los Angeles: IEEE, 2019. 5737–5743. [doi: 10.1109/BigData47090.2019.9006283]

[26] Fagbola TM, Thakur SC. Towards the development of artificial intelligence-based systems: Human-centered functional requirements and open problems. In: Proc. of the 2019 Int'l Conf. on Intelligent Informatics and Biomedical Sciences (ICIIBMS). Shanghai: IEEE, 2019. 200–204. [doi: 10.1109/ICIIBMS46890.2019.8991505]

[27] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys, 2022, 54(6): 115. [doi: 10.1145/3457607]

[28] Meenakshi K, Maragatham G. A review on security attacks and protective strategies of machine learning. In: Hemanth DJ, Kumar VDA, Malathi S, Castillo O, Patrut B, eds. Emerging Trends in Computing and Expert Technology. Cham: Springer, 2020. 1076–1087. [doi: 10.1007/978-3-030-32150-5_109]

[29] Zhang Y, Cai Y, Zhang M, Li X, Fan YF. A survey on privacy-preserving deep learning with differential privacy. In: Proc. of the 3rd Int'l Conf. on Big Data and Security. Shenzhen: Springer, 2022. 18–30. [doi: 10.1007/978-981-19-0852-1_2]

[30] Tiwari K, Shukla S, George JP. A systematic review of challenges and techniques of privacy-preserving machine learning. In: Shukla S, Unal A, Kureethara JV, Mishra DK, Han DS, eds. Data Science and Security. Singapore: Springer, 2021. 19–41. [doi: 10.1007/978-981-16-4486-3_3]

[31] Madhusudhanan S, Nair RR. Converging security threats and attacks insinuation in multidisciplinary machine learning applications: A survey. In: Proc. of the 2019 Int'l Seminar on Research of Information Technology and Intelligent Systems (ISRITI). Yogyakarta: IEEE, 2019. 217–222. [doi: 10.1109/ISRITI48646.2019.9034665]

[32] Chai CL, Wang JY, Luo YY, Niu ZP, Li GL. Data management for machine learning: A survey. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(5): 4646–4667. [doi: 10.1109/TKDE.2022.3148237]

[33] Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In: Proc. of the 41st Int'l Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Opatija: IEEE, 2018. 210–215. [doi: 10.23919/MIPRO.2018.8400040]

[34] Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, Srivastava M, Preece A, Julier S, Rao RM, Kelley TD, Braines D, Sensoy M, Willis CJ, Gurram P. Interpretability of deep learning models: A survey of results. In: Proc. of the 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud &

Big Data Computing, Internet of People and Smart City Innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI). San Francisco: IEEE, 2017. 1–6. [doi: 10.1109/UIC-ATC.2017.8397411]

[35] Vollert S, Atzmueller M, Theissler A. Interpretable machine learning: A brief survey from the predictive maintenance perspective. In: Proc. of the 26th IEEE Int'l Conf. on Emerging Technologies and Factory Automation (ETFA). Vasteras: IEEE, 2021. 1–8. [doi: 10.1109/ETFA45728.2021.9613467]

[36] Ha T, Dang TK, Le H, Truong TA. Security and privacy issues in deep learning: A brief review. SN Computer Science, 2020, 1(5): 253. [doi: 10.1007/s42979-020-00254-4]

[37] Gezici B, Tarhan AK. Systematic literature review on software quality for AI-based software. Empirical Software Engineering, 2022, 27(3): 66. [doi: 10.1007/s10664-021-10105-2]

[38] França HL, Teixeira C, Laranjeiro N. Techniques for evaluating the robustness of deep learning systems: A preliminary review. In: Proc. of the 10th Latin-American Symp. on Dependable Computing (LADC). Florianópolis: IEEE, 2021. 1–5. [doi: 10.1109/LADC53747.2021.9672592]

[39] Liu B, Ding M, Shaham S, Rahayu W, Farokhi F, Lin ZH. When machine learning meets privacy: A survey and outlook. ACM Computing Surveys, 2022, 54(2): 31. [doi: 10.1145/3436755]

[40] Guan ZY, Bian LX, Shang T, Liu JW. When machine learning meets security issues: A survey. In: Proc. of the 2018 IEEE Int'l Conf. on Intelligence and Safety for Robotics (ISR). Shenyang: IEEE, 2018. 158–165. [doi: 10.1109/IISR.2018.8535799]

[41] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy, 2021, 23(1): 18. [doi: 10.3390/e23010018]

[42] Liu JX, Meng XF. Survey on privacy-preserving machine learning. Journal of Computer Research and Development, 2020, 57(2): 346–362 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20190455]

[43] Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 41–67 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[44] Ji SL, Li JF, Du TY, Li B. Survey on techniques, applications and security of machine learning interpretability. Journal of Computer Research and Development, 2019, 56(10): 2071–2096 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20190540]

[45] He YZ, Hu XB, He JW, Meng GZ, Chen K. Privacy and security issues in machine learning systems: A survey. Journal of Computer Research and Development, 2019, 56(10): 2049–2070 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20190437]

[46] Tan ZW, Zhang LF. Survey on privacy preserving techniques for machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 2127–2156 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[47] Liu RX, Chen H, Guo RY, Zhao D, Liang WJ, Li CP. Survey on privacy attacks and defenses in machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 866–892 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[48] Liu WY, Shen CY, Wang XF, Jin B, Lu XJ, Wang XL, Zha HY, He JF. Survey on fairness in trustworthy machine learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(5): 1404–1426 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6214.htm [doi: 10.13328/j.cnki.jos.006214]

[49] Pessach D, Shmueli E. A review on fairness in machine learning. ACM Computing Surveys, 2023, 55(3): 51. [doi: 10.1145/3494672]

[50] Hu YP, Kuang WX, Qin Z, Li KL, Zhang JL, Gao YS, Li WJ, Li KQ. Artificial intelligence security: Threats and countermeasures. ACM Computing Surveys, 2023, 55(1): 20. [doi: 10.1145/3487890]

[51] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Computing Surveys, 2019, 51(5): 93. [doi: 10.1145/3236009]

[52] Zhang H, Babar MA, Tell P. Identifying relevant studies in software engineering. Information and Software Technology, 2011, 53(6): 625–637. [doi: 10.1016/j.infsof.2010.12.010]

[53] ISO. ISO/IEC TR 24027:2021 Information technology—Artificial intelligence (AI)—Bias in AI systems and AI aided decision making. Int'l Organization for Standardization, 2021.

[54] ISO. ISO/IEC TR 24029-1:2021 Artificial intelligence (AI)—Assessment of the robustness of neural network—Part 1: Overview. Int'l Organization for Standardization, 2021.

[55] IEEE. IEEE 3652.1–2020 IEEE guide for architectural framework and application of federated machine learning. IEEE, 2021. [doi: 10.1109/IEEESTD.2021.9382202]

[56] IEEE. IEEE P7001/D4-2021 IEEE approved draft standard for transparency of autonomous systems. IEEE. 2021. https://ieeexplore.ieee.

org/document/9574622

[57] Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proc. of the 18th Int'l Conf. on Evaluation and Assessment in Software Engineering. London: ACM, 2014. 38. [doi: 10.1145/2601248.2601268]

[58] Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology, 2006, 3(2): 77–101. [doi: 10.1191/1478088706qp063oa]

[59] Charmaz K. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Thousand Oaks: Sage, 2006.

[60] Jia CJ, Cai Y, Yu YT, Tse TH. 5W+1H pattern: A perspective of systematic mapping studies and a case study on cloud software testing. Journal of Systems and Software, 2016, 116: 206–219. [doi: 10.1016/j.jss.2015.01.058]

[61] Hart G. The five W's: An old tool for the new task of audience analysis. Technical Communication, 1996, 43(2): 139–145.

[62] Pan ZD, Kosicki GM. Framing analysis: An approach to news discourse. Political Communication, 1993, 10(1): 55–75. [doi: 10.1080/10584609.1993.9962963]

[63] ISO. ISO/IEC 25010:2011 Systems and software engineering —Systems and software quality requirements and evaluation (SQuaRE)—System and software quality models. Int'l Organization for Standardization, 2011.

[64] Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K. An empirical comparison of model validation techniques for defect prediction models. IEEE Trans. on Software Engineering, 2017, 43(1): 1–18. [doi: 10.1109/TSE.2016.2584050]

[65] Wu SH, Zhang C, Wang FT. Extracting software security concerns of problem frames based on a mapping study. In: Proc. of the 24th Asia-Pacific Software Engineering Conf. Workshops. Nanjing: IEEE, 2017. 121–125. [doi: 10.1109/APSECW.2017.29]

附中文参考文献:

[12] 中国信息通信研究院, 京东探索研究院. 可信人工智能白皮书. 2021. http://www.caict.ac.cn/kxyj/qwfb/bps/202107/t20210708_380126.htm

[42] 刘俊旭, 孟小峰. 机器学习的隐私保护研究综述. 计算机研究与发展, 2020, 57(2): 346–362. [doi: 10.7544/issn1000-1239.2020.20190455]

[43] 纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41–67. http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[44] 纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071–2096. [doi: 10.7544/issn1000-1239.2019.20190540]

[45] 何英哲, 胡兴波, 何锦雯, 孟国柱, 陈恺. 机器学习系统的隐私和安全问题综述. 计算机研究与发展, 2019, 56(10): 2049–2070. [doi: 10.7544/issn1000-1239.2019.20190437]

[46] 谭作文, 张连福. 机器学习隐私保护研究综述. 软件学报, 2020, 31(7): 2127–2156. http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[47] 刘睿瑄, 陈红, 郭若杨, 赵丹, 梁文娟, 李翠平. 机器学习中的隐私攻击与防御. 软件学报, 2020, 31(3): 866–892. http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[48] 刘文炎, 沈楚云, 王祥丰, 金博, 卢兴见, 王晓玲, 查宏远, 何积丰. 可信机器学习的公平性综述. 软件学报, 2021, 32(5): 1404–1426. http://www.jos.org.cn/1000-9825/6214.htm [doi: 10.13328/j.cnki.jos.006214]

李功源(1999－), 男, 硕士生, CCF 学生会员, 主要研究领域为人工智能的软件工程.

杨雨豪(1999－), 男, 硕士生, 主要研究领域为人工智能的软件工程.

刘博涵(1991－), 男, 博士, 助理研究员, CCF 专业会员, 主要研究领域为软件过程, 过程仿真建模, 机器学习, 软件资源库, 经验软件工程.

邵栋(1976－), 男, 副教授, CCF 专业会员, 主要研究领域为软件研发效能, 软件过程, 敏捷软件开发, DevOps, 高科技市场理论, 软件工程教育, 区块链, 大数据.