论人工智能算法黑箱的法律风险及应对策略

●于海梅

[摘要]人工智能技术，尤其是深度学习和机器学习算法，使其具有了一定的自主性，但也导致了内部决策过程的不透明，即算法黑箱。算法的不公开、不透明，被称为算法黑箱。算法黑箱问题带来的主要挑战，包括隐私保护泄露、责任归属认定模糊及现有法律体系的适用性问题。算法黑箱问题不仅使用户难以理解和信任人工智能系统的决策，也给法律监管和伦理审查带来了重大挑战。基于此，本文提出了分层分类监管、监管沙盒、自律机制和多元共治等应对策略，以促进人工智能算法的健康发展和良好治理。

[关键词]人工智能；算法黑箱；法律风险；应对策略

人工智能算法黑箱的风险

（一）隐私泄露和数据歧视

人工智能的开发和应用，涉及大量用户个人信息和隐私数据的收集与使用。算法黑箱导致用户无法对其决策过程进行监督和控制，无法有效控制和保护自身隐私，可能会使用户的隐私权面临侵犯风险。

此外，数据收集和处理过程中的偏差也可能导致数据歧视。算法决策是一个在掌握数据属性、人类偏好等基础上的深度学习过程。若数据本身存在错误或数据参数、属性不够全面，就会形成歧视性的算法结果。

（二）责任归属的模糊性

当算法决策行为导致损害时，确定责任归属变得异常复杂。人工智能系统通常涉及多个参与者，包括开发者、数据提供者和使用者等，当决策行为导致错误或损害时，如何确定每个参与者的责任份额成为不得不面临的难题。

（三）透明度和可解释性缺失

算法黑箱使得算法决策不透明，不仅导致公众对决策结果的不信任，也给监管和司法审查带来了挑战。特别在医疗、司法和金融等关键领域，算法的决策结果需要能够被理解和审查，以确保其符合伦理标准和法律要求。虽然学者们提出了各种可解释人工智能 AI（XAI）技术，如 LIME 和 SHAP，试图为人工智能决策提供人类可理解的解释，但其具体实施仍存在争议。

人工智能算法黑箱给法律治理模式带来的挑战

人工智能算法黑箱不仅带来了具体的法律风险，更给整个法律治理模式带来了严峻的挑战。

（一）法律效力的削弱

在法律适用方面，算法的不可解释性直接影响了法律适用的准确性和公正性，可能导致法律适用过程中出现重大偏差。正如 Burrell（2016）所指出的，即使是系统的设计者也可能无法完全解释系统的某些决策。算法黑箱的这些特性，不仅挑战了法律的可预测性原则，也使得在发生争议时难以确定责任归属。

（二）法律责任认定的模糊化

在人工智能系统中，一个决策可能是多个因素综合作用的结果，当这些系统的行为导致不利后果时，其决策的具体过程和原因变得更难追溯，增加了因果关系的认定困难。同时，人工智能系统没有人类意义上的“主观心理状态”，这使得基于过错的责任认定受到挑战。Scherer（2016）提出的“合理算法”标准，试图将责任认定的焦点从主观心理状态转移到客观的技术标准上。但这种转变需要法律界和技术界的密切合作，以制定出既科学又公平的评判标准。然而，客观化的责任认定标准也会引发新的问题，比如，如何在保持法律公正性的同时不过度抑制技术创新。

（三）法律保障的缺失

在传统法律程序中，公开性和可审查性是保障程序正义的关键要素。然而，人工智能系统决策的不透明，使得利益相关方无法评估决策的公正性和合法性，对程序正义产生威胁。同时，当个人受到人工智能系统不利决定影响时，由于无法了解决策依据，其寻求法律救济的能力受到严重限制。对此，有学者提出“反事实解释权”概念，要求人工智能系统能够提供“如果……就会……”类型的解释，帮助个人理解如何改变自己的情况以获得不同的结果。事实上，

“反事实解释权”的实现需要人工智能系统的设计者和使用者在系统开发之初就考虑到这一需求，这可能会增加人工智能系统的开发成本和复杂性。除此之外，平等保护的困境也使人工智能系统在无意中增加了社会不平等的风险。

（四）法律价值的冲突

第一，效率与公平的权衡。人工智能能够显著提高决策效率，但这种效率的提升通常以牺牲公平和个性化为代价。在刑事司法领域，人工智能辅助的风险评估工具可以快速处理大量数据，但可能忽视个案的具体情况，导致不公平的结果。因此，在法律框架中如何设定效率和公平的权重，并设计相应的监督和纠正机制是一个复杂的法律和政策问题。

第二，创新与监管的平衡。对人工智能过度监管可能抑制其创新，而监管不足则可能导致社会风险。“自适应监管”理论强调，监管应该随着技术的发展而不断调整，以保持其相关性和有效性。实现这一目标需要建立更灵活的法律框架，如采用原则性立法结合动态监管指引的方式，或者建立定期审查和更新人工智能相关法律法规的机制，这些措施的实施可能会面临法律确定性和可预测性的挑战。

第三，透明与保密之间的矛盾。针对人工智能算法黑箱的特性，提高人工智能的透明度有助于增加公众信任和加大法律监督力度，但也可能损害商业秘密和个人隐私，体现了透明与保密之间的矛盾。

Q 人工智能算法黑箱的法律应对策略

（一）分层分类监管：平衡 AI 创新与风险控制的精细化策略

分层分类监管策略旨在根据人工智能的风险程度和应用领域采取差异化的监管措施，以平衡创新与风险控制的需求。该策略的核心在于对人工智能算法进行风险分级，并针对不同风险级别制定相应的监管措施，从而实现监管资源的优化配置和监管效果的最大化。欧盟的《人工智能法案》提供了可借鉴模式。该法案将人工智能算法分为四个风险层级，并对不同层级采取相应的监管措施，体现了风险导向的监管理念。

2023 年 8 月 15 日，我国《生成式人工智能服务管理暂行办法》（以下简称《办法》）正式实施。《办法》对生成式人工智能提出了包容审慎和分类分级监管的要求。然而，该《办法》对分级分类监管的规定仍显模糊和笼统，未能明确划分风险的级别与标准，在可操作性方面有待进一步完善。分层分类监管的框架主要包括：第一，建立科学、动态的风险评估框架。这种框架不仅需要考虑技术风险，还需要纳入社会、伦理等方面的风险评估。第二，制定差异化监管措施以更好地适应不同行业和领域的特殊需求。第三，建立跨部门协调机制，有效避免监管重叠或监管真空问题。

（二）监管沙盒：AI 创新与监管协同的实验场

在分层分类监管的基础上，监管沙盒可以作为评估和验证人工智能系统的有效工具。监管沙盒为人工智能创新提供了一个受控的实验环境，允许企业在有限的范围内测试新技术，同时为监管机构提供了解和评估新技术的机会。

2015 年，英国金融行为监管局（FCA）将监管沙盒模式引入金融科技领域。监管沙盒的核心原理是，由公共机关建立一个类似于“安全空间”的测试平台，即在一个相对封闭和安全的环境中，对人工智能系统进行有限的和可控的试验和测试，以评估人工智能系统的性能和效果，以及可能带来的风险和影响。同时，沙盒试验也为人工智能系统的提供者和使用者提供一定的法律豁免和优惠，以鼓励人工智能的探索和创新。

监管沙盒的实施可以分为申请核准－测试－退出与总结三个阶段。首先，企业提交测试申请，由监管部门根据提交的申请材料进行尽职调查与资格审查，重点评估项目的创新突破意义、对消费者福利的提升程度等。其次，企业与监管部门商定测试方案，进入沙盒测试。测试过程中，监管部门进行全程监测，实时提出合规性评价与指导。同时，企业需要履行信息报告义务并接受监督。若发现损害消费者利益或有碍金融稳定的现象，监管部门会立即终止测试，确保问题不会扩散到沙盒外部。最后，企业在测试结束后必须主动退出沙盒。监管部门对测试的项目进行评估，并向企业反馈评估结果。2023 年，我国提出既要鼓励企业在数据要素应用上的首创精神，又要建立有效的数据沙箱机制，防范化解重大风险。全国多地也积极探索将数据创新融通应用纳入监管沙盒，这为人工智能算法黑箱治理的变革提供了可行方案。

（三）自律机制：AI 行业自治与道德约束的内生动力

对人工智能算法黑箱的法律治理而言，不论其私法权利及其救济体系如何设计，也不论监管机制如何完善，算法自律都是控制风险的第一要务。自律机制强调行业内部的自我约束和管理，在人工智能领域可以作为正式法律规制的有效补充，更灵活、及时地应对技术变革带来的挑战。

自律机制的具体实施需要多管齐下。具体来讲，可以通过由行业相关专业人员组成的行业自律组织，制定和实施算法黑箱透明度和可解释性的自律规范和措施。如算法黑箱的伦理准则、行业标准、社会责任等，以规范算法黑箱的行为和影响，同时也为算法黑箱的提供者和使用者提供一定的指导和支持。其中，算法自律相关规范应当明确算法伦理，将算法伦理的具体要求嵌入制度中，推动算法伦理法律化。这有助于以算法伦理为衡量标准推动算法规范化，依靠法律的强制力保障算法不偏离伦理轨道。

同时，加强对算法设计者和使用者的伦理教育，增强其伦理意识，确保算法的设计和应用符合人类道德观念的要求，也是自律机制中不可或缺的一环。通过自律机制的建立和完善，可以在一定程度上弥补正式法律规制的不足，为人工智能算法黑箱的治理提供更为灵活和有效的补充。

（四）多元共治：AI治理的跨界协作与社会共识

为应对算法黑箱带来的不透明性、不公平性以及其他潜在风险，传统单一主体的治理往往难以全面、有效地解决问题。因此，需要引入多元共治机制，通过主管部门、企业、行业自律组织、社会公众等多个主体的共同参与和协作，形成合力，共同应对算法黑箱带来的挑战。多元共治不仅是一种治理模式，更是一种价值理念，体现了在复杂技术环境下实现社会共识的努力。通过多元共治，可以更好地平衡技术创新、经济发展、社会公平和伦理道德等多重目标。这种平衡不仅有助于提高人工智能治理的质量，也能增强公众对人工智能技术的信任和接受度。

多元共治需要从多角度入手。第一，主管部门应加大对算法技术的监管力度，制定相关法律法规和政策来规范算法技术的使用和发展。同时，主管部门还应建立专门的监管机构来负责算法技术的监管工作，确保监管效果的有效性和公正性。第二，企业应自觉遵守法律法规和政策要求，加大内部管理和自我监管力度。通过建立健全的算法技术管理制度和流程规范来确保算法技术的合规运营和安全性。此外，企业还应积极履行社会责任和义务，推动算法技术的健康有序发展。第三，社会组织、媒体和公众等第三方力量应积极参与算法技术的监督和评估工作。通过公开透明的方式揭露算法黑箱带来的问题和风险，推动相关企业和主管部门采取有效措施进行改进和完善。鉴于算法技术的全球性和跨国性特点，主管部门和企业应与国际社会进行合作和对话，以促进算法黑箱的全球治理和协调。

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(06)：92-99.

[2]陈醇.私法制度中的代数算法黑箱及其应对[J].法学评论，2022，40(01)：55-68.

[3]徐凤.人工智能算法黑箱的法律规制——以智能投顾为例展开[J].东方法学，2019(06)：78-86.

[4]欧达婧.“算法黑箱”与“算法公开”天然冲突的解决路径——以我国算法解释与算法备案的适用为中心[J].南海法学，2023，7(02)：68-77.

[5]丁晓东.算法与歧视从美国教育平权案看算法伦理与法律解释[J].中外法学，2017，29(06)：1609-1623.

[6]刘友华.算法偏见及其规制路径研究[J].法学杂志，2019，40(06)：55-66.

基金项目：

山东省人文社会科学课题项目，项目名称：风险社会视野下人工智能算法“黑箱”的法律规制，项目编号：2023-XWKZ-018。

作者简介：

于海梅(1979－)，女，汉族，山东潍坊人，硕士，讲师，潍坊学院，研究方向：经济法。