Interdisciplinary Science Letters 交叉科学快报, 2025, 9(2), 157-161
Published Online March 2025 in Hans. https://www.hanspub.org/journal/isl
https://doi.org/10.12677/isl.2025.92021

Hans汉斯

论算法黑箱的安全伦理风险及其规制路径

邵凡翔，许胜晴

江苏海洋大学法学系，江苏 连云港

收稿日期：2025年2月18日；录用日期：2025年3月11日；发布日期：2025年3月20日

摘　要

人工智能时代算法是重要的技术生产力基础，也是推动人工智能技术进步的重要因素。但是，因专业性、复杂性、不确定性以及政府或企业出于保密和知识产权需要而带来的算法黑箱问题也引发了诸多安全伦理风险。本文在分析算法黑箱基本概念和类型的基础上，分析其引发的主要安全伦理风险。论文认为算法黑箱主要涉及数据和隐私保护风险、算法偏见和滥用风险以及算法应用安全风险等。基于相关治理理论和实践需求，论文提出强化算法应用的数据和隐私保护、加强算法合理设计与应用的监管、完善算法应用安全的预防性机制等方面的建议。

关键词

算法黑箱，安全风险，伦理风险，规制路径

On the Safety and Ethic Risks of Algorithm Black Box and Regulation Paths

Fanxiang Shao, Shengqing Xu

Department of Law, Jiangsu Ocean University, Lianyungang Jiangsu

Received: Feb. 18th, 2025; accepted: Mar. 11th, 2025; published: Mar. 20th, 2025

Abstract

In the era of artificial intelligence, algorithm is an important technical productivity base and an important factor in promoting the progress of artificial intelligence technology. However, algorithmic black box due to professionalism, complexity, uncertainty, and the need for confidentiality and intellectual property rights by governments or enterprises also causes many security ethical risks. Based on the analysis of the basic concepts and types of algorithm black box, this paper analyzes the main security ethical risks caused by it. The paper considers that the algorithm black box mainly involves the risk of data and privacy protection, the risk of algorithm bias and abuse, and the risk of algorithm application security. Based on the relevant governance theory and practice requirements, this paper

文章引用: 邵凡翔, 许胜晴. 论算法黑箱的安全伦理风险及其规制路径[J]. 交叉科学快报, 2025, 9(2): 157-161.
DOI: 10.12677/isl.2025.92021

puts forward some suggestions on strengthening the data and privacy protection of algorithm application, strengthening the supervision of rational algorithm design and application, and improving the preventive mechanism of algorithm application security.

Keywords

Algorithmic Black Box, Security Risk, Ethical Risk, Regulatory Path

1. 引言

随着技术的进步和发展，人工智能越来越广泛地应用于人类的生产生活实践。人工智能技术日益深化，给人类带来的便利和效率也逐渐提升，但是与之相应的技术复杂度和不确定性问题也随之凸显。其中，人工智能所依赖的算法也因其高度的专业化和复杂性而引发算法黑箱问题。如何有效应对算法黑箱问题，在人工智能时代依然保持人的主体性、自主性和尊严成为愈发重要和现实的研究问题。算法黑箱带来的安全伦理问题是技术发展局限性的体现，也是技术设计和应用的人为性的结果。算法黑箱安全伦理风险治理亟需完善相应的治理政策法律，而明确相关风险则是建立有效规制路径的重要前提。

2. 算法黑箱概念和类型

算法从一般的计算方法到计算机领域的模型、指令、代码，经历了由简单到复杂、由数学到工程的不断演化发展。尽管在基本内涵上较为一致，但算法在不同领域有着不同的定义形式。算法在计算机科学领域主要为使用计算机执行计算或解决问题时的一系列指令[1]。黑箱是一种隐喻，是对主体而言的一种隐蔽式的存在或者活动。从控制论的角度而言，黑箱是相对于较为清晰的输入端和输出端而言的系统内部的构成、功能及其运行方式。这种内部形态和运行不为主体所掌握，也不能轻易获知。算法黑箱具有相对性，对其技术开发者而言，算法黑箱并非完全不可获知或理解，但对其使用者或者被使用者而言，则往往呈现黑箱形态[2]。人工智能自主学习一般基于归纳逻辑。通过对海量数据的归纳与学习，人工智能算法可以提取不同类型数据的共同规律，学习不同序列数据之间的关联概率，从而生成相应内容[3]。因算法形态的技术性和应用的社会性，算法具有符号代码、保密信息和权力表达三重属性。就存在形态而言，算法以符号代码的形式存在，对于普通公众而言其难以理解，从而形成认知上的壁垒。就保密信息而言，算法黑箱还涉及企业或政府出于政务信息保密或者商业秘密、知识产权等方面的原因不公开算法内容，从而形成信息透明度方面的算法黑箱。权利表达主要涉及政府决策和公共领域资源配置等问题，涉及相关算法的运行过程和相应的权力运作等[2]。

3. 算法黑箱的安全和伦理风险

算法黑箱是算法风险中的内生性风险，是引发其他风险如算法滥用、算法操纵、算法霸权和算法问责等问题基础[4]。基于算法黑箱涉及的安全和伦理风险所威胁的内容、作用的方式以及应用的领域，相关风险可基本划分为数据和隐私风险、算法偏见与滥用风险以及应用安全的风险。

3.1. 数据和隐私保护风险

算法是进行数据计算和处理程序的模型支撑，其运行需要以数据的输入为基础。数据与算法是驱动

社会发展数字化、智能化的关键。机器算法嵌入大数据可以高效提取数据价值、进行仿生模拟并进而在不同场景下作出行为感知和决策判断[5]。数据是算法应用的重要基础，而数据本身具有的多重属性则为算法应用带来潜在风险。数据涉及各类信息，具有商业属性、技术属性、法律属性、社会属性和个人属性等多元属性。算法应用过程中所获取的数据涉及用户相关人身、财产、著作以及各类生产生活活动等数据。在个人信息方面，相关数据还涉及个体生物特征、人格权、隐私权和财产权等事项。算法在通过数据实现其价值和功能的同时也带来了数据安全和隐私保护问题。主动或者被动应用算法将使得相关数据由保密、隐私或者其他非公开形式进入到算法程序，从而在数据占有主体或存储形态上产生了扩展，增加了相关数据的泄露风险[6]。例如近年来频繁出现相关企业工作人员秘密下载客户信息和其他主体信息牟利的案例，涉及经营中涉及的客户信息、商业秘密等。各类人工智能工具也成为数据的收集者，一旦泄露将危机个人安全甚至公共安全。除人为因素外，网络空间的各类风险也会引发机器学习模型本身、训练数据集甚至原始数据发生泄露。

3.2. 算法偏见与滥用风险

随着人工智能的发展，通过人工智能提供更加高效、便捷的公共服务成为相关技术应用的重要发展方向。人工智能及其算法为技术进步提供了助力，但也因其技术局限性和应用中的监管问题而对公共领域安全带来算法偏见和滥用现象。数据偏见是指因数据本身存在的偏差(如样本选择偏差、数据收集偏差等)导致的算法模型在处理数据时产生不公平结果的现象。在社会服务或者商业领域，算法应用是提升管理、运行和服务效率和精确性的技术性辅助措施，但其本质上仍属于机械式的运行和输出过程。算法的应用结果受到一系列主客观因素的影响。就客观方面而言，算法模型可能受到数据本身的不完整性、异常性甚至虚假数据的影响，而在模型运行或者输出方面产生歧视性问题。在主观方面，算法的设计者或者使用着自身处于经济或者其他动因可能对算法设计施加主观偏见或倾向性内容，从而干扰或者操控算法的输出结果使其具有歧视性或者选择性。例如，在医疗诊断中，近年来已出现大量智能医学诊疗算法，涉及疾病的预测、诊断以及治理等各个方面[7]。医疗算法的应用对疾病诊断和治疗带来新的技术支撑，但因算法基于技术和计算的判断与职业人员基于专业和经验的判断往往存在张力。这种张力既可以成为更加精准诊断和治疗疾病的助力，也可能成为进一步加剧医患纠纷、增加医疗风险的原因。与商业领域的知情权类似，医疗算法也会产生“算法利维坦”现象，影响患者的基本权利。例如，医疗数据样本本身的局限性、算法的内在设计缺陷以及算法设计者可能存在价值偏见等问题会影响患者得到合理、公平的诊疗服务[8]。因数据、技术因素的嵌入，交互式人机共决机制容易导致医疗责任认定陷入困境，且进一步增加医患关系的复杂性[9]。另外，算法带来的便利性和准确性也存在被滥用的风险。例如，在商业领域，算法的应用能够用于剥夺消费者剩余，产生价格歧视。实践中存在经营者运用算法刻画消费者习惯以及结合各类营销方法，使得新老用户在价格待遇方面形成显著差别，从而形成大数据杀熟情况。通过算法设计对老顾客采取高定价策略，并对新客户采取低价策略，进而吸引更多新客户，同时对习惯性老客户榨取消费者剩余，形成市场份额和利润的最大化[10]。

3.3. 算法应用安全风险

算法设计理论上需服务于人类的安全和福祉，但与众多新兴技术类似，以算法为核心的数字信息和人工智能技术也蕴含着潜在的重大安全风险。算法的应用可能带来广泛的社会经济或者特定主体的生命财产损害。此类应用安全风险与算法本身的漏洞、网络安全和相关主体的意志有关。程学旗等(2024)提出依据人机融合的程度，智能算法安全包括算法自身的一元内生性安全、人机二元应用性安全、人机共生的复杂社会系统中多元系统性安全[11]。一元内生性安全风险与算法执行具有内在适用边界的任务不当

时所涉及的物理世界中的事故有关，如自动驾驶、自动交易和其他替代性操作，一旦存在漏洞或受到干扰将造成外在损害。人机二元应用性安全交互式智能服务提供过程中所引发的安全问题，包括算法压榨、信息泄露等。人机共生的多元系统性安全涉及物理空间和网络空间与人机共同参与社会活动，如有算法介入的网络社交平台、金融交易系统等。人工智能算法面临不可解释性、后门攻击、逆向攻击、投毒攻击、逃逸攻击、对抗样本攻击等方面的潜在技术安全问题与挑战[12]。例如：通过在神经网络中植入后门，攻击者可对特定算法实现干扰和控制，形成神经网络算法的后门攻击，而且此类方法逐渐演化，更加具有隐蔽性和现实性；通过让深度学习模型学习有毒数据特征来改变模型的决策边界，破坏神经网络模型的完整性和可用性，影响相关算法在推荐、医保、教育等领域的应用。

4. 算法黑箱安全伦理风险应对对策

4.1. 强化算法应用的数据和隐私保护

对算法应用可能带来的数据和隐私保护风险受到国内外各类监管者的重视。相关区域和国家也出台了一系列代表性监管法律法规。例如《欧盟通用数据保护条例》赋予数据主体对数据处理和数据画像等行为的反对权、知情权、访问权、更正权和删除权等，用户有权获知对数据处理的事实、原因、基本算法逻辑、预期分析后果和可能产生的风险[13]。我国的《个人信息保护法》也赋予了企业在处理数据时需确保合法、透明、正当的义务以及个人对信息的访问、修改和删除权。为进一步加强算法数据和隐私保护，还需完善上述权利和义务的系统化保障机制。例如，对用户知情同意规则予以规范化并建立审查机制，避免用户因知情同意条款的冗长或隐蔽而忽视个人信息保护。基于数据最小化原则，审查相关算法使用数据范围的合理性，并监督其匿名化过程。

4.2. 加强算法合理设计与应用的监管

为促进算法的公平合理与合法利用，应加强对算法设计和应用的监管，抑制算法设计者或应用者利用算法黑箱违法违规的主观动机和矫正算法自身局限性带来的不合理后果。算法设计和应用监管问题的主要应对措施是加强算法透明化治理。例如，欧盟《人工智能法案》规定提高高风险人工智能系统的开发和使用的透明度，要求高风险人工智能系统以及部分使用这些系统的公共实体在欧盟高风险人工智能系统数据库中注册。欧盟《数字服务法》建立在线平台的透明机制和明确的问责框架，杜绝算法偏见威胁平台用户权益，确保平台对其算法负责，要求数字平台授权研究人员访问其算法机制和数据系统，以提高其内部运营的透明程度；超大型平台则有义务通过采取基于风险的行动和对其风险管理系统进行独立审计来防止滥用其系统[14]。进一步加强算法透明度需要区分客观与主观带来的算法不透明问题。首先，对客观上难以解释的算法可通过算法审计机制予以规制。算法审计在实验中模拟用户与算法的交互过程并收集算法结果从而对其诊断评估[15]。对于主观上不愿披露和解释算法的情况，需平衡商业秘密和知识产权保护与透明度治理的关系。相关主体应当保证算法不存在违规违法设计，并在产生争议时承担证明责任。由于算法黑箱会增加修正性参数的识别和修正难度，可通过逆向监管措施予以矫正。例如，在算法推荐领域可通过模拟实验推算算法逻辑和运行效果，并以之为依据实施监管或者处罚。这种逆向监管不要求对算法进行解释或者公开其相关参数，而是以其运行结果作出判断。

4.3. 完善算法应用安全的预防性机制

预防性原则是技术风险治理的重要原则。算法应用可能带来的人身财产损害以及其他安全问题需通过预防性机制设计避免风险的发生。由于算法应用往往具有广泛性，采取事前预防措施相较于事后的补救措施而言更具有成本收益的合理性。算法应用的预防机制主要包括对算法产品投入使用前的评估检验

和应用过程中的持续监督。对此可借鉴欧盟的人工智能分级治理系统，基于不同算法的风险情况设置不同的监管措施，并对人类安全造成不可接受风险的人工智能系统采取严格的禁限措施。此类系统主要涉及有目的地操纵技术、利用人性弱点或根据各类个人特征等进行评价的系统等。为保证算法的风险评估的独立性，可通过专家咨询制度或设立相关的风险治理机构形成评估机制，明确风险评估的指标体系[16]。对于人工智能所实施的各类操作可能引发的风险，相关服务的提供者应该嵌入人工智能故障预警机制并构建完善的远程监督和控制体系，并确定人为控制相较于机器控制的优先性。

5. 结论

算法是人工智能发展的基础，其专业性和技术性使之难以被社会公众所完全理解并引发了算法黑箱问题。从社会智力的角度，算法黑箱因其隐蔽性而难以监管，因而容易产生安全和伦理风险。算法黑箱的安全伦理风险与其带来的社会生产效率和生活便利的提升相伴而行，其治理机制也需在多种利益衡量中寻求微妙平衡。算法黑箱治理应注重保证基本权利的维护。相应的机制设计需要以各类透明度和责任制度为依托。把算法设计和适用纳入有效的监管框架是促进相关行业健康发展的必然要求，也是在人工智能时代保证社会安全稳定和提升人类福祉的题中之义。

基金项目

本研究为 2024 年江苏省基础研究专项资金(软科学研究)项目“江苏人工智能安全及伦理风险治理框架研究”(BR2024027)的阶段性成果。

参考文献

[1] 谭九生, 范晓韵. 算法“黑箱”的成因、风险及其治理[J]. 湖南科技大学学报(社会科学版), 2020, 23(6): 92-99.

[2] 李春生. 技术治理中的算法“黑箱”及应对策略[J]. 青年记者, 2021(22): 5.

[3] 张爱军, 郭镇毓. 生成式人工智能的技术神话、祛魅与认知进路——以 Sora 模型为例[J]. 理论建设, 2025, 41(1): 74-89.

[4] 徐伟, 韦红梅. 数智时代算法推荐风险的法律治理[J]. 科技与法律, 2024(3): 27-36.

[5] 于文轩, 魏炜. 数据开放中的算法依赖: 发展模式与驱动路径[J]. 理论与改革, 2023(2): 109-122.

[6] 赵旖. 互联网平台算法推荐的风险检视与法律规制[J]. 产业创新研究, 2024(4): 47-49.

[7] 高慧珊. 算法推荐技术引发的用户隐私风险及应对研究[J]. 新媒体究, 2024, 10(12): 45-48, 62.

[8] 张荣, 徐飞. 智能医学算法决策的伦理困境及风险规制[J]. 医学与哲学, 2022, 43(8): 10-15.

[9] 宋冠澎, 王启帆. 医疗人工智能算法决策的伦理风险及规制策略[J]. 中国医学伦理学, 2024, 37(9): 1080-1086.

[10] 谢永江, 杨永兴, 刘涛. 个性化推荐算法的法律风险规制[J]. 北京科技大学学报(社会科版), 2024, 40(1): 77-85.

[11] 程学旗, 陈薇, 沈华伟, 等. 智能算法安全: 内涵、科学问题与展望[J]. 中国科学院院刊, 2024, 39(11): 1-10.

[12] 彭长根, 何兴, 谭伟杰, 等. 人工智能算法安全研究现状与对策[J]. 贵州师范大学学报: 自然科学版, 2022(6): 1-16, 134.

[13] 陈荣. 欧盟《通用数据保护条例》的亮点与启示——兼论隐私与个人信息保护[J]. 西部金融, 2019(10): 85-90.

[14] 崔文波, 张涛, 马海群, 等. 欧盟数据与算法安全治理: 特征与启示[J]. 信息资源管理学报, 2023(2): 30-41.

[15] 徐明华, 魏子瑶. 算法伦理的治理新范式: 算法审计的兴起、发展与未来[J]. 当代传播, 2023(1): 80-86.

[16] 曾雄, 梁正, 张辉. 中国人工智能风险治理体系构建与基于风险规制模式的理论阐述: 以生成式人工智能为例[J/OL]. 国际经济评论: 1-22. https://kns.cnki.net/kcms2/article/abstract?v=u3gpgSR0TKef-WMi8iAWRH8QALu1HWNxbaMkD8RT8YKfQyB_2i7Nu8TAa9jrncTbRygwxzWNOhAwAe5MXHgESolsim-mijXOoMbI3ZsdWVrG3D4RrOLTDZSDN7aAD6wb9jP6J6nK4OWC1_R-C-6foqGwBYG6Ozq4qaX44h-ZuI-harZl0W9mlYDoteNY_w7WoYK&uniplatform, 2025-02-26.