可信AI的前世今生

陶大程

引用本文:
陶大程. 可信AI的前世今生[J]. 智能系统学报, 2021, 16(4): 0–0.
$stringUtils.citationAuthorFormate(${article.authorEnNames},", ",", et al"). [J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 0–0.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202107047

陶大程，京东集团探索研究院首任院长，澳大利亚科学院院士，ACM/AAAS/IEEE Fellow，主要研究方向为人工智能与可信人工智能。在相关领域发表学术论文 200 余篇，并多次获得顶级国际会议和刊物的最佳论文奖与时间验证奖。2021 年荣获 IEEE Computer Society Edward J McCluskey 技术成就奖，2018 年荣获 IEEE ICDM 研究贡献奖，2015 年和 2020 年两度荣获澳大利亚尤里卡奖，2015 年荣获悉尼科技大学校长奖章，2020 年荣获悉尼大学校长杰出研究贡献奖。

卷首语 Foreword

可信 AI 的前世今生

陶大程

人工智能并不是一个新的概念，从 1950 年的图灵之问开始，到今天产业的蓬勃发展。随着人工智能广泛的产业落地带来的诸多问题，AI 面临越来越多的可信挑战，例如 AI 系统的不确定性导致潜在的安全问题；可解释性的缺乏限制了 AI 更广泛的应用与赋能；AI 系统如何在使用数据的同时保护用户隐私等。构筑可信 AI 已成为全球共识，2016 年欧盟颁布了《通用数据保护条例（GDPR）》；2017 年 12 月，IEEE 提出了《人工智能的伦理设计准则》，之后澳洲、美国、新加坡等都提出了相关的政策、指南或白皮书。国内，何积丰院士于 2017 年 11 月香山科学会议上首次提出了“可信人工智能”的概念。2017 年 12 月，工业和信息化部发布了《促进新一代人工智能产业发展三年行动计划》。在此之后，中国的科技公司纷纷提出了可信人工智能发展规划。2019 年 10 月，京东集团就首次在乌镇世界互联网大会上提出京东践行“可信赖 AI”的六大维度；京东探索研究院在 2021 年 4 月已将“可信人工智能”正式列为主要研究方向之一，并于同年 7 月联合中国信通院完成撰写、发布国内首本《可信人工智能白皮书》。

可信 AI 的研究涉及方方面面，为实现可信 AI，首要任务是找到合适的方法来定量分析、量化人工智能算法、模型、系统的稳定性、可解释性、隐私保护能力及公平性。如果人工智能在以上四个方面的“可信”度量上都达到很高的共识水平，就有更好的机会做到明确责任、透明可信，并且推动人工智能在产业的进一步落地。

可信 AI- 稳定性：人工智能系统抵抗恶意攻击或者环境噪声并且能够做出正确决策的能力。高性能 AI 系统，能在保障用户安全的同时更好地服务社会。可以通过攻击算法的攻击成功率等方式来度量稳定性能。现有稳定性技术的提升方法包括对抗训练、样本检测等。

可信 AI- 可解释性：指人工智能系统所做出的决策需要让人能够理解。可解释性的提升，不仅有助于构建更高性能的 AI 系统，更能促进 AI 技术在更广泛行业的赋能与落地。可解释性的度量除了模型可解释性之外，还包含训练样本的可解释性、测试样本的可解释性等。其研究内容包括有效性分析、样本检测、显著性分析等方面。

可信 AI- 隐私保护：指人工智能系统不能把个人的隐私信息或者群体的隐私信息进行泄露。AI 系统在为用户提供精准服务的同时，也要注重保护用户的隐私。隐私保护非常重要，衡量隐私保护可以使用差分隐私、隐私攻击等多种方式。我们可以通过联邦学习、多方计算、同态加密等手段来提升用户的隐私保护。

可信 AI- 公平性：指人工智能系统需要公平地对待所有用户。公平的 AI 系统能包容人与人之间的差异，为不同的用户提供相同质量的服务。目前可以使用个体公平性和群体公平性等进行公平性度量。公平性保障算法包括预处理方法、处理中方法、后处理方法等。

关于可信 AI 稳定性、可解释性、隐私保护能力及公平性的度量及技术的提升还有待进一步研究。此外，可信 AI 研究的方方面面亦互相联系，并不孤立，因此还需要从整体出发来开展可信 AI 研究。要想实现最终的可信人工智能，需要找到统一的综合治理框架，即要构建可信 AI 的一体化理论，然后来帮助我们实现有效的可信治理。可信 AI 一体化研究将是未来的重要趋势。可信能力评测将是未来人工智能产业落地中非常重要的一个环节，从理论和实践的层面持续开展可信 AI 研究，将推动人工智能产业进入一个新的浪潮。