算法影响评估制度的实践路径分析*

王宇晓[1,2]，吴少卿[2]，张静怡[1,2]
（1. 中国信息通信研究院，北京 100010；2. 移动应用创新与治理技术工业和信息化部重点实验室，北京 100010）

摘 要：算法影响评估制度作为国内外算法治理的基础性制度，是当前阶段算法治理的主要方案，但在应用实施过程中仍存在评估对象泛化、评估方法透明性不足的问题，需进一步细化。聚焦分析算法影响评估制度在应用实施过程中应关注的评估对象、评估内容和评估方法，提出在实际评估过程中应根据初步风险对算法评估对象分类分级，加强对以机器学习为基础的复杂算法的评估，重点关注算法的应用环节，重点评估算法输入的数据与特征、输出的目标与结果、干预与保障措施以及个人权利响应。

关键词：算法影响评估；算法风险；算法治理；分类分级
中图分类号：TN919；D922.16　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2024.10.009
引用格式：王宇晓，吴少卿，张静怡. 算法影响评估制度的实践路径分析［J］. 网络安全与数据治理，2024，43（10）：57-61.

Practical solutions of algorithm impact assessment system

Wang Yuxiao[1,2]，Wu Shaoqing[2]，Zhang Jingyi[1,2]
（1. China Academy of Information and Communications Technology，Beijing 100010，China；
2. Key Laboratory of Mobile Application Innovation and Governance Technology，MIIT，Beijing 100010，China）

Abstract：As the basic system proposed in the first stage of algorithm governance，the algorithm impact assessment system is the main solution for algorithm governance at the current stage. However，during the application and implementation process，there are still problems of generalization of assessment objects and insufficient transparency of assessment methods，which need to be further elaborated. This article focuses on analyzing the objects，assessment content and assessment methods that should be assessed during the application and implementation of the adjustment impact assessment system. During the actual assessment process，it is proposed to initially classify and grade the adjustment assessment objects based on risks，and strengthen the use of machine learning-based algorithms. Evaluation focuses on the application stage of the algorithm，focusing on the input data and feature complexity of the algorithm，the output goals and results，intervention and safeguard measures，and individual rights responses.

Key words：algorithmic impact assessment；algorithmic risk；algorithmic governance；classification and grading

0 引言

随着人工智能技术的迅速发展，算法广泛应用于社会生活的方方面面，与此同时，算法带来的安全风险引起了广泛关注[1]。面对日益攀升的算法治理压力，算法影响评估制度（Algorithm Impact Assessment）作为重要的治理方法被提上中外立法议程，成为当下算法治理实践中瞩目的焦点[2]。算法影响评估制度是指依据系统制定的衡量标准对算法自动化决策系统的系统设计、应用流程和数据使用等内容进行全面评估，以明确该系统的风险等级和影响范围的一种算法治理实践[3]。本文主要剖析在算法实际应用过程中，算法影响评估制度应重点关注的评估对象与范围、评估内容中的要点和细化方法，以及对推动算法影响评估制度落地提出相关建议。

1 研究背景

有学者认为，算法影响评估制度起源于20世纪90年代的隐私影响评估制度，由欧盟《通用数据保护条例》（General Data Protection Regulation，GDPR）创设的数据保

*基金项目：工业和信息化部2022年产业技术基础公共服务平台（2022-234-226）

护影响评估制度（Data Protection Impact Assessment，DPIA）发展而来[3]。随着人工智能技术的发展和广泛应用，算法作为重要技术底座受到普遍关注，算法本身的安全风险及其衍生风险频发，算法影响评估制度逐渐被单独提出，并得到了世界各国立法与实践的响应[4]。例如2019年加拿大政府颁布《自动化决策指令》（Directive on Automated Decision-Making），系统化创建算法影响评估指标[5]。2023年美国国家标准与技术研究所发布《人工智能风险管理框架》，明确人工智能系统、服务和产品的风险评估方法和标准。2024年欧盟《人工智能法案》顺利通过，其中针对高风险类人工智能系统提出要进行第三方符合性评估，具有系统性风险的通用人工智能模型应进行模型评估。当前，世界各国对算法影响评估制度在算法治理中的作用均有共识，美国和加拿大的算法影响评估制度聚焦于政府公权力机构在行政事务处理中的自动化决策算法，而欧盟的实践则从人权保护出发，算法评估范围既包含政府公权力所使用的算法，也包括企业所使用的商业性算法。

在我国，算法影响评估制度已经作为一项核心制度工具被明确提出并实施[6]。例如，2024年2月发布的国家标准《生成式人工智能服务安全基本要求》，其中对模型算法提出安全评估的要求。2023年公布的《生成式人工智能服务管理暂行办法》，提出提供具有舆论属性或者社会动员能力的生成式人工智能服务，应当按照国家有关规定开展安全评估。《上海市推动人工智能大模型创新发展若干措施（2023—2025年）》中提出对大模型驱动的互联网信息服务，加强合规指导，履行安全评估、算法备案等相关程序[7]。在具体的立法层面，我国已出台的法律法规里虽然没有直接规定算法影响评估制度，但间接规定了算法服务提供者有对算法应用进行评估的义务。如《个人信息保护法》第55条借鉴GDPR中DPIA制度，规定了我国的个人信息保护影响评估制度，并将自动化决策作为其中重要的评估情形；2023年实施的《互联网信息服务深度管理规定》，要求深度合成服务提供者和技术者应定期审核、评估、验证生成合成类算法机制机理；2022年颁布生效的《互联网信息服务算法推荐管理规定》第8条则要求算法推荐服务提供者应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等，第27条要求具有舆论属性或者社会动员能力的算法推荐服务提供者应当按照国家有关规定开展安全评估。

从评估对象上来看，我国立法实践中所确立的算法影响评估制度更接近欧盟，其评估对象主要针对企业所使用的商业性算法。基于此本文主要探讨企业在落实算法影响评估制度时的方法和路径。

2 算法影响评估对象与范围

2.1 评估对象的明确

算法影响评估制度的评估对象是算法（Algorithm），但目前算法并没有明确、权威的定义，其概念的内涵和外延还有不确定性，如自动化决策（Automated Decision-Making）、数据画像（Profiling）、人工智能（Artificial Intelligence）等均被归类为算法。因此在实际开展算法影响评估过程中，存在评估对象泛化、评估对象不准确、关键问题被忽视等现象，评估对象应该进一步说明。

算法影响评估制度是受数据保护影响评估制度启发发展而来的，在针对自动化决策的影响评估基础上发展出了算法影响评估制度。也正是因此，在前期，算法影响评估与针对自动化决策系统的数据处理活动影响评估几乎等同，提起算法关注度更多的是自动化决策。但随着人工智能技术的发展，算法影响评估制度开始逐步独立出来。

在开展评估过程中，自动化决策是在个人信息处理过程中采取的措施，其中重要组成部分是数据画像。欧盟《关于自动化个人决策目的和数据画像目的准则》中规定，数据画像由三要素组成：它必须是一种自动化的处理形式；它的实施必须是针对个人数据的处理；它的目的必须用来评估关于某个自然人的私人方面[8]。我国《个人信息保护法》规定，自动化决策是指通过计算机程序自动分析、评估个人的行为习惯、兴趣爱好或者经济、健康、信用状况等，并进行决策的活动。自动化决策过程一般涉及收集使用和处理个人信息或对个人信息主体进行评估或分析等[9]。而算法影响评估制度下的算法，超出了自动化决策系统的范畴，被评估算法可能不涉及处理个人信息或对个人信息主体的评估和分析，例如《互联网信息服务算法推荐管理规定》所规定的生成合成类算法、检索过滤类算法、精选排序类算法，并非针对个人信息保护所规定的自动化决策。

2.2 评估范围

但算法影响评估制度所要评估的算法是什么的问题仍需厘清。在不同专业领域内，对算法存在不同的解释，例如，在计算机科学中，算法是指一个被定义好的、计算机可施行其指示的有限步骤或次序，常用于计算、数据处理和自动推理，也有的定义描述算法为模型分析的一组可行的、确定的和有穷的规则[10]。在数学中，算法被认为是用于计算的方法，通过这种方法可以达到预期的计算结果，或者是对特定问题的求解步骤的一种精确描述方法。从这些定义可以看出，只要是为实现某一目

标而明确设定的一系列步骤或策略，都可能被称为算法。广义的算法既包括在人工智能（AI）、机器学习（Machine-Learning）等基础上产生的复杂算法，也包括简单的函数公式、自动化的程序工具，甚至包括非数学领域的决策程序如平台运营策略。

但算法影响评估制度并不针对上述广义范围内的所有算法。广义算法中简单函数公式、自动化程序或策略等不应是算法影响评估制度规范重点。算法影响评估制度需要重点关注、加强评估的是利用机器学习、人工智能技术等复杂数据处理方式进行计算，输出相应预期结果的自动化系统。以机器学习技术为基础的算法与传统规则工具不同。传统规则工具是设计出一套逻辑规则，机器按该流程执行。而在机器学习或人工智能算法中，系统可以根据输入数据不断自我归纳、调优，寻找相关性，完成规则的制定。机器学习的自我学习能力，使得算法逐渐获得了自我决策、自我设定规则的能力，算法可做的事情越来越接近真实的人类。算法逐渐从“辅助地位”转变为“决策地位”，人类由曾经的绝对决策方，有时变成了被决策的一方，甚至开始担心变成被操纵的一方。算法影响评估制度的本质正是要应对这种角色异化所带来的新型风险，因此对此类算法进行重点评估十分有必要。

2.3 评估风险的分级

在对算法开展评估的过程中，为避免造成评估资源的浪费，应先对评估对象分类分级，将算法影响评估与算法分级相结合。根据算法风险进行初步区分，可以将评估资源集中于高风险算法，提高低风险算法的评估效率。

算法分级分类是当前算法治理的有效手段。如欧盟的《人工智能法案》将 AI 系统风险分成了四层，分别为不可接受、高风险、有限风险和极小风险，针对不同风险级别的算法匹配了相应的管理强度，如极小风险的人工智能系统可不进行任何干预。我国的《互联网信息服务算法推荐管理规定》要求建立算法分类分级安全管理制度，根据算法的内容类别、用户规模、算法技术处理的数据重要程度、对用户行为的干预程度等对算法进行分类分级管理。

在开展算法影响评估时，可参照风险分级分类管理的思路，对待评估算法分级管理。根据风险程度，采取不同的管理措施。例如对高风险级别算法进行严格的评估审查，对低风险级别算法进行一般评估或备案审核。判断算法风险等级可以综合考虑以下因素：其一，算法应用的服务和功能是否具有重要性，是否在产品的核心功能或服务中使用，服务是否展示在首页、弹窗等重点区域。其二，算法所处理的数据是否敏感，如涉及处理生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等敏感个人信息，或处理不满十四周岁未成年人的个人信息。其三，算法是否对用户思想认知产生影响，如可能直接干预影响用户行为。其四，算法结果是否会危害用户或国家安全，如影响个人征信、医疗服务、教育或就业资格等利益，可能涉及国家和社会公共安全。

3 算法影响评估的内容与要点

从算法影响的类型来看，算法风险大体上可以分为社会风险和技术风险两个方面。欧盟于 2019 年发布的《可信赖人工智能伦理指南》（Ethics Guidelines for Trustworthy AI）提出，可信赖人工智能需符合三个基本条件：合法性、合伦理性、鲁棒性，其中合法性和合伦理性侧重于算法策略的社会风险，而鲁棒性则指算法的准确性、完整性、可用性等安全技术风险[11]。因此，算法影响评估的评估内容应从社会风险和技术风险两方面出发，并充分协调、综合考虑两部分评估结论。

其次，算法影响评估除针对算法技术细节本身外，应多关注算法服务和应用环节。一是如果算法仅是在测试环境或系统中应用，没有部署于任何实际服务中，则无法对用户个体或者社会产生影响，其危害性几乎没有[12]。二是机器学习技术原理使得算法模型可以实现“自我学习”，其中技术原理复杂，难以直接被人理解，而算法服务或应用环节则更加直接、清晰，对其进行评估所产生的规制效果更加显著，能有效应对算法角色的异化。三是算法技术在不断地快速迭代更新，其原理本身呈现专业性和复杂性的特点[13]，如果将评估重点落在算法技术本身上，则对评估方法、技术、人员及时效的要求提高。

因此，在开展算法影响评估时，应聚焦于评估算法输入的数据与特征，输出目标、结果及应用场景，干预与保障措施，用户权益响应四方面，如图 1 所示。

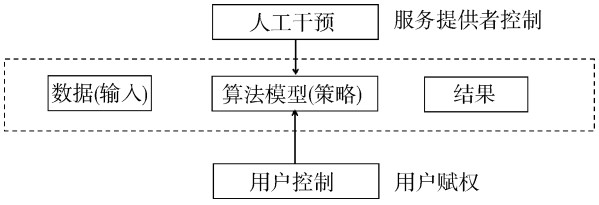

图 1 算法影响评估过程

3.1 输入的数据与特征

输入的数据与特征的评估，是指算法服务提供者应评估在算法模型训练、测试和具体部署时，输入算法和通过算法处理的各类数据的风险，主要考虑合法性、必

要性和数据质量方面的评估。首先，输入数据合法性的评估，是指需要判断待处理的数据来源是否合法、数据源本身是否存在风险。如数据源是否侵犯他人的知识产权、商业秘密，是否涉及重要数据。当输入数据中包含个人信息的情况下，个人信息的收集使用是否具备合法性基础，是否超出原有的“告知同意”授权范围。输入数据的训练、标注环节是否具有安全性，流程、标准和内容是否可能导致歧视。以上都是合法性评估的重要环节。其次，必要性是指需要评估处理的数据是否超出必要的范围，是否遵循最小必要的原则，输入的特征是否可能导致过拟合问题，考察处理的个人信息是否可以进行去标识化或匿名化。然而，一般在实际测试或部署环境中，模型训练阶段难以判断某类数据或特征是否对模型有价值，往往需要在测试阶段进一步验证，因此在模型训练阶段可以适当放宽或者平衡必要性的评估标准。最后，数据质量主要评估选取的用于训练模型的数据及特征样本是否具有多样性。在训练模型阶段，特征样本的选择如果不够准确、广泛，就可能导致算法的偏见，例如面部识别模型如果选取的训练数据中不存在某类面部类型时，面部识别系统处理结果表现不佳。因此，需要在数据训练阶段就评估数据样本的选择是否单一，是否真实准确、完整可用，是否与模型目标相关和匹配。

3.2 输出目标、结果及应用场景

输出目标、结果及应用场景的评估，是指算法服务提供者评估算法是否达成设计阶段设定的目标，算法应用后算法输出的结果所可能产生的风险，以及算法实际应用场景是何种情况，主要包括对目的正当性、个人权益侵害性以及公共利益侵害性的评估。目的正当性一般分为两个层次：一是算法模型本身的计算目标；二是算法模型部署应用后所预期达到的目标。例如个性化推荐的算法服务，前者的目标可能是指预测用户的点击率，为用户推荐、匹配感兴趣的内容，而后者则可能是增加用户的留存。在实际的业务操作中，两个目标共存但有所差异，为实现不同业务目标而部署的提供同类预测结果的算法模型，其最终产生的影响可能不同。个人权益的侵害性，指的是算法可能影响的个人权益。我国个人信息保护法重点关注自动化决策是否会对个人权益产生重大影响，重大影响一般包括法律影响或近似重大影响，近似重大影响包括导致合同解除，或影响个人的征信、医疗服务、求职和求学等情形[14]。除此之外，算法对个人权益的影响还包括可能导致歧视、偏见、剥削等有违公平、公正原则的结果，例如电商平台“大数据杀熟”或其他对消费者不合理的差别待遇等。算法还可能导致用户沉迷，产生信息茧房，以及侵害未成年人、老年人等特殊群体的权益。公共利益的侵害性主要体现在信息内容安全、市场竞争秩序以及社会动员等方面。信息内容安全主要是评估算法是否导致违法、不良或低俗信息等不可控传播的风险，以及是否保障算法推荐内容的全面性、客观性和多样性。我国《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》更多从信息内容安全的角度对算法进行规制。算法对市场竞争秩序影响则指的是大型平台为了限制竞争、影响舆论，通过算法操纵榜单或者检索结果排序、控制热搜或者精选等干预信息呈现，实施自我优待[15]。社会动员层面的风险则可能影响国家安全层面，例如剑桥分析事件对美国大选的影响[16]。

3.3 干预与保障措施

干预与保障措施的评估，是指算法服务提供者对算法结果进行的人工干预，以及保障算法模型运行安全的各项措施。为了防止算法对人类个体进行决策和操纵，对算法最终决策结果进行适当的人工干预有一定的必要性。对算法的干预控制首先应以算法的可解释性为基础，如果算法的决策结果完全由技术人员都无法进行有效解释的算法黑箱产生，那么其可能带来的偏见和歧视风险也无从识别和纠正。在算法可解释性的基础上，以个性化推荐算法为例，常见干预措施一般包括过滤消重、插排打散等。过滤消重是指对算法召回内容进行过滤或消重，使得违规违法、不良信息、重复内容等不被推荐。插排打散是指防止同类内容过于密集，对算法推荐的召回内容进行一定顺序的调整。除此之外，多样的人工运营手段也是对算法进行干预的有效措施。安全保障措施则是指算法服务提供者通过完备的管理和技术手段来确保算法模型运行的安全鲁棒性。在确保安全性的同时，算法服务提供者在部署上线前还应当通过各类测试或检验方式对算法决策的准确性进行验证，常见的测试方法包括离线数据集测试、A/B 测试等。

3.4 用户权利响应

用户权利响应的评估，是指算法服务提供者评估算法应用中对用户权益保护的要求。算法服务提供者应当提供给个人权利如知情权，用户可了解算法应用过程中如何处理用户个人信息、如何作出用户权益；选择权，用户可选择使用算法与否、如何使用等；控制权，用户可拒绝算法产生的自动化决策等。知情权与算法服务透明度义务息息相关，算法服务提供者应当以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制，明确算法在处理个人信息时是否保障个人信息主体的查询、删除、撤回、可携带等权利。拒绝自动化决策则是指算法服务提供者需要向个人提供不针对其个人特征的

选项，或者向个人提供便捷的拒绝方式，以及当算法是通过自动化决策方式作出对个人权益有重大影响的决定时，需要响应个人要求对决策结果进行说明，并拒绝仅通过自动化决策的方式作出决定的权利。除了上述基本的控制权利之外，还可以考察评估算法服务提供者是否提供其他补充的控制权以保障个体的权益，以及是否提供了个体权利请求和反馈的渠道。

4 结论

算法影响评估制度是当前阶段算法治理的重要工具，它与算法可解释性、算法归责等制度协同，共同搭建起算法综合治理体系。目前，各国对算法及其影响评估规定仍有待细化，行业最佳实践方案也在探索阶段，然而，随着人工智能的迅速发展，算法治理成为保障技术发展与应用的重要手段，有效地开展算法治理能够促进产业高质量发展。本文介绍了在开展算法评估实践过程中应考虑的评估对象、范围、评估内容和要点以及应重点考虑的问题。首先，算法评估对象和范围应当细化，按照分类分级思路明确评估对象，重点加强对以机器学习为基础的高风险算法开展评估，对于低风险算法，可采取备案、监测等方式，避免将算法影响评估制度泛化。其次，算法评估重点要更多地聚焦在算法服务环节，厘清评估内容和要点，不应只把技术评估结果作为风险判断依据，实际在算法应用及业务实践过程中，可能存在直接或间接的风险隐患。未来在算法治理和实践方面，应重点考虑明确以上主要问题。

参考文献

[1] 谢琳，曾俊森. 算法影响评估制度的定位与完善 [J]. 数字经济与法治，2023（1）：57－82，294.

[2] 张欣. 算法影响评估制度的构建机理与中国方案 [J]. 法商研究，2021，38（2）：102－115.

[3] 张惠彬，仲思睿. 数字经济时代算法推荐技术的应用风险与规范进路 [J]. 杭州师范大学学报（社会科学版），2022，44（5）：122－130.

[4] 张凌寒. 算法评估制度如何在平台问责中发挥作用 [J]. 上海政法学院学报（法治论丛），2021，36（3）：45－57.

[5] 石佳友，曾佳. 个人信息保护影响评估：制度内涵与完善路径 [J]. 西北工业大学学报（社会科学版），2022（4）：90－102.

[6] 赵宏. 公共决策适用算法技术的规范分析与实体边界 [J]. 社会科学文摘，2023（6）：11－13.

[7] 上海市经济和信息化委员会. 关于印发《上海市推动人工智能大模型创新发展若干措施（2023－2025 年）》的通知 [EB/OL]. [2024－06－17]. https：//app. sheitc. sh. gov. cn/jsjb/695961. htm.

[8] European Commission. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (wp251rev. 01) [EB/OL]. [2024－06－30]. https：//ec. europa. eu/newsroom/article29/items/612053/en.

[9] 陈林林，严书元. 论个人信息保护立法中的平等原则 [J]. 华东政法大学学报，2021，24（5）：6－16.

[10] 郭少飞. 算法法人的理论证立及构成要素探析 [J]. 东方法学，2021（5）：93－109.

[11] 张全洁. 人工智能的伦理准则研究 [D]. 苏州：苏州大学，2020.

[12] 张欣. 免受自动化决策约束权的制度逻辑与本土构建 [J]. 华东政法大学学报，2021，24（5）：27－40.

[13] 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.

[14] 宋华健. 反思与重塑：个人信息算法自动化决策的规制逻辑 [J]. 西北民族大学学报（哲学社会科学版），2021（6）：99－106.

[15] 钛媒体. 网信办出手规范算法推荐，点名大数据杀熟、操纵榜单、流量造假、诱导沉迷 [EB/OL]. （2021－08－20）[2024－06－30]. https：//tech. ifeng. com/c/894tp3E3R8v.

[16] REESE H. The Cambridge Analytica whistleblower on how American voters are “primed to be exploited” [EB/OL]. [2024－06－17]. https：//www. vox. com/the-highlight/2019/10/28/20932790/chris-wylie-cambridge-analytica-facebook-trump 2020/2020－01－12.

（收稿日期：2024－07－04）

作者简介：

王宇晓（1991－），通信作者，女，硕士，工程师，主要研究方向：移动安全、数据安全、个人信息保护。E-mail：wangyuxiao@ caict. ac. cn。

吴少卿（1990－），男，硕士，主要研究方向：个人信息保护、数据安全、大模型合规。

张静怡（1995－），女，学士，助理工程师，主要研究方向：个人信息保护、数据治理。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部