软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(8):3774−3792 [doi: 10.13328/j.cnki.jos.006592]

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

人工智能系统可信性度量评估研究综述*

刘 晗[1,2], 李凯旋[1,2], 陈仪香[1,2]

[1](华东师范大学 软件工程学院, 上海 200062)

[2](上海市高可信计算重点实验室 (华东师范大学), 上海 200062)

通信作者: 陈仪香, E-mail: yxchen@sei.ecnu.edu.cn

摘 要: 近年来, 人工智能技术突飞猛进, 人工智能系统已经渗透到人们生活中, 成为人们生活中不可或缺的一部分. 然而, 人工智能系统需要数据训练模型, 数据扰动会对其结果造成影响. 并且随着人工智能系统业务多样化, 规模复杂化, 人工智能系统的可信性愈发受到人们的关注. 首先, 在梳理不同组织和学者提出的人工智能系统可信属性基础上, 提出人工智能系统的 9 个可信属性; 接着, 从数据可信性、模型可信性和结果可信性分别介绍现有的人工智能系统数据、模型、结果可信性度量方法, 设计人工智能系统可信证据收集方法. 其次, 总结当前人工智能系统的可信度量评估理论与方法. 然后, 结合基于属性的软件可信评估方法与区块链技术, 建立一个人工智能系统可信度量评估框架, 包括可信属性分解及可信证据获取方法、联邦式可信度量模型与以及基于区块链的人工智能系统可信度量评估架构. 最后, 讨论人工智能系统可信度量技术面临的机遇和挑战.

关键词: 人工智能系统; 可信性; 度量评估

中图法分类号: TP18

中文引用格式: 刘晗, 李凯旋, 陈仪香. 人工智能系统可信性度量评估研究综述. 软件学报, 2023, 34(8): 3774–3792. http://www.jos.org.cn/1000-9825/6592.htm

英文引用格式: Liu H, Li KX, Chen YX. Survey on Trustworthiness Measurement for Artificial Intelligence Systems. Ruan Jian Xue Bao/Journal of Software, 2023, 34(8): 3774–3792 (in Chinese). http://www.jos.org.cn/1000-9825/6592.htm

Survey on Trustworthiness Measurement for Artificial Intelligence Systems

LIU Han[1,2], LI Kai-Xuan[1,2], CHEN Yi-Xiang[1,2]

[1](Software Engineering Institute, East China Normal University, Shanghai 200062, China)

[2](Shanghai Key Laboratory of Trustworthy Computing (East China Normal University), Shanghai 200062, China)

Abstract: In recent years, artificial intelligence (AI) has rapidly developed. AI systems have penetrated people's lives and become an indispensable part. However, these systems require a large amount of data to train models, and data disturbances will affect their results. Furthermore, as the business becomes diversified, and the scale gets complex, the trustworthiness of AI systems has attracted wide attention. Firstly, based on the trustworthiness attributes proposed by different organizations and scholars, this study introduces nine trustworthiness attributes of AI systems. Next, in terms of the data, model, and result trustworthiness, the study discusses methods for measuring the data, model, and result trustworthiness of existing AI systems and designs an evidence collection method of AI trustworthiness. Then, it summarizes the trustworthiness measurement theory and methods of AI systems. In addition, combined with attribute-based software trustworthiness measurement methods and blockchain technologies, the study establishes a trustworthiness measurement framework for AI systems, which includes methods of trustworthiness attribute decomposition and evidence acquisition, the federation trustworthiness measurement model, and the blockchain-based trustworthiness measurement structure of AI systems. Finally, it describes the opportunities and challenges of trustworthiness measurement technologies for AI systems.

Key words: artificial intelligence system; trustworthiness; measurement

---

* 基金项目: 华东师范大学-华为可信创新实验室项目 (15902-412312-19214/013); 上海市可信工业互联网软件协同创新中心项目
本文由"智能系统的分析和验证"专题特约编辑明仲教授、张立军教授和秦胜潮教授推荐.
收稿时间: 2021-09-03; 修改时间: 2021-10-14; 采用时间: 2022-01-10; jos 在线出版时间: 2022-01-28
CNKI 网络首发时间: 2023-01-19

人工智能自 1956 年诞生以来, 经历了从繁荣到衰退再到繁荣的螺旋式发展过程, 出现了 3 次发展高潮[1]. 第 3 次高潮起源于辛顿 (Hinton) 在 2006 年提出的深度学习概念[2], 2016 年围棋机器人 AlphaGo[3]以该模型为核心的算法战胜了人类顶级棋手引发广泛关注, 极大地推动了人工智能的研究与应用高潮. 近年来, 欧美日等国持续加大对人工智能基础理论和应用的重点投入, 以保持其在人工智能技术方面的领先地位. 我国政府在 2017 年发布了《新一代人工智能发展规划》, 将人工智能正式列入国家发展战略[4], 学术界和产业界也掀起了人工智能研发热潮, 李国杰院士和陆汝钤院士等多位人工智能科学家在多个场合呼吁重视人工智能的发展态势[5], 华为、百度、腾讯、科大讯飞等人工智能领域企业也在不断增强其人工智能方面的研发力量, 人工智能在图像识别、数据挖掘、自然语言处理、推荐算法、信息检索、语言识别和自动驾驶等领域均得到了不同程度的应用.

然而, 随着人工智能不断融入我们的日常生活, 人工智能系统的表现还不尽如人意. 例如, 近年来, 许多学者发现, 在图像识别领域, 人工智能模型对训练数据非常敏感, 当训练数据受到扰动时可能输出不恰当的结果, 这种对抗性扰动已经成为人工智能模型, 尤其是神经网络模型的梦魇: 对抗攻击[6−8]. 而且, 在自动驾驶领域也有研究发现, 除了人为的对抗攻击, 极端光照条件也会影响视觉系统识别, 从而影响自动驾驶[9]. 另一方面, 如果训练数据被偏见性地标注, 相应机器学习模型的输出结果往往也会具有一定偏见性[10]. 复旦大学管理学院企业管理系孙金云教授研究团队 2021 年发表的“2020 打车软件出行状态调研报告”表明, 打车软件通过“差异化的偏差信息”提高了平台自身的收益[11]. 这些数据一旦被滥用于机器学习算法, 不仅可能导致偏见性的结果, 还可能导致隐私泄露问题的出现. 这类事件频繁出现, 越来越多的专家和学者开始关注人工智能系统的可信性[1,5,12]. 系统的“可信性”是基于传统的“安全”“可靠”等概念产生的, 简单来说是指一个系统在实现给定目标时, 其行为及其结果符合人们的期望[13]. 刘克等人认为软件系统“可信性”是人类心中对客观对象各属性较为整体的反映[14].

在政府机构和有关学者的引导下, 学术界和产业界都愈发关注人工智能系统可信性[15], 构造可信人工智能已经成为现代人工智能发展和应用的重要趋势和必然选择. 然而, 构造可信人工智能需要人们对人工智能系统的可信性有着清晰的认知, 因而如何综合评估判断人工智能系统的可信性已经成为可信人工智能研究中的一个重要问题. 因为人工智能系统可信性由其多维属性反映, 并且人工智能系统的可信性问题需要从其训练数据可信性、学习模型可信性和预测结果可信性 3 个方面来考虑, 所以对人工智能系统可信性的综合评估较为困难. 本文从人工智能的可信属性入手, 总结梳理人工智能应有的可信属性; 接着讨论了数据、模型和结果可信性的度量方法, 为度量人工智能系统可信性提供证据基础. 在此基础上, 本文讨论了现有的人工智能可信度量模型. 然后, 本文提出了一个基于可信属性的人工智能系统可信度量框架, 期待推动人工智能可信度量的研究与发展. 本文将从以下几个方面讨论人工智能系统可信评估的研究现状和面临的挑战.

(1) 人工智能系统的可信属性

可信性是人类心中对客观对象各属性较为整体的反映, 传统软件的可信属性通常包含可靠性、安全性、可用性、正确性等诸多属性[14], 人工智能系统作为软件系统的一种, 其可信属性包括部分传统软件的可信属性. 但是, 人工智能系统本身的特性又导致其具有普通软件所不具有的其他属性, 例如描述对抗数据干扰能力的鲁棒性, 没有偏见的公平性等. 因此, 如何准确描述这些属性对人工智能系统的可信评估有着重要作用.

(2) 人工智能系统数据、模型和结果可信性与人工智能系统度量模型

人工智能系统的可信性依赖于大量可靠的数据支撑、恰当的模型选择和符合预期输出结果, 即其可信性依赖于其训练数据可信性、学习模型可信性和预测结果可信性. 数据可信性是训练出可信模型的基础, 模型可信性又是得到符合人们心理预期结果的基础, 三者相辅相成, 相互依赖, 共同影响着人工智能系统的可信性. 为了实现对人工智能系统可信性的度量, 需要从数据、模型和结果 3 方面的度量结果收集可信证据. 此外, 现有对人工智能系统度量的工作也为人工智能系统可信性度量模型建立打下了坚实的基础.

(3) 基于可信属性的人工智能系统可信度量评估体系

传统软件可信评估理论经过数十年的发展, 已经被证明在传统软硬件系统开发中是确保系统可信必不可少的技术. 例如列车控制系统和航空飞行系统开发国际标准中均要求将该方法贯穿于整个系统的开发周期以获得较高的可信等级认证[16]. 尤其是基于可信属性的软件可信度量评估, 现在已经有了较为成熟的研究成果, 并被用于各

种安全攸关软件的可信性度量中[17]. 然而, 由于目前人工智能系统本身特有的性质, 使得不能直接将软件可信评估理论简单地移植到人工智能系统, 而是要针对人工智能系统本身的可信性进一步研究可信度量评估理论与技术. 本文结合人工智能系统可信性现状和软件可信评估理论提出了一个人工智能系统可信度量评估框架, 具体包括可信属性分解及可信证据获取方法、联邦式可信度量模型与基于区块链的人工智能系统可信度量评估架构. 该框架还将保障数据可信性常用的区块链技术融入可信度量全过程, 从而可以保障整个度量过程的不可篡改性.

1 研究现状和动机

1.1 人工智能现状

1950 年, 计算机科学之父图灵在论文《计算机器与智能》中描述了智能的概念, 并提出机器智能的测试方法, 即“图灵测试”[18]. 随后在 1956 年的达特茅斯会议上, 美国 10 位学者正式提出了人工智能概念, 达特茅斯会议也成为第 1 次人工智能研讨会, 1956 年被称为“人工智能元年”[19]. 人工智能的发展自其诞生以来经历了 3 次高潮. 2006 年辛顿提出的深度学习算法的概念[2], 不仅掀起了深度学习研究与应用的热潮, 而且促进了人工智能第 3 次高潮的兴起. 人工智能发展的潮起潮落引领着人工智能不断健康地发展, 也使得人工智能技术健康地服务人类. 现在, 人工智能技术在各个领域已经取得了丰硕的研究成果, 图像识别[20]、自然语言处理[21]、计算机视觉[22]、自动驾驶[23]、推荐系统[24]等技术中处处可见其身影.

1.2 可信人工智能发展战略规划

如何保障人工智能的健康发展已经成为当前社会和国家关注的主要问题之一, 许多政府和学者都倡导科学发展人工智能. 2016 年 9 月, 英国下议院提出应对机器人发展带来的伦理、法律和道德问题[25]. 2018 年 3 月, 欧洲政治战略中心提出要解决人工智能在发展过程中出现的偏见问题, 研究人工智能的道德准则[26]. 2018 年 4 月, 欧盟确立人工智能的伦理和法律框架[27]. 2018 年 12 月, 欧盟委员会的人工智能高级专家组 (high-level expert group on artificial intelligence, AIHLEG) 发布了《可信人工智能伦理指南草案》, 该指南提出了一个可信人工智能框架, 总计 10 项可信人工智能要求和 12 项实现人工智能的技术和非技术方法[28].

2016 年 10 月, 美国国家科学技术委员会 (NSTC) 探讨了人工智能潜在的公共问题[29,30]. 2017 年年初, 美国国家科学基金会、国防部高级研究项目局 (DARPA) 启动可解释人工智能计划, 发展可解释、可信的人工智能技术[31]. 2018 年 4 月, 美国国防部发表《国防部人工智能战略》, 旨在促进美国人工智能安全[32]. 2018 年 9 月 DARPA 启动了 20 亿美元的 AINext 战略, 明确发展第 3 代人工智能基础理论和技术, 重点内容包括鲁棒、对抗、高效知识推理, 以及更强能力的人工智能理论和技术[33].

2018 年 6 月, 新加坡成立人工智能伦理委员会, 以帮助政府制定伦理标准[34]. 同一时间, 印度政府发布《人工智能国家战略》, 其中重点涉及了道德隐私方面的问题[35]. 我国政府从 2015 年开始就大力推进人工智能技术发展[36], 2016 年发展人工智能技术被列入“十三五”发展规划[37]. 2017 年国务院和工信部提出新一代人工智能发展规划[4,38], 强调人工智能健康发展, 2018 年国家标准化管理委员会发布的人工智能标准化白皮书中提到关注人工智能的安全、伦理和隐私问题[39].

1.3 软件可信性研究现状

1.3.1 软件可信性概念

随着科技的不断发展, 计算机在人类社会的普及程度越来越高, 无论是经济、军事还是社会生活中的方方面面, 人们都越来越离不开软件, 对软件的功能需求也在不断增加. 随着软件系统日益复杂化, 开发规模不断扩大, 或多或少都存在的软件缺陷使得人们越来越难以控制软件质量. 各种各样的软件事故不断给用户带来严重的损失, 软件经常不按人们期望的方式工作[17,40], 使得人们逐渐失去对软件的信任, 可信性的概念应运而生[14].

“可信性”是在软件传统的“可靠”“安全”等概念的基础上发展而来的. 1972 年, Anderson 提出可信系统的概念[41], 这是人们对计算机系统可信性的首次探索. 但是此时人们大多关注硬件的可信性, 直到美国国防部制定可信计算机标准时才提到软件可信性的概念[42]. Laprie 指出可信性与可靠性是两个不同的概念, 前者比后者要复杂得多[43].

美国科学与技术委员会 NSTC 则认为即使在系统在危险情况下, 即系统本身存在错误, 环境存在危险或者系统遭到其他人的致命攻击, 设计者、实现者和使用者都能保障系统的大部分功能, 使其不会失效, 则该系统是高可信的[44]. 美国国家研究委员会 NRC 认为一个系统即使在运行环境出现问题、操作人员失误操作、系统受到外界的致命攻击或者系统的设计和实现存在 Bug 的情况下, 也能够按照原来设定的预期运行, 得到预期的结果, 那么该系统是可信的[45]. 德国奥尔登堡研究生院的研究发现可信性应该包括正确性、安全性、服务质量、保密性和隐私性[46]. 在我国国家自然科学基金委“可信软件基础研究”重大研究计划中, 何积丰等人[14]认为可信性是人类心中对客观对象各属性较为整体的反映, 提出了可信软件 (trustworthy software) 是指软件系统的动态行为及其结果总是符合人们预期, 并在受到干扰时不会失效, 并能提供连续服务的软件, 这里的“干扰”包括操作错误、环境影响和外部攻击等. 陈火旺院士则认为高可信性质包括可靠性、防危性、安全性、可生存性、容错性和实时性[47].

1.3.2 软件可信性发展计划

软件可信性保证对整个软件产业, 特别对安全攸关软件研发的影响日益加深. 各国政府、研究机构以及各大公司都对软件可信性研究提出了相应的研究计划. 美国自然科学基金会从 2005 年开始便在可信计算研究领域投入约数亿美元[48], 政府的“网络与信息技术研究发展计划”中重点强调“可信软件”相关领域[49]; 欧洲于 2006 年 1 月启动了名为“开放式可信计算 (open trusted computing)”的研究计划, 旨在开发开源可信计算软件, 已有 23 个研究机构参加[50]; 欧盟于 1997 年 12 月通过的“第五框架计划”[51]和 2002 年 11 月通过的“第六框架计划”[52]都把高可信软件作为软件技术发展的重点. 构造可信软件已成为现代软件技术发展和应用的重要趋势和必然选择. 我国国家中长期科学和技术发展规划纲要 (2006–2020 年) 中将可信计算机的研究作为发展重点[53]; 2007 年, “863”计划开展“高可信软件生产工具及集成环境”重点项目[54]; 国家自然科学基金委于 2007 年提出开展“可信软件基础研究”重大研究计划[14], 并将“软件的可信性度量与评估”列为 4 个重要核心问题之首.

1.3.3 软件可信性度量评估

软件可信度量评估有助于软件可信性保障, 因而其研究成果具有重大意义. 国内外一系列学者专家专注于此, 并取得了丰硕的研究成果. Marascas 等人使用问卷调查的方法, 结合多元统计分析方法, 将人的主观评价和软件的客观度量结果相结合, 形成综合的度量结果[55−57]. 美国国家标准与技术研究院 NIST 提出一种自上而下的软件可信性评估框架, 使用形式化方法来确定软件可信值[58]. Alexopoulos 等人将贝叶斯概率和 DS 证据理论结合, 使用模块化的思想, 量化相关风险并对软件组件可信性进行度量[59]. Cho 等人提出了一个系统级别的可信度度量框架, 从可信属性的角度度量软件可信性, 该框架包含安全性、信任、弹性和敏捷性度量标准, 称为 STRAM 框架[60].

国内的许多科研团队同样在软件可信度量领域取得了丰硕的研究成果. 杨善林院士团队采用专家打分的效用值结合 DS 证据理论来计算软件可信性[61,62]. 郑志明和李未院士团队使用动力学分析, 研究软件在动态开放环境下的行为统计学特征, 建立软件关于可信属性的缩小化最优统计分析方法[63−65]. 王怀民院士团队对可信软件的概念做了进一步规范, 也从可信属性的角度建立软件可信属性模型和软件可信等级, 给出一种基于验证的可信证据模型[66]. 王德鑫等人在王青教授团队建立了支持软件过程可信评估的可信证据, 从开发过程数据来评估软件可信性[67]. 陈仪香教授团队同样面向多维属性, 基于公理化的方法建立面向软件多维属性的软件可信度量模型, 并提出软件可信性分配模型与增强规范[17,68,69]. 他们还从软件源代码角度, 建立基于 Extensive 结构的软件可信性度量模型[70].

1.4 人工智能系统可信度量动机研究

在人工智能技术发展初期, 其应用主要是专家系统、定理证明、问题求解等领域[71], 这些领域任务相对较为简单, 初期人工智能技术足以应付这些问题. 随着人工智能技术发展, 人们对人工智能系统和智能软件的需求不仅仅停留在这些简单问题上, 图像识别、语音识别、无人驾驶、推荐系统等越来越多的领域开始应用人工智能技术, 由此人工智能走向了第 3 次发展高潮.

然而, 和软件发展一样, 人工智能系统愈发复杂化的同时, 其可信性越来越受到人们的关注. 2015 年 8 月, 印

度一名配件公司员工由于离焊接金属板机器人太近而被机器人杀死[72]. 2016 年 8 月, 浙江大学徐文渊教授带领的团队在 Defon 黑客大会上对特斯拉自动驾驶技术进行攻击, 导致其自动驾驶失效[73]. 2018 年 3 月 Uber 自动驾驶汽车在行驶过程中没有识别路上行人, 撞倒了该行人并导致其死亡[74]. 人工智能系统不仅面临着传统软件所存在的可信问题, 而且存在人工智能特有的问题. 文献 [75] 发现医疗诊断工具对黑人患者分配较少的医疗资源, 并且对许多白人患者, 该医疗诊断工具拒绝使用黑人患者已使用过的治疗设备. 同样, 2018 年 5 月, 亚马逊公司智能音箱出现故障, 错误地把俄勒冈州一名女士和其丈夫的对话发到了他人邮箱[76]. 这类事件使得人们对人工智能系统的可信性愈发关注, 各个国家也在人工智能发展战略中提到发展可信人工智能.

人工智能系统可信度量可以帮助人工智能系统的开发者和使用者了解人工智能的可信程度, 使他们对人工智能系统的可信程度有一个明确的认知. 开发者可以通过可信度量得出的结果进一步改进开发的系统, 使得该系统可以保证用户安全的同时, 让用户有更好的智能体验. 用户也可以根据可信度量的结果对自己使用的智能系统的可信程度有所了解, 在使用过程中可以放心享受智能带来的快乐和方便. 因而, 对人工智能系统进行可信度量是人工智能系统健康发展的必然选择.

2 人工智能可信属性

目前, 越来越多的组织和学者关注人工智能系统的可信性, 他们意识到, 人工智能系统可信性不同于传统软件的可信性, 它的可信属性除了包含传统软件所需要的可信属性外, 还应包含许多人工智能本身特性而产生的相关性质. 例如, 人工智能模型的黑盒特性需要关注人工智能的可解释性与透明性.

欧盟委员会人工智能高级专家组提出可信人工智能伦理指南草案[28], 为可信人工智能提出了可追责性、数据治理、普惠性、人工智能自主性的管控、非歧视、尊重和强化人类自治、隐私保护、鲁棒性、安全性、透明性共 10 个基本要求, 这些要求给人工智能提出了可追责性、普惠性、自主性、公平性、隐私性、鲁棒性、安全性、透明性共 8 个可信属性. 美国和 OECD 组织认为, 可信人工智能需要拥有包容性增长、可持续发展和福祉、以人为本的价值观与公平、透明度和可解释性、鲁棒性以及安全性和防危性[77], 这些要求对应了可持续发展、价值观、公平性、透明性、可解释性、鲁棒性、安全性、防危性 8 个可信属性. 美国国家标准与技术研究所 NIST 对人工智能提出其需要拥有互操作性、安全性、可靠性、鲁棒性、公平性和可解释性共 6 个可信属性的要求[78]. IBM 公司则认为, 人工智能系统应拥有公平性、鲁棒性、透明度和问责制、价值观、可解释性、隐私[79]. 这意味着他们认为, 人工智能系统的可信性应具有公平性、鲁棒性、透明性、可追责性、价值观、可解释性、隐私性 7 个可信属性.

许多学者也对人工智能应用的可信属性提出了自己的见解. Singh 等人认为可信人工智能不应具有偏见, 应保证公平性, 模型具有可解释性与透明度、应对对抗性攻击的鲁棒性, 系统本身还要做到隐私保护, 保证安全性的同时也要具有得体性[80]. Singh 等人对人工智能提出了公平性、可解释性、鲁棒性、隐私性、安全性、得体性 6 个可信属性. Fujii 等人则认为人工智能系统要保证数据的完整性、模型的鲁棒性、系统的高质量、过程的敏捷性以及满足客户的期望[81], 他们更加强调完整性、鲁棒性、系统质量以及敏捷性. 而对于 Chatila 等人而言, 他们关注的是人工智能系统的透明性、可验证性、可解释性、安全性与鲁棒性[82]. Ashoori 等人也提出了影响人们信任人工智能的 7 个因素[83], 分别是决策风险、决策者、训练方法、模型的可解释性、训练和测试集的说明、社会透明性与模型置信度. 这些因素强调了可信人工智能的公平性、鲁棒性、可解释性、透明性与置信度. 我国何积丰院士也提出人工智能应具有鲁棒性、自我反省性、自适应性和公平性[12]. 表 1 中列出了各个不同组织机构和学者对可信属性的见解. 这些可信属性从不同的角度提出了对人工智能系统可信性的要求. 从传统软件可信性的角度提出了人工智能系统可信性应满足的要求: 隐私性、安全性、防危性、可靠性、系统质量和敏捷性; 从人工智能模型的黑盒特性对其行为结果产生的影响提出了: 公平性、鲁棒性、可解释性、透明性、可追责性、可验证性和置信度等属性; 从人工智能要为人类服务的角度提出了可信需求: 普惠性、可持续发展和互操作性; 从人工智能应具有人的特征提出了: 自主性、价值观、得体性、自适应性和自我反省性.

表 1 可信属性对比

属性	欧盟[28]	OECD[77]	NIST[78]	IBM[79]	Singh等人[80]	Fujii等人[81]	Chatila等人[82]	Ashoori等人[83]	何积丰[12]
鲁棒性	√	√	√	√	√	√	√	√	√
公平性	√	√	√	√	√	√	—	√	√
可解释性	—	√	√	√	√	—	√	√	—
隐私性	√	—	—	√	√	—	—	—	—
安全性	√	√	√	—	√	—	√	—	—
防危性	—	√	—	—	—	—	—	—	—
可追责性	√	—	—	√	—	—	—	—	—
透明性	√	√	—	√	—	—	√	√	—
普惠性	√	—	—	—	—	—	—	—	—
自主性	√	—	—	—	—	—	—	—	—
价值观	—	√	—	√	—	—	—	—	—
可靠性	—	—	√	—	—	—	—	—	—
互操作性	—	—	√	—	—	—	—	—	—
可持续发展	—	√	—	—	—	—	—	—	—
得体性	—	—	—	—	√	—	—	—	—
敏捷性	—	—	—	—	—	√	—	—	—
系统质量	—	—	—	—	—	√	—	—	—
可验证性	—	—	—	—	—	—	√	—	—
置信度	—	—	—	—	—	—	—	√	—
自适应性	—	—	—	—	—	—	—	—	√
自我反省性	—	—	—	—	—	—	—	—	√

人工智能系统作为软件系统的一种, 其可信性需要满足传统软件的要求, 而作为人工智能系统, 其可信性又要强调其本身的特性和人的特征. 因而, 本文认为人工智能系统的可信属性应包含可靠性、隐私性、安全性、防危性、公平性、鲁棒性、可解释性、自适应性和自我反省性共 9 个属性. 其中, 可靠性要求人工智能系统能提供可靠服务, 其数据、结果等应是可靠的; 隐私性要求人工智能系统能保护其所拥有和使用的数据隐私不被泄露; 安全性要求人工智能系统可以抵抗外来因素, 保护系统信息完整性、机密性和可用性; 防危性要求人工智能系统失效时不会产生不可接受的风险; 公平性要求人工智能系统可以公平地对待所有使用者; 鲁棒性要求人工智能系统可以在受到扰动时输出正确的结果; 可解释性要求人工智能系统中的模型可解释, 其判断过程可以被人类所理解; 自适应性要求人工智能系统在新环境下可以适应, 输出正确的结果; 自我反省性要求人工智能系统对自身性能或错误能够有所感知. 这些属性不仅结合了人工智能系统本身的特性和传统软件可信属性, 而且也从为人类服务的角度考虑, 更加准确地反映了人工智能系统的可信性特征, 为度量人工智能系统的可信性奠定多维属性基础.

3 人工智能系统可信度量证据

人工智能系统可信证据是指可从人工智能系统中提取且用于衡量人工智能系统可信性的相关指标. 因为人工智能系统的可信性问题可以从其训练数据的可信性、学习模型的可信性和预测结果的可信性 3 个方面来考虑, 本节从这 3 个方法讨论了相关度量方法, 并设计了人工智能系统可信证据的收集方法.

3.1 训练数据可信性

数据是人工智能系统中不可或缺的一部分, 训练数据集的可信性直接影响着人工智能系统的可信性, 有效保障和评估数据集的可信性有助于对人工智能系统可信性度量. 数据来源于数据源, 一些学者从数据源的可靠性角度对数据的可信性进行了评估. Ge 等人从关联和比较多数据源的角度出发, 研究数据可信度评估问题[84]. 他们基于数据一致性程度与数据可靠性关联关系提出了一种识别多个数据源一致信息的计算方法, 采用两步过程来计算一组项目的数据一致性程度, 这些项目由不同平台上的多组用户进行评级, 首先构建多源深度信念网络 (MSDBN)

来识别隐藏在多源评级数据中的常见原因, 然后通过将单个来源与从潜在原因中导出的重构数据进行比较来计算每个项目的一致性得分. Tabibian 等人针对在线知识库数据可靠性和数据源可信度问题, 从用户和专业编辑者对知识库内容可靠性的噪声评估数据入手, 以噪声评估数据所留下的“时间痕迹”为线索提出了一个“时间点过程建模框架”, 并将这些“时间痕迹”与数据可靠性和数据源可信度的健壮性、公平性和可解释性概念联系起来, 基于凸优化技术从“时间痕迹”数据中学习模型的参数[85]. Fogliaroni 等人阐述了志愿地理信息 (VGI) 数据的质量评价问题, 并针对该问题提出了一个基于版本更新的 VGI 质量指标量化模型, 该模型依赖于出处信息, 为地理特征和 VGI 系统的贡献者分别推导可信度和声誉得分[86]. Zhang 等人提出了一个在多个数据源提供声明和支持性证据, 并且每个声明可能由多个数据源产生的环境中估计数据源可信度的通用框架 JELTA, 并开发了一系列概率模型, 共同估计来源以及断言的可信度[87].

人工智能系统所需的数据在数据采集、数据存储等各个环节存在隐私泄露的风险[88], 近年来关于数据隐私泄露度量的研究也炙手可热, 这些学者的研究对数据的隐私性和安全性的度量打下了基础. Liu 等人提出了一种轻量级隐私保护信任评估方案, 用于协同车辆安全应用中的分布式数据融合, 该方案能够在较低计算、通信和存储开销基础上, 较好地平衡信任评估和隐私保护, 从而促进协同车辆安全应用中分布式数据融合[89]. Xu 等人针对智能交通研究中车辆拥挤传感器系统中恶意节点生成虚假事件报告问题, 提出了一种轻量级辅助车辆拥挤感知框架 TPSense, 保证数据可信性和用户隐私: 将数据可信度评估问题转化为极大似然估计问题, 并通过基于期望最大化 TEEM 算法求解, 完成事件报告可信度评估和车辆可靠性评估, 达到在路侧单元上过滤虚假事件报告的目的[90]. Tsou 等人首次利用差异隐私的噪声估计评估数据泄露风险, 使用数据噪声量作为桥梁来评估多个属性 (数值或二进制数据) 的数据泄露风险, 并制定差分隐私和匿名化之间的关系, 将两者关联起来[91].

区块链作为保障数据不可被篡改的技术, 近年来一直被用于提高数据可信性, 为数据的安全性提供了保障. Ardagna 等人在可信物联网保障评估的基础上, 创新性地提出了一种基于服务的可信证据收集原子方法, 并将其作为实现可信物联网环境的基础[92]. 该方法将收集证据和汇总证据的方式联系, 或根据证据做出可信决策, 平衡所提供证据的可信度水平及其性能. 该方法采用智能合约技术要求从每个智能设备收集的数据必须首先进行评估, 只有在满足最低保证要求的情况下才能投入使用, 并使用区块链作为数据存储库, 存储可信证据收集和评估的所有交易. Zhang 等人将联邦学习的模型质量参数作为衡量候选员工可信声誉的指标, 以实现联邦学习过程中可信员工的选择: 使用交叉熵来计算联邦学习员工的可信声誉, 给出归一化公式, 并利用区块链技术抗篡改和不可抵赖性, 设计区块链平台来管理记录下的声誉, 从而提高联邦学习的可信性[93]. Distefano 等人采用分布式账本技术实现一个以车辆为中心的信息系统, 通过网络分发数据, 同时确保可信度. 该文面向以可信车辆为中心的智慧交通系统领域, 构建并实现一个信息系统: 使用分布式数据库 MongoDB 和区块链多链技术来分别存储该系统中的数据和相关元数据, 以确保整个系统的可信度和性能/可扩展性之间的平衡[94].

3.2 学习模型可信性

人工智能系统模型是人工智能系统的关键要素之一, 是人工智能系统的灵魂所在, 其可信性对人工智能系统的可信性有着至关重要的影响. 有关人工智能系统模型可信性度量的定义众说纷纭, 莫衷一是, 本文将人工智能系统模型的可信性定义为人工智能系统模型本身所具有的可信属性, 如可解释性、鲁棒性、隐私保护能力等属性, 因而对人工智能系统模型可信性的度量即是对这些模型本身性质的度量.

近年来, 由于大多数机器学习模型的黑盒特性, 越来越多的学者注重模型可解释性的度量. Bau 等人提出网络解剖的方法来度量模型可解释性, 网络解剖依赖于密集标记的数据集合, 这些数据集合被标记上了颜色、材质、纹理、场景等诸多标签, 在给定 CNN 模型的基础上, 使用网络解剖寻找语义神经元, 通过语义神经元的数量及其所有神经元的比例来度量模型解释性的分数[95]. Slack 等人通过用户研究实验的方式来评估可解释性, 他们设计了 1 000 名参与者参与的用户研究实验, 系统地比较了决策树、逻辑回归和神经网络 3 类模型的可解释性[96]. Sanneman 等人提出了一个基于人类用户信息需求的可解释人工智能系统框架, 框架包括了可解释人工智能系统的 3 个级别, 定义了可解释的人工智能系统应该支持哪些关于人工智能的算法和流程信息[97]. Rosenfeld 尝试对可

解释人工智能模型量化, 并提出了 4 个量化指标来度量模型的可解释性, 指标涉及解释后的模型和实际模型间的性能差异, 解释方式中所使用的规则数目, 以及模型在解释时需要的特征数目和解释模型的稳定性[98]. Lin 等人提出了一个系统性自动评估人工智能系统的可解释性框架, 通过检查人工智能模型是否能够检测到输入中存在的后门, 形成输出特定的预测结果, 并使用 3 种度量指标量化人工智能系统的可解释性, 整个过程无需人工干预, 可以自动量化人工智能模型的可解释性[99].

鲁棒性是近年来人工智能模型所面临的另一个难题, 许多学者也在尝试度量人工智能模型的鲁棒性. Ruan 等人将全局鲁棒性定义为在测试数据集中的最大安全半径的数学期望, 并提出了一种基于 Hamming 距离的深度神经网络全局鲁棒性评估算法, 通过迭代计算最大安全半径上下界来近似度量深度神经网络的全局鲁棒性[100]. Yu 等人通过损失可视化来定性解释对抗攻击和防御机制, 并建立了评价神经网络模型鲁棒性的量化指标. 该指标通过结合一种新的正则化方法, 在任何状态下不变地展示神经网络模型的鲁棒性[101]. 还有许多学者使用不同的对抗攻击形式来分析模型鲁棒性, 例如文献 [102−104] 等都面向特定领域构造了对抗性示例来分析模型鲁棒性.

人工智能模型需要大量数据训练模型, 通过一些模型的输出可以倒推训练集中某条目标数据的部分或全部属性值, 因而可能会对造成数据隐私泄露[88,105], 许多学者也在度量模型的隐私保护能力方面做了研究. Song 等人提出了“基准隶属度推理隐私风险”和一种基于预测熵修正的推理攻击方法: 用基准攻击来补充现有基于神经网络的攻击, 以有效地度量隐私风险; 并介绍了一种用于细粒度隐私分析的新方法: 通过构造和派生一个称为“隐私风险评分”的新度量, 以度量个体样本成为训练成员的可能性, 帮助敌手识别具有高隐私风险的样本, 并以高置信度执行成员推理攻击; 考虑上述攻击方法的隐私防御, 将现有的总体隐私分析和其提出的细粒度隐私分析相结合, 以系统地衡量隐私风险[106]. Ma 等人针对运动员成绩记录聚类过程中的数据隐私泄露问题, 使用隐私感知的近似近邻搜索技术 SimHash, 通过分析分布在不同云平台上的运动得分记录, 对相似的球员进行聚类, 实验证明该方法能有效解决基于运动成绩记录的球员聚类中存在的数据量庞大和隐私泄露问题[107].

此外, Yang 分析了对抗性样本产生的根本原因, 提出了机器学习模型的一个新性质, 即保真度, 用来描述模型所学知识与人类所学知识之间的差距[108], 从而保障人工智能系统的可靠性. Yang 对保真度做出了明确的定义, 并提出了一种方法, 使用一组传统的机器学习模型作为评判标准来计算保真度. de Bie 等人提出了一种评估和解释回归预测模型可信度的方法 RETRO-VIZ[109], 分析了人工智能系统的可靠性与可解释性. 该方法由两部分组成: 一部分是 RETRO, 用来定量估计预测的可信度; 另一部分 VIZ 则是一种帮助使用者理解预测可信度的可视化解释. 他们在 117 个实验中使用了该方法, 发现 RETRO-VIZ 分数与预测误差呈现负相关. 他们还对用户做了相关调研, 大多数人认为其解释有助于用户理解.

3.3 预测结果可信性

自人工智能技术诞生以来, 人们就对其结果的可信性非常关心. 本文将人工智能系统模型可信性和结果可信性分开, 将人工智能系统结果可信性的度量定义为通过人工智能的结果来反应人工智能系统整体可信性的方法. 在人工智能的分类任务中, 学者们把分类结果分为真正例 (*TP*)、假正例 (*FP*)、真反例 (*TN*) 和假反例 (*FN*), 分别代表其判断结果有多少是判断正确的正例、判断错误的正例、判断正确的反例和判断错误的反例, 通过错误率和精度两种方法来度量人工智能系统的分类结果是否可信[110]. 其中, 错误率 (*ErrorRate*) 是指在分类任务中分类错误的样本数占总样本数的比例, 即:

$$ErrorRate = \frac{FP+FN}{TP+FP+TN+FN}.$$

精度 (*Accuracy*) 则是分类正确的样本数占总样本数的比例, 即:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}.$$

然而错误率和精度虽然是常用的两种度量, 但是对于大多数人工智能的分类任务而言, 不能仅使用人工智能系统输出结果是否正确这一指标来判断结果是否可信, 大多数任务会更加注重人工智能判断的结果有多少比例是正确的. 因此, 学者们对于这类问题, 又定义查准率 (*PrecisionRate*), 真正例占正例的百分比, 即:

$$PrecisionRate = \frac{TP}{TP+FP}$$

和查全率 (*RecallRate*), 真正例占预测正确的百分比, 即:

$$RecallRate = \frac{TP}{TP+FN},$$

从而判断结果的可信性. 在此基础上, 人们又提出了 P-R 曲线[110]、ROC 曲线[111]与 AUC[112]来进一步可视化判断人工智能系统预测结果的可信性.

而对于回归任务而言, 最常用度量则是均方误差, 其对应欧氏几何中的距离度量. 例如在线型回归中, 均方误差就是所有样本到直线上的平均欧式距离, 这个距离可以很好地表示回归任务结果的可信性. 此外, 对于一些需要真实体现出误差的任务, 人们会使用平均绝对误差来表示回归任务结果的好坏.

近年来, 也有学者在置信度方面做出了研究. Jha 等人为深度神经网络 (DNN) 提出了一种新的置信度度量, 基于归因的置信度度量, 它可以表征是否可以信任输入上的 DNN 的输出, 该方法通过对给定高维输入的邻域内进行归因驱动抽样来计算结果置信度, 从而度量模型预测的一致性, 它不需要访问训练数据或进行额外校准, 因此具有很好的实用性[113]. Waa 等人为支持决策系统定义了一种可解释的置信度框架, 该框架认为置信度应该满足 4 个性质, 分别是准确、能够解释单个置信度值、使用透明的算法和提供可预测的置信度值[114]. 基于这 4 个性质, 他们定义了基于案例推理的回归分析置信度度量, 在具有不同机器学习模型的若干分类任务上评估了其准确性, 稳健性和普遍性. 此外, 作者还在海域内动态定位领域应用可解释置信度概念.

3.4 人工智能可信证据收集方法

人工智能系统训练数据的可信性、学习模型的可信性和预测结果的可信性共同影响着人工智能系统整体可信性, 所以人工智能系统的可信证据收集可以从数据、模型、结果可信性度量方面考虑. 本节在前文介绍的现有数据、模型、结果可信性度量研究基础上, 提出人工智能系统可信证据收集方法.

在人工智能系统训练数据可信性方面, 大多研究认为数据源可靠性、数据隐私泄露风险和数据不可篡改性会对其可信性造成影响. 因而, 对于人工智能数据可信证据, 其收集方法考虑量化数据源可靠性、训练数据隐私泄露风险和训练数据不可篡改程度.

在人工智能系统学习模型可信性方面, 人工智能系统面临学习模型不可解释、对抗攻击影响、模型隐私泄露和模型自身保真度问题. 因而, 对于人工智能模型可信证据, 其收集方法考虑分析模型可解释与和转化程度、模拟对抗攻击的影响、量化模型隐私泄露风险和度量模型保真度.

在人工智能系统预测结果可信性方面, 研究者们对其可信度度量研究已经有一定的成果, 涉及了一系列指标. 因而, 对于人工智能结果可信证据, 其收集方法考虑综合预测结果的错误率、精度、查准率、查全率、误差程度和置信度.

4 人工智能系统可信度量模型

在之前的部分, 我们讨论了人工智能系统的可信属性, 总结梳理了人工智能系统的 9 个可信属性. 然后, 为了度量人工智能系统可信性, 本文从训练数据可信性、学习模型可信性和预测结果可信性 3 方面讨论了度量方法, 为可信性度量提供可信证据. 近年来, 有部分学者对人工智能系统的可信度量评估进行了研究, 本节将着重介绍这些研究.

文献 [115,116] 从软件质量的度量出发, 对人工智能系统进行度量. Shepperd 等人回顾了软件度量的相关概念, 并进一步介绍了人工智能和软件工程的关系, 他认为软件度量的许多度量指标都可以被应用在专家系统的管理和质量保证方面, 包括软件结构度量等, 但是由 Boehm[117]提出的软件成本估算方法 COCOMO 无法直接应用于专家系统[115]. Nakamichi 等人重点研究了机器学习系统所要求的质量特性, 以满足企业信息系统的质量要求. 他们通过确定与机器学习系统需求规范相关的 22 个问题, 包括环境/用户、系统/基础设施、模型和数据, 将传统的软件质量标准 ISO25010 的质量扩展到机器学习系统所拥有的特性, 提出了评价质量特性的指标和测量方法, 并通过

一个行业的实证研究, 验证了所提出模型和测量方法的可行性[116].

也有一些学者针对不同类型的人工智能系统提出了度量模型. Cheng 等人提出了一个基于主观逻辑的神经网络可信量化框架 DeepTrust, 其构建了人工智能算法的概率逻辑描述, 同时考虑了数据集和内部算法工作的可信度[118]. DeepTrust 同时适用于分类和回归问题, 其输入值不影响可信度的计算. DeepTrust 不仅能够在训练数据和训练过程可访问的前提下, 对训练阶段神经网络的意见进行量化, 而且在给定预训练神经网络的情况下, 它还可以用于神经网络决策或输出的可信量化. 神经网络预测的意见和可信度量化提供了对输入数据和神经网络内部工作可信度的评估, 并且在神经网络过拟合时非常有效. Uslu 等人则首次提出了一种可信指标框架来评估人工智能系统在食品、能源和水管理决策中的验收标准, 在评估专家选择每个人工智能系统提出的最合适的解决方案中, 计算从这些人工智能系统提出的解决方案到专家给出的最佳参考方案的距离, 并使用提出的信任框架计算系统的可信接受度[119]. 此外, 他们通过信任系统来聚合了多个专家的度量评估结果. Chowdhury 等人使用车载单元 (OBU) 组件、GPS 数据和安全信息来确定自动驾驶车辆的可信性, 通过计算和处理相关的不确定性, 利用确定性逻辑和主观逻辑开发了评估可信值的理论模型, 并对提出的模型产生的可信值进行了对比分析[120]. Chowdhury 等人还在模拟城市交通平台 SUMO 和澳大利亚历史交通数据 VicRoads 验证提出的模型, 其中, 基于特定逻辑的模型产生的可信值比主观逻辑产生的可信值对被破坏的组件更加敏感. Chattopadhyay 等人对机器人进行可信评估, 分析了 Zumi 和 Cozmo 社交机器人的内部缺陷, 并使用 IEEEA/IS 可信 AI 指南对其总体设计进行了评估, 研究该社交机器人是否与可信人工智能框架描述的原则相一致[121].

总体而言, 学者们近几年来开始愈发关注人工智能系统可信性, 并且尝试对其可信性进行评估, 从而开发出更加可信的人工智能产品. 目前, 大多数研究聚焦在如何改善和提高人工智能系统可信性, 而人工智能可信度量评估并没有受到很大关注, 因此也没有像软件可信度量评估那样, 经过数十年发展积累丰厚的研究成果. 因此, 人工智能可信度量评估领域还需要进行深入系统的研究.

5 基于属性的人工智能系统可信度量评估框架

尽管目前人工智能系统在可信性度量评估已经有部分研究成果, 但鲜有研究从基于属性的软件可信度量的角度度量人工智能系统的可信性. 人工智能系统作为软件系统, 其可信量化评估也可以使用软件可信度量理论来实施. 本文提出一个人工智能可信度量评估框架, 将软件可信度量理论应用于人工智能系统的可信度量评估, 而且为了保证度量过程数据的可信性不被篡改, 本文将保证数据可信性常用的区块链技术应用于框架中, 整个框架包括人工智能系统可信属性分解与证据获取方法、联邦式可信度量模型以及基于区块链的人工智能可信评估架构 AITrust 这 3 个部分. 本文期望该框架可以为人工智能系统可信度量研究人员提供参考, 以推动人工智能系统可信度量研究进一步发展.

5.1 人工智能系统可信属性分解与证据获取方法

“可信性”是客观对象诸多属性在人们心目中的一个综合反映[14], 而每个属性在不同时段的含义又有所不同. 因此, 为对人工智能系统进行可信评估, 首先需要将人工智能系统的整体可信属性分解, 结合系统所采用的智能模型及其领域特性, 根据人工智能系统在系统构建不同阶段的可信需求, 对可信属性赋予合适的含义, 解释其包含的不同含义, 将可信属性进一步分解出可信子属性. 例如, 把人工智能系统构建划分为系统开发, 系统验证和系统测试 3 个阶段, 各阶段可信属性划分的可信子属性不一定相同, 在人工智能系统的构建阶段, 可解释性主要指人工智能系统的数据准确能被人类所理解, 模型的判别或者推理过程可以转化成具备逻辑关系的规则以及代码易读易理解, 因此把可解释性划分为数据准确性、模型可转化性和代码易读性; 在系统验证阶段, 可解释性主要指模型的判别或者推理过程可以转化成具备逻辑关系的规则、代码易读易理解以及系统输出结果能被人类理解和分析, 因此把可解释性划分为模型可转化性、代码易读性和结果可分析性; 在系统测试阶段, 可解释性主要指人工智能系统的数据准确能被人类所理解, 模型的判别或者推理过程可以转化成具备逻辑关系的规则以及系统输出的结果能被人类理解和分析, 因此把可解释性划分为数据准确性、模型可转换性和结果可分析性. 图 1 中给出了可解释性在

3 个阶段的子属性划分对比.

根据不同属性划分出的子属性, 依据其具体含义, 对子属性中包含的可信证据规范进行获取, 并把可信证据规范划分为度量元. 表 2 给出了在系统构建阶段模型可转换性子属性下获取的可信证据与其度量元.

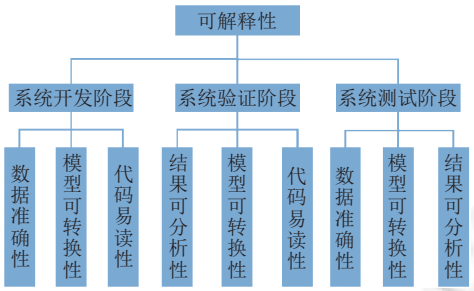

图 1 可解释性 3 个阶段子属性划分对比

表 2 系统构建阶段可信证据示例

可信属性	可信子属性	可信证据	度量元
可解释性	模型可转换性	系统所用模型判别或者推理过程可以转化成具备逻辑关系的规则	A: 系统所用模型判别或者推理过程可以转化成具备逻辑关系的规则 B: 系统所用模型判别或者推理过程可以转化成规则 C: 系统所用模型判别或推理过程可以被解释 D: 系统所用模型判别或推理过程无理论解释

5.2 联邦式可信度量模型

由于人工智能系统在构建的各个阶段往往分布在不同开发团队与用户群体中, 各过程数据可能不互通, 导致数据孤岛问题. 此时的人工智能系统可信度难以量化评估, 有效融合各阶段的可信性度量结果, 形成较为准确的可信度量值是人工智能系统可信度量评估的另一挑战. 本节给出一个在人工智能开发、验证、测试阶段的联邦式可信度量模型结构图, 用于计算融合各阶段的可信度量结果, 其结构如图 2 所示. 构建、验证和测试阶段在属性 i 上分别拥有度量值 y_{ji} 和该属性在此阶段所占的权重 α_{ji}, 联邦式可信度量模型建立融合公式 f,g 用于融合 3 个阶段下的度量值和权重, 得到属性 i 的可信度量值. 然后, 联邦式可信度量模型给出计算公式 F, 用于在得到各个属性的度量值后, 计算出系统可信度量值.

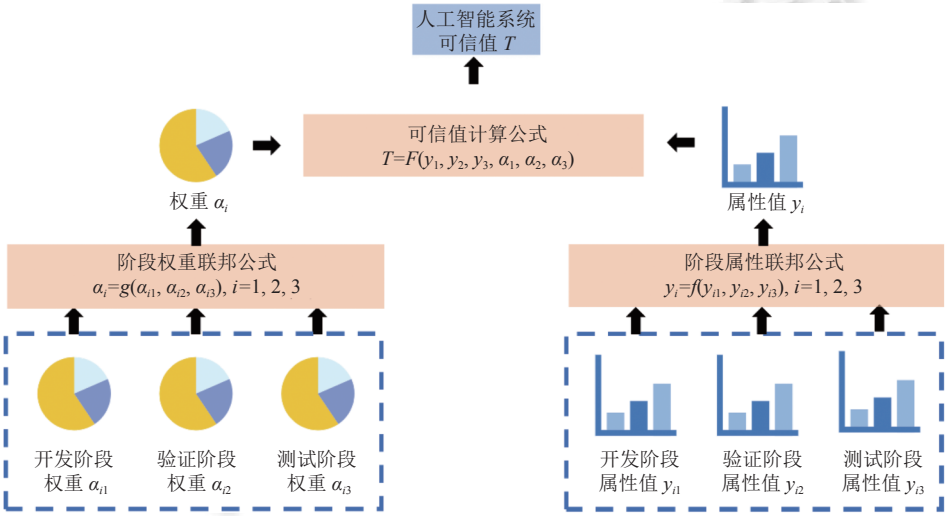

图 2 联邦式可信评估结构图

5.3 基于区块链的人工智能系统可信评估架构 AITrust

区块链技术是保障数据可信性常用的技术之一, 区块链本身的特性保证了处于其中数据的不被篡改与可追溯. 而在对人工智能系统进行评估时, 需要涉及多个阶段的数据, 例如系统构建阶段的数据, 系统测试阶段的测试数据等. 为了确保这些数据的准确性与不被篡改, 从而确保度量结果的可信性, 本节将区块链技术融入人工智能可信度量评估, 给出拥有在人工智能系统开发、验证、测试 3 个阶段的构建过程的基于区块链的人工智能可信评估架构, 其由一列可信区块 $B_1, B_2, B_3, ..., B_t$ 组成, 每个可信块 $B_i = (T_{ri}, h_{i-1}, e_i)$, 其中, T_{ri} 交易至少包含: ① 人工智能系统, ② 人工智能构建模型, ③ 验证模型, ④ 测试方法, ⑤ 评估体系, ⑥ 人工智能可信级别, 并按照 Merkle 树表示, h_{i-1} 是一指向前块的密码哈希 ($h_{i-1} = hash(B_{i-1})$), e_i 是一个校验码, 存储评估竞争过程中的元数据, 用来校验区块的有效性. 人工智能系统可信评估区块链的建立过程涉及至少 4 个类型的角色: 发布人工智能系统的发布者、对人工智能系统进行系统验证的验证者、对人工智能系统进行系统测试的测试者与对人工智能系统进行可信评估的评估者. 人工智能系统可信评估区块链结构如图 3 所示.

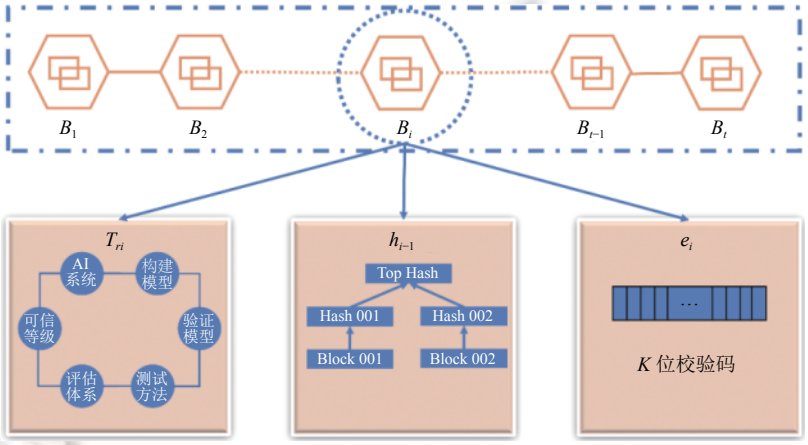

图 3 基于区块链的人工智能系统可信评估架构

当发布者发布了一个人工智能系统后, 其相关的开发数据会被存入其发布的区块链中, 测试者和验证者分别对该人工智能系统进行测试和验证, 并把测试与验证的结果和相关数据存储其发布的区块链中, 评估者从他们建立的区块链中获取系统构建、系统测试与系统验证阶段的数据进行评估, 得到该系统的可信等级, 然后增加可信块将其写入区块链中, 从而可以依靠区块链技术保证整个评估过程与评估结果的可追溯性和可信性.

6 总结与展望

本文对人工智能系统的可信性度量进行了综述性的分析和研究. 首先对人工智能技术和软件可信度量的背景进行了讨论, 分别对两个方向的研究现状进行了论述, 并阐述了人工智能系统可信度量的动机和可信人工智能的战略与规划. 然后, 本文对人工智能系统的可信属性进行详细探讨, 比较了各个组织和学者给出的可信属性, 提出了人工智能系统基础的 9 个可信属性. 接着, 分别从训练数据可信性、学习模型可信性、预测结果可信性方面讨论了相关度量方法, 设计了人工智能系统可信证据收集方法. 然后, 讨论了现有的人工智能系统可信度量方法, 结合基于属性的软件可信度量技术, 本文建立一个人工智能系统可信度量评估框架, 该框架包括可信属性分解与证据获取方法, 联邦式可信度量模型和基于区块链的人工智能系统可信评估架构 AITrust, 将基于属性的软件可信度量技术应用于人工智能系统的可信评估, 并且使用区块链技术保障了评估过程的可信性.

人工智能系统可信性已经获得越来越多人的关注, 人工智能系统可信度量评估可以保障人工智能系统的可信性, 为人工智能系统开发与使用人员提供相应参考. 然而, 人工智能系统可信度量的发展仍然面临着许多问题和挑战.

(1) 针对人工智能系统的可信属性, 各个机构与学者关注点不尽相同, 本文虽然讨论并总结梳理出了 9 个人工智能基础的可信属性, 然而对于面向不同领域的人工智能系统, 使用不同机器学习模型的人工智能系统, 其可信属性可能会有所扩展, 且其度量模型也有可能不同. 因此, 人工智能可信属性的完备性值得关注, 从而设计面向不同领域的不同系统、不同模型、甚至不同数据的可信属性集合以及其度量模型, 形成更加完备的人工智能可信属性模型.

(2) 就人工智能系统的可信证据而言, 在人工智能系统的训练数据可信性、学习模型可信性和预测结果可信性度量方面虽然有许多研究, 但是这些研究过于杂乱, 量纲难以统一, 很难形成统一的可信证据度量模型, 为后续可信度量模型的建立造成一定难度. 因而, 统一可信证据量纲, 形成科学合理的可信证据模型同样值得关注.

(3) 目前, 面向人工智能系统的可信度量模型研究还比较初步, 相关的研究还比较少, 本文提出了一个面向属性的人工智能系统可信度量框架, 但其中的可信度量模型还需进一步完善, 从而建立科学合理的人工智能系统可信度量模型, 使其度量结果具有可信性. 并且, 本文建立的人工智能系统可信度量框架是面向可信属性, 仅从静态角度来度量可信性, 并未考虑人工智能系统运行时的动态变化与用户反馈. 因此, 建立科学合理且全面的人工智能系统可信度量评估模型是人工智能可信度量评估的又一挑战.

(4) 在建立人工智能系统可信度量方法的基础上, 开发人工智能系统可信度量评估综合工具, 使得人工智能系统可信度量评估工程化, 把人工智能系统的度量评估融入人工智能系统的实际开发和使用过程中, 提高人工智能系统的可信性, 为人工智能系统更好地服务人类奠定基础.

希望通过本文的讨论、探索、思考和提出的人工智能系统可信度量评估框架, 为人工智能系统可信度量评估的发展和应用开辟更加广阔的道路.

References:

[1] Fang BX. Artificial Intelligence Safety and Security. Beijing: Publishing House of Electronics Industry, 2020. 1–10 (in Chinese)

[2] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]

[3] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484–489. [doi: 10.1038/nature16961]

[4] Notice of the State Council on printing and distributing the development plan of new generation artificial intelligence, GF-[2017] No. 35 Development plan of new generation artificial intelligence. 2017 (in Chinese). http://ai.cumt.edu.cn/info/1024/1052.htm

[5] Institutes of Science and Development, Chinese Academy of Sciences, Research Support Center for Science Popularization and Education Faculty of Chinese Academy of Sciences. Commentary science and technology hotpots in China 2019. Beijing: Science Press, 2020. 89–137 (in Chinese).

[6] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: ICLR, 2014.

[7] Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao CW, Prakash A, Kohno T, Song D. Robust physical-world attacks on deep learning visual classification. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1625–1634. [doi: 10.1109/CVPR.2018.00175]

[8] Xu H, Ma Y, Liu HC, Deb D, Liu H, Tang JL, Jain AK. Adversarial attacks and defenses in images, graphs and text: A review. Int'l Journal of Automation and Computing, 2020, 17(2): 151–178. [doi: 10.1007/s11633-019-1211-x]

[9] 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]

[10] Cazzell A. Why algorithmic fairness is elusive. 2019. https://hackernoon.com/why-algorithmic-fairness-is-elusive-sf7v323b

[11] Sun JY. Research report on the current situation of taxi-hailing software in 2020. School of Management, Fudan University-Fudan Young Entrepreneur Education and Research and Development Center. 2021. 3. 40–56 (in Chinese). https://mp.weixin.qq.com/s/G2VzL9QJJU4Acsl8VvtGhw

[12] He JF. Secure and trustworthy artificial intelligence. Information Security and Communication Secrecy, 2019(10): 5–8 (in Chinese with English abstract).

[13] Yang W. The Fundamental Research for Trustworthy Software. Hangzhou: Zhejiang University Press, 2018. 1–26 (in Chinese).

[14] Liu K, Shan ZG, Wang J, He JF, Zhang ZT, Qin YW. Overview on major research plan of trustworthy software. Bulletin of National Natural Science Foundation of China, 2008, 22(3): 145–151 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-8217.2008.03.005]

[15] CCF Formal Professional Committee. Research progress and trend of formal verification technology for artificial intelligence system. China Computer Science and Technology Development Report from 2019 to 2020. 2020. 491–539 (in Chinese).

[16] ISO/IEC 15408-2: 2020 information technology-security techniques-evaluation criteria for IT security Part 2: Security functional components. 2020.

[17] Chen YX, Tao HW. Software Trustworthiness Measurement Evaluation and Enhancement Specification. Beijing: Science Press, 2019. 1–19 (in Chinese).

[18] Turing AM. I.—Computing machinery and intelligence. Mind, 1950, LIX(236): 433–460. [doi: 10.1093/mind/LIX.236.433]

[19] McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. AI Magazine, 2006, 27(4): 12–14. [doi: 10.1609/aimag.v27i4.1904]

[20] Jang YK, Cho NI. Generalized product quantization network for semi-supervised image retrieval. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3417–3426. [doi: 10.1109/CVPR42600.2020.00348]

[21] Jia W, Dai D, Xiao XY, Wu H. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACM, 2019. 1399–1408. [doi: 10.18653/v1/P19-1135]

[22] Li YJ, Liu MY, Li XT, Yang MH, Kautz J. A closed-form solution to photorealistic image stylization. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 468–483. [doi: 10.1007/978-3-030-01219-9_28]

[23] Li PL, Chen XZ, Shen SJ. Stereo R-CNN based 3D object detection for autonomous driving. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7636–7644. [doi: 10.1109/CVPR.2019.00783]

[24] Pi Q, Bian WJ, Zhou GR, Zhu XQ, Gai K. Practice on long sequential user behavior modeling for click-through rate prediction. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019. 2671–2679. [doi: 10.1145/3292500.3330666]

[25] House of Commons, Science and Technology Committee. Robotics and artificial intelligence. 2016. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

[26] European Political Strategy Centre. The age of artificial intelligence: Towards a European strategy for human-centric machines. 2018. https://ec.europa.eu/jrc/communities/sites/jrccties/files/epsc_strategicnote_ai.pdf

[27] Communication Artificial Intelligence for Europe. 2018. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[28] Ethics guidelines for trustworthy AI. 2019. https://www.i-programmer.info/programming/artificial-intelligence/12702-ethics-guidelines-for-trustworthy-ai-.html

[29] NSTC (National Science and Technology Council). Preparing for the future of artificial intelligence. 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[30] NSTC (National Science and Technology Council). The national artificial intelligence research and development strategic plan. 2016. https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

[31] NSF (National Science Foundation). National robotics initiative 2.0: Ubiquitous collaborative robots (NRI-2.0). 2019. https://www.nsf.gov/pubs/2019/nsf19536/nsf19536.pdf

[32] Summary of the 2018 department of defense artificial intelligence strategy. 2019. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

[33] Algorithmic warfare: DARPA's 'AI Next' program bearing fruit. 2019. https://www.nationaldefensemagazine.org/articles/2019/7/2/algorithmic-warfare-darpas-ai-next-program-bearing-fruit

[34] AI. SG: New national programme to catalyse, synergise and boost Singapore's, artificial intelligence capabilities. 2017. https://www.nrf.gov.sg/docs/default-source/modules/pressrelease/201705031442082191-press-release-(ai.pdf

[35] National strategy for artificial intelligence #AIFORALL. 2019. https://www.niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf

[36] The State Council. The State Council's guiding opinions on actively promoting the "Internet plus" action. 2015 (in Chinese). http://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm

[37] The State Council. The 13th five year plan of national science and technology innovation. 2016 (in Chinese). http://www.gov.cn/

zhengce/content/2016-08/08/content_5098072.htm

[38] Ministry of Industry and Information Technology. Three year action plan for promoting the development of new generation AI industry (2018–2020). 2017 (in Chinese). http://www.cac.gov.cn/1122114520_15132987738211n.docx

[39] China Institute of Electronic Technology Standardization. White paper on artificial intelligence Standardization (2018 Edition). Beijing: National Standardization Management Committee. 2018 (in Chinese).

[40] Jin Z, Wrote; Ye LL, Trans. History of Fatal Software Errors. Beijing: Posts & Telecom Press Company, 2016. 1–7 (in Chinese).

[41] Anderson JP. Computer security technology planning study, volume II. Technical Report ESD-TR-73-51, Deputy for Command and Management Systems, HQ Electronics Systems Division (AFSC), 1972.

[42] Lipner SB. The birth and death of the orange book. IEEE Annals of the History of Computing, 2015, 37(2): 19–31. [doi: 10.1109/MAHC.2015.27]

[43] Laprie JC. Dependable computing and fault tolerance: Concepts and terminology. In: Proc. of the 25th Int'l Symp. on Fault-tolerant Computing, 'Highlights from Twenty-Five Years'. Pasadena: IEEE, 1995. 2. [doi: 10.1109/FTCSH.1995.532603]

[44] National Science and Technology Council. Research challenges in high confidence systems. In: Proc. of the 1997 Committee on Computing Information and Communications Workshop. Alexandria, 1997.

[45] Schneider F. Trust in Cyberspace. Washington: National Academy Press, 1999. 154–169.

[46] Hasselbring W, Reussner R. Toward trustworthy software systems. Computer, 2006, 39(4): 91–92. [doi: 10.1109/MC.2006.142]

[47] Chen HW, Wang J, Dong W. High confidence software engineering technologies. Acta Electronica Sinica, 2003, 31(S1): 1933–1938 (in Chinese with English abstract). [doi: 10.3321/j.issn:0372-2112.2003.z1.001]

[48] National Science Foundation. Computer Systems Research (CSR) nsf05629. 2006. https://www.nsf.gov/pubs/2005/nsf05629/nsf05629.htm

[49] Executive Office of the President, National Science and Technology Council. The networking and information technology research and development (NITRD) program 2012 strategic plan. 2012. https://www.nitrd.gov/PUBS/strategic_plans/2012_NITRD_Strategic_Plan.pdf

[50] European Union. Open Trusted Computing (OpenTC). 2009. https://www.itas.kit.edu/english/projects_webe05_otc.php

[51] European Union. Fifth RTD framework programme, 1998–2002. 2002. https://cordis.europa.eu/programme/id/FP5

[52] European Union. Sixth framework programme. 2006. https://europeanlaw.lawlegal.eu/6th-framework-programme/

[53] The State Council. Outline of the national medium and long-term science and technology development plan. 2006 (in Chinese). http://www.gov.cn/jrzg/2006-02/09/content_183787.htm

[54] Ministry of Science and Technology. High-confidence software production and integrated environment project (863 plan) application. 2009 (in Chinese). http://www.edu.cn/ke_yan_yu_fa_zhan/gai_kuang/xin_wen_gong_gao/200904/t20090413_372055.shtml

[55] Taibi D, Del Bianco V, Carbonare DD, Lavazza L, Morasca S. Towards the evaluation of OSS trustworthiness: Lessons learned from the observation of relevant OSS projects. In: Proc. of the 20th IFIP Int'l Conf. on Open Source Systems. Milano: Springer, 2008. 389–395. [doi: 10.1007/978-0-387-09684-1_37]

[56] Del Bianco V, Lavazza L, Morasca S, Taibi D. Quality of open source software: The qualipso trustworthiness model. In: Proc. of the 5th IFIP Int'l Conf. on Open Source Systems. Skövde: Springer, 2009. 199–212. [doi: 10.1007/978-3-642-02032-2_18]

[57] Lavazza L, Morasca S, Taibi D, Tosi D. Predicting OSS trustworthiness on the basis of elementary code assessment. In: Proc. of the 2010 ACM-IEEE Int'l Symp. on Empirical Software Engineering and Measurement. Bolzano-Bozen: ACM, 2010. 36. [doi: 10.1145/1852786.1852834]

[58] Boland T, Cleraux C, Fong E. Toward a preliminary framework for assessing the trustworthiness of software. Gaithersburg: National Institute of Standards and Technology, 2010.

[59] Alexopoulos N, Habib SM, Schulz S, Mühlhäuser M. M-STAR: A modular, evidence-based software trustworthiness framework. arXiv:1801.05764, 2018.

[60] Cho JH, Xu SH, Hurley PM, Mackay M, Benjamin T, Beaumont M. STRAM: Measuring the trustworthiness of computer-based systems. ACM Computing Surveys, 2019, 51(6): 128. [doi: 10.1145/3277666]

[61] Yang SL, Ding S, Chu W. Trustworthy software evaluation using utility based evidence theory. Journal of Computer Research and Development, 2009, 46(7): 1152–1159 (in Chinese with English abstract).

[62] Ding S, Yang SL, Fu C. A novel evidential reasoning based method for software trustworthiness evaluation under the uncertain and unreliable environment. Expert Systems with Applications, 2012, 39(3): 2700–2709. [doi: 10.1016/j.eswa.2011.08.127]

[63] Zheng ZM, Ma SL, Li W, Wei W, Jiang X, Zhang ZL, Guo BH. Dynamic characteristics and evolution complexity of software trustworthiness. Science China (Information Sciences), 2009, 39(9): 946–950 (in Chinese with English abstract).

[64] Zheng ZM, Ma SL, Li W, Jiang X, Wei W, Ma LL, Tang ST. Complexity and dynamic statistical analysis method of software trustworthiness. Science China (Information Sciences), 2009, 39(10): 1050–1054 (in Chinese with English abstract).

[65] Zhang X, Li W, Zheng ZM, Guo BH. Optimized statistical analysis of software trustworthiness attributes. Science China Information Sciences, 2012, 55(11): 2508–2520. [doi: 10.1007/s11432-012-4646-z]

[66] Ding XL, Wang HM, Wang YY, Lu G. Verification oriented trustworthiness evidence and trustworthiness evaluation of software. Journal of Frontiers of Computer Science & Technology, 2010, 4(1): 46–53 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.2010.01.005]

[67] Wang DX, Wang Q, He J. Evidence-based software process trustworthiness model and evaluation method. Ruan Jian Xue Bao/Journal of Software, 2017, 28(7): 1713–1731 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5102.htm [doi: 10.13328/j.cnki.jos.005102]

[68] Tao HW, Chen YX, Wu HY. A reallocation approach for software trustworthiness based on trustworthy attributes. Mathematics, 2020, 8(1): 14. [doi: 10.3390/math8010014]

[69] Tao HW, Wu HY, Chen YX. An approach of trustworthy measurement allocation based on sub-attributes of software. Mathematics, 2019, 7(3): 237. [doi: 10.3390/math7030237]

[70] Tao HW, Zhao J. Source codes oriented software trustworthiness measure based on validation. Mathematical Problems in Engineering, 2018, 2018: 6982821.

[71] Minsky M, Papert S. Artificial Intelligence Progress Report. Cambridge: MIT Press, 1972.

[72] A man killed by a robot welding car parts at factory in India. 2015. https://www.ishn.com/articles/102997-a-man-killed-by-a-robot-welding-car-parts-at-a-factory-in-india

[73] Hacker conference shows Tesla software attack technology: Collision accident will be induced. 2016 (in Chinese). https://www.youxia.org/2016/08/23441.html

[74] NTSB. Accident investigations: Car with automated vehicle controls crashes into pedestrian. https://www.ntsb.gov/investigations/Pages/HWY18FH010.aspx

[75] Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. Obstetrical & Gynecological Survey, 2021, 76(1): 5–7. [doi: 10.1097/01.ogx.0000725672.30764.f7]

[76] Machkovech S. Amazon confirms that Echo device secretly shared user's private audio. 2018. https://arstechnica.com/gadgets/2018/05/amazon-confirms-that-echo-device-secretly-shared-users-private-audio/

[77] OECD Legal Instruments. Recommendation of the council on artificial intelligence. 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[78] NIST. RFI on standards for reliable, robust, and trustworthy artificial intelligence. 2019. https://www.nist.gov/system/files/documents/2019/06/06/nist-ai-rfi-cset-001.pdf

[79] IBM. Trusted AI. 2019. https://www.research.ibm.com/artificial-intelligence/trusted-ai

[80] Singh R, Vatsa M, Ratha N. Trustworthy AI. In: Proc. of the 8th ACM IKDD CODS and the 26th COMAD. Bangalore: ACM, 2021. 449–453. [doi: 10.1145/3430984.3431966]

[81] Fujii G, Hamada K, Ishikawa F, Masuda S, Matsuya M, Myojin T, Nishi Y, Ogawa H, Toku T. Guidelines for quality assurance of machine learning-based artificial intelligence. Int'l Journal of Software Engineering and Knowledge Engineering, 2020, 30(11–12): 1589–1606. [doi: 10.1142/S0218194020400227]

[82] Chatila R, Dignum V, Fisher M, Giannotti F, Morik K, Russell S, Yeung K. Trustworthy AI. In: Braunschweig B, Ghallab M, eds. Reflections on Artificial Intelligence for Humanity. Cham: Springer, 2021. 13–39. [doi: 10.1007/978-3-030-69128-8_2]

[83] Ashoori M, Weisz JD. In AI we trust? Factors that influence trustworthiness of ai-infused decision-making processes. arXiv:1912.02675, 2019.

[84] Ge L, Gao J, Li XY, Zhang AD. Multi-source deep learning for information trustworthiness estimation. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Illinois: ACM, 2013. 766–774. [doi: 10.1145/2487575.2487612]

[85] Tabibian B, Valera I, Farajtabar M, Song L, Schölkopf B, Gomez-Rodriguez M. Distilling information reliability and source trustworthiness from digital traces. In: Proc. of the 26th Int'l Conf. on World Wide Web. Perth: Int'l World Wide Web Conf. Steering Committee, 2017. 847–855. [doi: 10.1145/3038912.3052672]

[86] Fogliaroni P, D'Antonio F, Clementini E. Data trustworthiness and user reputation as indicators of VGI quality. Geo-spatial Information Science, 2018, 21(3): 213–233. [doi: 10.1080/10095020.2018.1496556]

[87] Zhang Y, Ives ZG, Roth D. Evidence-based trustworthiness. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 413–423.

[88] 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]

[89] Liu ZQ, Ma JF, Weng J, Huang FR, Wu YD, Wei LF, Li YX. LPPTE: A lightweight privacy-preserving trust evaluation scheme for facilitating distributed data fusion in cooperative vehicular safety applications. Information Fusion, 2021, 73: 144–156. [doi: 10.1016/j.inffus.2021.03.003]

[90] Xu ZQ, Yang WD, Xiong ZG, Wang JY, Liu G. TPSense: A framework for event-reports trustworthiness evaluation in privacy-preserving vehicular crowdsensing systems. Journal of Signal Processing Systems, 2021, 93(2): 209–219. [doi: 10.1007/s11265-020-01559-6]

[91] Tsou YT, Chen HL, Chen JY. RoD: Evaluating the risk of data disclosure using noise estimation for differential privacy. IEEE Trans. on Big Data, 2021, 7(1): 214–226. [doi: 10.1109/TBDATA.2019.2916108]

[92] Ardagna CA, Asal R, Damiani E, El Ioini N, Pahl C. Trustworthy IoT: An evidence collection approach based on smart contracts. In: Proc. of the 2019 IEEE Int'l Conf. on Services Computing (SCC). Milan: IEEE, 2019. 46–50. [doi: 10.1109/SCC.2019.00020]

[93] Zhang QN, Ding QY, Zhu JM, Li DD. Blockchain empowered reliable federated learning by worker selection: A trustworthy reputation evaluation method. In: Proc. of the 2021 IEEE Wireless Communications and Networking Conf. Workshops (WCNCW). Nanjing: IEEE, 2021. 1–6. [doi: 10.1109/WCNCW49093.2021.9420026]

[94] Distefano S, Di Giacomo A, Mazzara M. Trustworthiness for transportation ecosystems: The blockchain vehicle information system. IEEE Trans. on Intelligent Transportation Systems, 2021, 22(4): 2013–2022. [doi: 10.1109/TITS.2021.3054996]

[95] Bau D, Zhou BL, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3319–3327. [doi: 10.1109/CVPR.2017.354]

[96] Slack D, Friedler SA, Scheidegger C, Roy CD. Assessing the local interpretability of machine learning models. arXiv:1902.03501, 2019.

[97] Sanneman L, Shah JA. A situation awareness-based framework for design and evaluation of explainable AI. In: Proc. of the 2nd Int'l Workshop on Explainable, Transparent Autonomous Agents and Multi-agent Systems. Auckland: Springer, 2020. 94–110. [doi: 10.1007/978-3-030-51924-7_6]

[98] Rosenfeld A. Better metrics for evaluating explainable artificial intelligence. In: Proc. of the 20th Int'l Conf. on Autonomous Agents and Multiagent Systems. London: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2021. 45–50.

[99] Lin YS, Lee WC, Celik ZB. What do you see?: Evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors. In: Proc. of the 27th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining. Singapore: ACM, 2021. 1027–1035. [doi: 10.1145/3447548.3467213]

[100] Ruan WJ, Wu M, Sun YC, Huang XW, Kroening D, Kwiatkowska M. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 5944–5952.

[101] Yu FX, Qin ZW, Liu CC, Zhao L, Wang YZ, Chen X. Interpreting and evaluating neural network robustness. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 4199–4205.

[102] Zheng XQ, Zeng JH, Zhou Y, Hsieh CJ, Cheng MH, Huang XJ. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 6600–6610. [doi: 10.18653/v1/2020.acl-main.590]

[103] Choi JH, Zhang H, Kim JH, Hsieh CJ, Lee JS. Evaluating robustness of deep image super-resolution against adversarial attacks. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 303–311. [doi: 10.1109/ICCV.2019.00039]

[104] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proc. of the 37th Int'l Conf. on Machine Learning. Vienna: JMLR, 2020. 206.

[105] 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]

[106] Song LW, Mittal P. Systematic evaluation of privacy risks of machine learning models. In: Proc. of the 30th USENIX Security Symp. USENIX Association, 2021. 2615–2632.

[107] Ma R, Li JQ, Xing BH, Zhao YY, Liu YW, Yan C, Yin H. A novel similar player clustering method with privacy preservation for sport performance evaluation in cloud. IEEE Access, 2021, 9: 37255–37261. [doi: 10.1109/ACCESS.2021.3062735]

[108] Yang ZQ. Fidelity: A property of deep neural networks to measure the trustworthiness of prediction results. In: Proc. of the 2019 ACM Asia Conf. on Computer and Communications Security. Auckland: ACM, 2019. 676–678. [doi: 10.1145/3321705.3331005]

[109] de Bie K, Lucic A, Haned H. To trust or not to trust a regressor: Estimating and explaining trustworthiness of regression predictions. arXiv:2104.06982, 2021.

[110] Zhou ZH. Machine Learning. Beijing: Tsinghua University Press, 2016. 23–51 (in Chinese).
[111] Spackman KA. Signal detection theory: Valuable tools for evaluating inductive learning. In: Proc. of the 6th Int'l Workshop on Machine Learning. New York: Morgan Kaufmann Publishers Inc., 1989. 160–163. [doi: 10.1016/B978-1-55860-036-2.50047-3]
[112] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1997, 30(7): 1145–1159. [doi: 10.1016/S0031-3203(96)00142-2]
[113] Jha S, Raj S, Fernandes SL, Jha SK, Jha S, Jalaian B, Verma G, Swami A. Attribution-based confidence metric for deep neural networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 11837–11848.
[114] van der Waa J, Schoonderwoerd T, Van Diggelen J, Neerincx M. Interpretable confidence measures for decision support systems. Int'l Journal of Human-computer Studies, 2020, 144: 102493. [doi: 10.1016/j.ijhcs.2020.102493]
[115] Shepperd MJ, Ince DC. Software metrics in software engineering and artificial intelligence. Int'l Journal of Software Engineering and Knowledge Engineering, 1991, 1(4): 463–476. [doi: 10.1142/S0218194091000305]
[116] Nakamichi K, Ohashi K, Namba I, Yamamoto R, Aoyama M, Joeckel L, Siebert J, Heidrich J. Requirements-driven method to determine quality characteristics and measurements for machine learning software and its evaluation. In: Proc. of the 28th IEEE Int'l Requirements Engineering Conf. (RE). Zurich: IEEE, 2020. 260–270. [doi: 10.1109/RE48521.2020.00036]
[117] Boehm BW. Software engineering economics. IEEE Trans. on Software Engineering, 1984, 10(1): 4–21.
[118] Cheng MX, Nazarian S, Bogdan P. There is hope after all: Quantifying opinion and trustworthiness in neural networks. Frontiers in Artificial Intelligence, 2020, 3: 54. [doi: 10.3389/frai.2020.00054]
[119] Uslu S, Kaur D, Rivera SJ, Durresi A, Durresi M, Babbar-Sebens M. Trustworthy acceptance: A new metric for trustworthy artificial intelligence used in decision making in food-energy-water sectors. In: Proc. of the 35th Int'l Conf. on Advanced Information Networking and Applications. Toronto: Springer, 2021. 208–219. [doi: 10.1007/978-3-030-75100-5_19]
[120] Chowdhury A, Karmakar G, Kamruzzaman J, Islam S. Trustworthiness of self-driving vehicles for intelligent transportation systems in industry applications. IEEE Trans. on Industrial Informatics, 2021, 17(2): 961–970. [doi: 10.1109/TII.2020.2987431]
[121] Chattopadhyay A, Ali A, Thaxton D. Assessing the alignment of social robots with trustworthy AI design guidelines: A preliminary research study. In: Proc. of the 11th ACM Conf. on Data and Application Security and Privacy. ACM, 2021. 325–327. [doi: 10.1145/3422337.3450325]

附中文参考文献:

[1] 方滨兴. 人工智能安全. 北京: 电子工业出版社, 2020. 1–10.
[4] 国务院关于印发新一代人工智能发展规划的通知 (国发[2017]35号). 2017. http://ai.cumt.edu.cn/info/1024/1052.htm
[5] 中国科学院科技战略咨询研究院, 中国科学院学部科学普及与教育研究支撑中心. 中国科技热点述评2019. 北京: 科学出版社, 2020. 89–137.
[11] 孙金云. 2020打车软件出行现状调研报告. 复旦大学管理学院—复旦青年创业家教育与研究发展中心. 2021. 3. 40–56. https://mp.weixin.qq.com/s/G2VzL9QJJU4Acsl8VvtGhw
[12] 何积丰. 安全可信人工智能. 信息安全与通信保密, 2019(10): 5–8.
[13] 可信软件基础研究项目组. 可信软件基础研究. 杭州: 浙江大学出版社, 2018. 1–26.
[14] 刘克, 单志广, 王戟, 何积丰, 张兆田, 秦玉文. "可信软件基础研究"重大研究计划综述. 中国科学基金, 2008, 22(3): 145–151. [doi: 10.3969/j.issn.1000-8217.2008.03.005]
[15] CCF形式化专业委员会. 人工智能系统的形式化验证技术研究进展与趋势. 见: 2019–2020中国计算机科学技术发展报告. 2020. 491–539.
[17] 陈仪香, 陶红伟. 软件可信性度量评估与增强规范. 北京: 科学出版社, 2019. 1–19.
[36] 国务院. 国务院关于积极推进"互联网+"行动的指导意见. 2015. http://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm
[37] 国务院. "十三五"国家科技创新规划. 2016. http://www.gov.cn/zhengce/content/2016-08/08/content_5098072.htm
[38] 工业和信息化部. 促进新一代人工智能产业发展三年行动计划(2018–2020). 2017. http://www.cac.gov.cn/2017-12/15/c_1122114496.htm
[39] 中国电子技术标准化研究院. 人工智能标准化白皮书(2018版). 北京: 国家标准化管理委员会, 2018.
[40] 金钟河, 著; 叶蕾蕾, 译. 致命Bug. 北京: 人民邮电出版社, 2016. 1–7.
[47] 陈火旺, 王戟, 董威. 高可信软件工程技术. 电子学报, 2003, 31(S1): 1933–1938. [doi: 10.3321/j.issn:0372-2112.2003.z1.001]
[53] 国务院. 国家中长期科学和技术发展规划纲要. 2006. http://www.gov.cn/jrzg/2006-02/09/content_183787.htm

[54] 科技部. 高可信软件生产及集成环境项目(863计划)申请. 2009. http://www.edu.cn/ke_yan_yu_fa_zhan/gai_kuang/xin_wen_gong_gao/200904/t20090413_372055.shtml

[61] 杨善林, 丁帅, 褚伟. 一种基于效用和证据理论的可信软件评估方法. 计算机研究与发展, 2009, 46(7): 1152–1159.

[63] 郑志明, 马世龙, 李未, 韦卫, 姜鑫, 张占利, 郭炳晖. 软件可信性动力学特征及其演化复杂性. 中国科学F辑: 信息科学, 2009, 39(9): 946–950.

[64] 郑志明, 马世龙, 李未, 姜鑫, 韦卫, 马丽丽, 唐绍婷. 软件可信复杂性及其动力学统计分析方法. 中国科学F辑: 信息科学, 2009, 39(10): 1050–1054.

[66] 丁学雷, 王怀民, 王元元, 卢刚. 面向验证的软件可信证据与可信评估. 计算机科学与探索, 2010, 4(1): 46–53. [doi: 10.3778/j.issn.1673-9418.2010.01.005]

[67] 王德鑫, 王青, 贺劼. 基于证据的软件过程可信度模型及评估方法. 软件学报, 2017, 28(7): 1713–1731. http://www.jos.org.cn/1000-9825/5102.htm [doi: 10.13328/j.cnki.jos.005102]

[73] 黑客大会展示特斯拉软件攻击技术: 将诱发碰撞事故. 2016. https://www.youxia.org/2016/08/23441.html

[88] 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]

[105] 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]

[110] 周志华. 机器学习. 北京: 清华大学出版社, 2016. 23–51.

刘晗(1997－), 男, 博士生, CCF 学生会员, 主要研究领域为软件可信性度量.

陈仪香(1961－), 男, 博士, 教授, CCF 杰出会员, 主要研究领域为软件可信性度量, 软件和硬件协同设计, 物联网, 实时协同规范语言设计.

李凯旋(1997－), 男, 博士生, CCF 学生会员, 主要研究领域为软件可信性度量, 软硬件协同设计.