可信人工智能标准体系建设研究

程方正　彭飞荣

（中国计量大学）

摘　要：人工智能的发展和进一步应用，已经成为不可逆的趋势。人工智能带来福利的同时，也带来一定的风险。通过可信人工智能标准体系构建的方法，可降低人工智能所带来的风险。可信人工智能要求人工智能系统具备可解释性、无歧视性和数据安全性。域外可信人工智能标准体系建设主要存在伦理概念标准建设和技术标准建设两种途径，并已经有相关的可信人工智能标准体系建设总纲。我国可以参照域外的经验，建设可信人工智能标准体系总纲，加强人工智能伦理标准体系建设和技术标准体系建设，确保人工智能系统的可信性。

关键词：可信人工智能，标准体系建设，人工智能伦理

DOI编码：10.3969/j.issn.1674–5698.2023.09.004

Research on the Construction of Trusted AI Standards System

CHENG Fang–zheng　PENG Fei–rong

（China Jiliang University）

Abstract: The development and further application of artificial intelligence (AI) has become an irreversible trend. While AI brings benefits, it also brings certain risks. The risk brought by AI can be reduced through the method of building a trusted AI standards system. Trusted AI requires that artificial intelligence systems have explainability, non-discrimination, and data security. There are mainly two ways to build an extraterritorial trusted AI standards system: the construction of ethical concept standards and the construction of technical standards, and there is already a relevant general outline for the construction of a trusted AI standards system. China can refer to the overseas experience to build a general outline of a trusted AI standards system, strengthen the construction of an AI ethical standards system and a technical standards system, and ensure the credibility of AI systems.

Keywords: trusted AI, standards system construction, AI ethics

1　可信人工智能标准体系建设的必要性

人工智能的进一步发展，已经成为了不可逆的趋势。近年来，欧洲、美国、日本等国家及地区持续加大投入关于人工智能基础理论和应用的研究，以此保持在此领域的技术领先地位。我国政府也在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。根据中国互联网协会所发布的《中国互联网发展报告（2021）》显示，我国的人工智能产业，市场规模已经达到3031亿元。

人工智能的发展虽促进了社会生产力、经济的

作者简介：程方正，硕士研究生，研究方向为知识产权法与网络法治。

彭飞荣，法学博士，副教授，国家知识产权培训（浙江）基地副主任，研究方向为知识产权法、标准化法。

发展，但也产生了新的问题。近年来，例如：特斯拉的自动驾驶事故、数据泄漏等事件的发生，社会各方开始重点关注人工智能系统的安全性问题。习近平总书记在2018年10月主持政治局第九次集体学习时就强调，“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控”。随后，我国在G20峰会上，首次提出可信人工智能概念，强调发展以人为本的可信人工智能，并得到国际社会的普遍认同[1]。

将可信人工智能的要求落实到具体的实践中去，是目前人工智能产业发展亟需要解决的问题。2021年7月，中国信息通信研究院和京东探索研究院联合制作了《可信人工智能白皮书》，分析并发掘如何实现可信人工智能的路径。《白皮书》指出，“可信人工智能不仅仅是企业单方面的实践和努力，需要多方协调参与，形成一个相互影响、相互支持、相互依赖的良性生态。这个生态主要包括标准体系、评估验证、合作交流。”[2]诚如《白皮书》中所言，政策和法律只能规定原则，具体的落地措施还需要依靠标准来实现。我国目前虽然已经有关于可信人工智能的相关标准，如：2021年4月出台的《信息安全技术人脸识别数据安全要求》和2021年9月25日出台的《新一代人工智能伦理规范》，以及国家标准化管理委员会等五部门在2020年7月27日所印发的《国家新一代人工智能标准体系建设指南》。上述的文件都有回应有关可信人工智能建设的问题，但其内容多半为原则性概念，且内容较为宏观，难以称为一套完整的标准体系。本文将通过论述可信人工智能标准的基本要求，并参照域外经验后，为我国可信人工智能标准体系的建设提供一些建议。

2　可信人工智能标准体系的基本要求

可信人工智能标准体系的设立目的为将可信人工智能伦理准则规范化、实践化，因此，在构建可信人工智能标准之前，需了解可信人工智能的基本要求。目前，人工智能的伦理原则的内容并不存在统一的标准，但可解释性、公平与偏见、安全和隐私和可问责性这4个方面为业界和学界较为公认的4个重要原则[3]。由于可问责性原则的实现，主要依赖于法律、行政法规等规范性文件，与标准内容并无过大的关联，因此，本文将不做详细的介绍。下文将从可解释性、公平与偏见和安全与隐私这3个维度展开，进一步介绍人工智能的基本要求。

2.1 可解释性

在过去的几年，人工智能的不透明，也就是所谓的“算法黑箱”已经成为热点问题。透明度为数据保护的一般原则，但人工智能由于其特性的限制，使得透明度的具体标准难以被制定。

人工智能存在技术壁垒、因果关系难以解释和高度动态性的问题，使得外部观察者对于系统本身造成认知约束。因此，有关人工智能透明度的问题，不应当侧重于获取有关人工智能系统尽可能详细的信息，而是侧重于人工智能系统的可解释性，故应当引入可解释概念，解决此类问题。有学者也曾提出过透明性（transparency）的概念，即要求产品方将人工智能的算法进行公开或者部分公开，以此解决“算法黑箱”的问题。但对于人工智能产品来说，采用可解释性的要求明显优于透明性。这是因为，就像对于人类有关行为的解释，并不需要了解其体内的神经元信号如何流通一样，对于人工智能有关行动的解释，也不必去了解人工智能系统的比特流。

人工智能的可解释性的概念与内涵，取决于不同领域和不同利益相关者，目前主要基于3个角度，即方法层面、用户需求角度和应用研究方面。方法层面主要侧重于算法模型的解释。用户需求层面，则期望AI决策能够透明化，熟悉其决策规则。AI决策在应用层面研究较为广泛，也较为深远，例如：智能医疗、无人驾驶甚至智能司法。因此，根据人工智能可解释性的要求，产品方应当提高有关算法的功能和缺陷的解释、AI决策化的流程图和决策规则以及所销售人工智能产品的潜在风险（与普通同种类产品有所不同的风险）。

2.2 无歧视性

无歧视的社会是我们向往追求的，歧视往往来源于主观偏见，并依据主观偏见做出了带有歧视性的决策。机器依据纯粹的事实和设定的规则

做出决策，因此，机器所作出的决策应当避免歧视性结果的发生。然而，事实并非如此，大量的证据证明，人工智能系统所作出的决策，仍可能具有歧视性。

人工智能系统的歧视性来源，可以分为3类，分别为有缺陷的数据收集、有缺陷的数据聚合和规范的无响应[4]。缺陷的数据收集是导致人工智能系统生成歧视性结果的一种常见原因，即数据输入偏差导致数据输出偏差。输入偏差的原因通常是因为数据的代表性不足。有缺陷的数据聚合是指原始的训练数据不存在偏差，但在之后的数据处理过程之中发生了偏差。此类问题的产生原因主要为数据标准前后不一致，从而影响后续人工智能的决策行为。

相较于有缺陷的数据收集、有缺陷的数据聚合，规范的无响应性的问题较为复杂且更加难以解决。规范无响应性是指，即使通过相应的规范构建防止此类歧视行为的发生，依旧可能发生歧视性的决策。人工智能的决策行为基于以往的统计数据所产生，也就是说，人工智能只能通过对过去事实的观察，以此预测未来的发生，但规范往往是“反事实”行为，目的在于调整现状，这二者间相互独立，因此在某些情况下可能会发生冲突。例如：根据数据统计表明，女性的平均寿命要高于男性，因此，保险公司有理由基于此数据，区别设置男性和女性的保险费。但这涉及性别歧视，违反公序良俗原则，因此需设立相关规定，禁止将性别作为决策的参考因素。然而，这时就可能发生规范的无响应性，即使设立相关规定，禁止保险公司将性别作为决策的参考因素，其性别歧视的行为依旧存在。这是因为人工智能系统即使不将性别作为决策的参考因素，但也可以通过其他数据的测量和评算，预测出具体客户的“性别”，并以此做出决策，从而使得“性别平等”的规范在实践中无效化。

人工智能的歧视治理并不能采用单一化的方式，应当通过多个层面进行治理，例如：通过“检查训练数据的公平性、检查模型输出的公平性和填写数据卡和模型卡”[5]的方式，并制定出相应的技术标准，提前预防歧视性行为的发生。

2.3 数据安全性

ChatGPT的概念刚产生不久，有关的人工智能的数据安全和网络安全问题就已经浮出水面，根据Fortinet北亚区首席技术顾问谭杰表示，ChatGPT等AI技术对网络和数据安全的威胁已经发生[6]。

目前，人工智能数据安全风险主要来自于数据投毒、数据深度伪造、数据过度采集、数据滥用分析等方面[7]。随着人工智能在各个领域的发展，数据资源得到更广泛的使用，但数据风险安全问题也愈发突出。数据安全问题不仅涉及到用户的隐私权益，也涉及到行业的运行规制，甚至关乎国家的安全问题。我国关于数据安全问题已经出台了一系列法律法规，例如：《个人信息保护法》《数据安全法》，但人工智能技术具有一定的特殊性，需要制定特别的规范，以此确保其数据安全性。

数据安全的问题应当从两种路径进行解决，第一种为制定相关的技术标准，从技术层面化解数据安全风险；第二种为制定相应的规范性文件，落实相关人员的责任，合理弥补受害者的损失。前一种规范为事前规范，可以最大程度减少数据安全风险，使得社会资源得到最大的利用。后一种规范为事后规范，事后规范并不能有效预防危险的发生，但由于在现实层面中，不可能从技术方面完全杜绝数据安全风险的发生。通过事后规范，可以从最大程度上填平损失，使得人工智能产业稳定、健康发展。

3 可信人工智能体系标准建设的域外情况

国外的人工智能技术和产业的发展起步较早，并在人工智能伦理和可行性方面已经有了较为丰富的研究，对于我国可信人工智能标准体系建设具有重要参考价值。

国际上有关人工智能标准建设的研究和发展，大致遵循两种路径，第一种路径为侧重建立人工智能伦理基础原则（以下简称“伦理原则路径”），例如：ISO/IEC JTC 1/SC 42/WG 3、IEC/SEG 10、IEEE-GIEAIS。另一种路径则试图通过对人工智

能技术的规制（以下简称“技术路径”），从而确保人工智能的可信性，此种路径的标准代表有ISO/IEC JTC1/SC42/WG3、IEEE P7000™[8]。

伦理原则路径的标准体系建设，主要通过提出一系列的概念，并要求人工智能产品必须具备各个概念中的相应标准。例如：SC42/WG3提出问责制、责任性、可解性、鲁棒性等一系列概念，总结出人权、劳动实践、自然环境等一系列影响核心社会责任议题。侧重于人工智能伦理基础原则的标准体系建设，其功能在于，为可信人工智能标准体系建设的发展方向提供具体要求，使得人工智能的可信度可以被评价。

人工智能可信性标准的落地，需要依赖具体的技术得以实现，只依靠评价体系建设，并不能从根源上解决人工智能可信性的问题，因此，需要从技术角度设立相关的标准。然而，由于人工智能相关技术存在复杂性和多样性的特点，国际标准化组织在该领域的技术类标准的研究，均处于研究阶段，并未完全成型。但也产生一些成果，对于我国的相关领域的建设，具有参考价值，例如：上文所列的IEEE P7000™。

建设伦理原则的路径和建设技术标准的路径并非相互矛盾，恰恰相反，二者互为手段和目的。因此，欧盟的相关委员会尝试将两种路径进行统筹，构建出一套系统性的人工智能可信性指南。2018年12月，欧盟委员会的人工智能高级专家组发布了《可信人工智能伦理指南草案》，该指南提出一个可信人工智能框架，强调伦理规范性和技术性，并提出总计10项可信人工智能的要求和12项用于实现可信人工智能的技术和非技术性方法，同时设计出一套评估清单，便于企业和监管方进行对照。随后，ISO（国际标准化组织）和IEC（国际电工委员会）在2020年出台ISO/IEC TR 24028《可信人工智能标准概述》，其目的在于分析可能影响人工智能系统可信度的一些关键因素，以此协助标准界确定AI领域的具体标准化差距。上述指南的提出，标志着可信人工智能的标准建设已经进入到结构性、体系性阶段。

4　我国可信人工智能标准体系建设的建议

我国已有大量关于人工智能技术的标准，但是缺乏关于人工智能可信性的相关标准。2020年国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部印发的《国家新一代人工智能标准体系建设指南》中第八项，提到了人工智能的安全/伦理标准，并在文中提出，需从8个方向进行标准体系的建设，但未提出具体的标准构建。现行的GB/T 41867–2022标准中，提出了有关人工智能伦理的相关概念，除此之外，并无任何与可信人工智能相关的标准。本文将结合人工智能的基本原则和部分域外标准规范，提出几点关于构建可信人工智能标准的建议。

4.1 设立人工智能可信性指南，构建可信人工智能标准框架

人工智能进一步的发展可能受制于用户群体对人工智能系统可靠性、有效性和公平性的质疑，若不能系统性解决社会层面对于人工智能的担忧，则会削弱市场对于人工智能的投入，从而影响人工智能产业的长远发展。

我国的可信人工智能标准体系，并不是将可信人工智能相关标准机械相加。标准体系是一定范围内的标准按其内在联系形成的科学有机整体，具有目的性、层次性。可信人工智能标准体系内的标准应当相互协调、相辅相成，绝不是孤立的、彼此之间毫无关联的。人工智能可信性指南的设立，是将关于可信人工智能标准规范有机结合，构建可信人工智能标准体系框架的基础条件。

本文认为，可以参考域外的经验，即参照ISO/IEC TR 24028标准，设立符合我国国情的可信人工智能标准指南。指南的内容不宜设立得过于具体，但要将人工智能系统可能存在的风险和现存的评价体系，尽可能罗列出来，例如：ISO/IEC TR 24028中第八节的内容，将人工智能可能存在的威胁性风险全部罗列出来，并在第九章中罗列出相关的解决途径。

4.2 建设人工智能伦理原则标准，设立人工智能风

险分类分级制度

目前，国际上许多关于人工智能伦理的研究活动侧重于建立人工智能伦理基础原则，形成非技术性指导文件。我国即将推行的GB/T 5271.31–2006标准包含描述人工智能可信性的概念，但相比于国外的部分标准，例如：与SC42/WG3中所提出的概念相比，还是显得有些不足。

我国也可以效仿IEC/SEG 10标准体系，建立人工智能系统评级和人工智能使用等级标准。在IEC/SEG 10标准体系中，通过透明性、鲁棒性等特征，将人工智能系统进行分级分类，并从控制程度、纠正程度进行使用等级分类。人工智能目前仍处于发展阶段，若实施较为严格的标准，不利于人工智能产业的发展。况且不同种类的人工智能系统，所带来的风险程度和风险可能性都不相同。目前，我国并未有系统性、全面性和可操作性的人工智能系统分级分类标准，因此需加快此类标准建设进度。

4.3 加强可信人工智能技术标准建设，构建场景化人工智能技术标准

基于技术标准的建设路径，是实现可信人工智能落地化的必然途径。可信人工智能的实现，仅凭概念化的标准建设，例如：提出鲁棒性、可解释性的概念，是完全不够的。人工智能系统应当属于产品的一种，而产品的完全性、可信性，关键在于技术手段是否应用合理。只有通过对产品的技术手段进行规制，才能从源头上保障产品的质量安全。而设立人工智能技术标准，就是一种较为科学的规制方式。

目前，人工智能产业发展迅速，不同领域内所应用的技术截然不同。若标准要约束和指导人工智能可信化的落地和实施，则必须进一步类型化、具体化和场景化。我国可信人工智能的技术标准目前集中于产品端，例如：智能工厂、智能制造，但缺少有关人工智能技术层面的规制，也缺少具体化和场景化的技术标准。2021年4月出台的《信息安全技术人脸识别数据安全要求》虽然弥补了这一方面的缺失，但距离可信人工智能标准体系的总目标，仍有很大的差距。

4.4 着重关注算法治理，规制机器学习技术

可信人工智能的治理关键在于算法规制，可信人工智能标准体系的基本要求，即可解释性、安全与隐私和公平与偏见都与人工智能系统的应用算法有关。人工智能系统的安全与隐私问题、公平与偏见问题，可能来源于系统训练时采用的数据存在瑕疵，也可能来源于对数据初次筛选时具有偏见，但最主要也是最复杂的源头，就是算法应用存在不合理之处。

机器学习是人工智能所采用的主要手段，我国尚未设立机器学习的相关技术标准。目前，ISO/IEC 23053–2022和ISO/IEC TS 4213–2022两个标准中，已经着手机器学习的性能和相关技术评估，在此方面，我国可以借鉴上述两项标准，构建符合我国国情的机器学习相关技术标准。

参考文献

[1] 何积丰. 安全可信人工智能[J]. 信息安全与通信保密, 2019(10):5–8.

[2] 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[R/OL], [2023.4.14],http://www.caict.ac.cn/kxyj/qwfb/bps/202107/P020210709319866413974.pdf .

[3] 孔祥维,王子明,王明征, 等. 人工智能使能系统的可信决策：进展与挑战[J]. 管理工程学报, 2022,36(06):1–14.

[4] 托马斯·威斯迈德,编. 人工智能与法律的对话[M].韩旭至,译. 上海: 上海人民出版社, 2020:116.

[5] 朱悦. AI如何理解AI治理：一名研究者与ChatGPT的问答[EB/OL].[2023.4.14]. https://redian.news/wxnews/166774.

[6] 樊雪寒. ChatGPT威胁数据安全是杞人忧天吗? 企业闻风而动[N]. 第一财经日报, 2023–2–26.

[7] 林伟. 人工智能数据安全风险及应对[J]. 情报杂志, 2022,41(10):105–111+88.

[8] 孙宁. 人工智能伦理与社会关注国际标准研究[J]. 信息技术与标准化, 2020(Z1):27–29+34.