算法黑箱与透明性悖论：可解释 AI 的哲学困境与突破路径

张朔境
中国司法大数据研究院　北京　100041

摘　要：以深度学习为代表的AI复杂模型性能卓越，却因决策不透明引发“算法黑箱”问题。社会对“可解释AI”（Explainable AI, XAI）需求迫切，却遭遇“透明性悖论”，即模型复杂性与可解释性存在根本张力。本文深入剖析该悖论背后的哲学困境，从认识论、本体论与价值论三个维度探讨。先界定“黑箱”与“解释”内涵，揭示XAI领域科学主义与实用主义解释观冲突、人类认知局限与机器智能涌现的鸿沟、透明性伦理摇摆这三种核心张力。接着批判审视主流XAI技术路径，提出“语境化解释”框架，将解释视为动态的社会 - 技术实践。最终认为，解决透明性悖论关键在于构建多元主体参与、权责清晰、能持续对话协商的治理生态系统。

关键词：算法黑箱；可解释AI；透明性悖论；认识论；算法治理

引言

2026年，人工智能深度融入社会，在信贷审批、医疗诊断、司法裁判等诸多关键决策领域发挥重要作用。但基于深度神经网络的AI系统像“黑箱”，人们难明其决策因果与推理逻辑。这削弱了人类信任与控制感，错误或偏见出现时追责、修正和预防困难重重。“可解释AI”（XAI）成为学界与业界目标，旨在让AI决策透明、可理解、可追溯。然而，XAI研究深入后浮现“透明性悖论”：性能强大、应用广泛的AI模型结构复杂难解释，易解释的简单模型处理现实问题又力不从心，我们似乎要在“有效”与“可理解”间二选一。本文认为，此悖论根源在于哲学困境，仅从工程角度解决恐难奏效。需跳出纯技术视域，从认识论、本体论、价值论三个维度剖析困境，探索更具建设性和可行性的突破路径。

1 概念澄明：“黑箱”与“解释”的多重面孔

1.1“黑箱”的谱系学

“黑箱”一词源于控制论，原指内部机制未知但能通过输入 – 输出关系研究的系统，在 AI 语境下它至少有三层含义。技术性黑箱，指模型内部参数、权重及计算路径极端复杂，超出人类心智直接解析能力，如大型语言模型决策是海量非线性变换产物，无法用简洁规则描述。制度性黑箱，是因商业机密、知识产权保护或国家安全等因素有意封闭，科技公司为竞争常拒绝公开核心算法细节，让外部监督难以开展[1]。认知性黑箱，即便提供所有技术细节，普通用户和部分专家也可能因缺乏专业知识而无法理解，它是主客体间认知鸿沟的体现。由此可见，“黑箱”既是技术现实，也是社会建构产物，XAI 的首要任务就是区分并应对不同类型的黑箱。

1.2“解释”的多元诉求

同样，“解释”也并非单一目标。不同的利益相关者对“解释”有着截然不同的期待：开发者/工程师需要的是调试性解释，旨在理解模型为何失败、如何改进其架构或训练过程。监管者/审计员需要的是合规性解释，用以验证模型是否符合法律法规（如GDPR中的“解释权”），是否存在歧视性偏见。终端用户/受影响者需要的是正当性解释，他们关心的是“为什么是我？”、“这个决定对我公平吗？”，期望获得一个能被其价值观和生活经验所接纳的理由。科学家/研究者则可能追求因果性解释，试图揭示数据背后的真实因果机制，而非仅仅是统计相关性。这些不同类型的解释，在内容、形式和深度上都存在巨大差异。一个能满足工程师调试需求的特征重要性图，对于一位被拒贷的用户而言，可能毫无意义甚至充满冒犯。因此，XAI的“解释”必须是语境敏感和受众导向的。

2 哲学困境：透明性悖论的三重维度

正是上述概念的复杂性，催生了XAI领域内难以调和的哲学困境，构成了透明性悖论的深层结构。

2.1 认识论困境：科学主义 vs. 实用主义的解释观

XAI领域的认识论冲突，集中体现在两种解释观的对立上。科学主义解释观认为，真正的解释必须揭示系统内部真实的、客观的因果机制。它追求一种“上帝视角”下的完备性与保真度（Fidelity）。在此范式下，理想的XAI方法应能精确还原模型的内部计算逻辑，任何近似或简化都是对真理的背叛。这种方法论深受物理学等硬科学的影响，强调解释的客观性和普遍性。实用主义解释观则认为，解释的价值在于其效用，而非其与“真实”的契合度。只要一个解释能够帮助特定用户达成其目标（如建立信任、做出决策、进行申诉），那么它就是有效的。在

此视角下，一个高度简化的、甚至带有一定“虚构”色彩的类比或故事（如LIME生成的局部线性近似），只要能被用户理解和接受，就具有充分的解释力。这两种解释观的冲突构成了XAI的核心张力[2]。科学主义者批评实用主义方法（如许多后验解释器）是“对黑箱的另一个黑箱”，其生成的解释可能与模型的真实行为相去甚远，甚至产生误导。而实用主义者则反诘，科学主义的完美解释在复杂模型面前要么不可得，要么其复杂程度使其自身成为一个新的黑箱，从而失去了向人类沟通的意义。这种认识论上的分裂，使得XAI社区在评估方法优劣时缺乏统一标准，也导致了技术发展路径的迷茫。

2.2 本体论困境：人类中心主义与机器智能的“他者性”

本体论困境触及了人与AI的根本关系。传统上，我们将解释视为一种主体间性（Intersubjectivity）活动，即一个有意识的主体（解释者）向另一个有意识的主体（被解释者）阐明其意图、理由或信念。这种模式预设了解释双方共享一套基本的认知框架、语言体系和价值观念。然而，AI，特别是当前的深度学习模型，并不具备人类意义上的意识、意图或信念。它的“智能”是一种涌现（Emergence）现象，是海量数据和复杂计算交互产生的结果，而非源于一个可被内省的“心智”。这意味着，当我们要求AI“解释”其决策时，我们实际上是在向一个“他者”（The Other）——一个在本体论上与我们截然不同的存在——强加一个人类中心主义的理解框架。这种错位带来了两个难题。其一，解释的拟人化陷阱。为了便于理解，XAI方法常常赋予AI以人类特质，如“注意力”、“偏好”或“理由”。这种修辞虽有助于沟通，但也可能掩盖了AI决策的真正本质——一种非意向性的、基于统计模式匹配的计算过程，从而产生虚假的熟悉感和控制感。其二，认知鸿沟的不可逾越性。人类的认知是符号化、离散化和叙事化的，而深度学习模型的表征往往是高维、连续且分布式（Distributed）的。试图将后者强行翻译成前者，必然会丢失大量信息，造成解释的失真。我们或许永远无法像理解另一个人那样去“理解”一个深度神经网络，因为二者运行在完全不同的本体论平面上。

2.3 价值论困境：透明性的工具理性与价值理性之争

最后，关于为何需要XAI的价值论争论，也充满了张力。透明性究竟是一种工具价值（Instrumental Value）还是一种内在价值（Intrinsic Value）？工具理性视角认为，透明性本身并无好坏，其价值完全取决于它能否服务于其他更高阶的目标，如提升模型性能（通过更好的调试）、增强用户信任（从而促进采纳）、满足法律合规要求、或实现更有效的问责。在此框架下，如果某种不透明的模型能完美地、无偏见地完成任务，那么追求透明性就是多余的，甚至是低效的。透明性只是达成目的的手段[3]。价值理性视角则主张，在涉及人类重大利益的决策领域，透明性本身就是一种道德要求和民主基石。它关乎人的尊严、自主权和知情同意权。一个无法被质询、无法被理解的权力（即使是算法权力），本质上是专断的。正如哲学家伊曼努尔·康德所强调的，人应当被视为目的本身，而非手段。剥夺个体理解影响其生活的决策逻辑的权利，就是将其工具化。因此，无论透明性是否能带来直接的工具性收益，它都具有不可让渡的内在价值。这一价值论困境直接影响了XAI的优先级设定和资源分配。在商业环境中，工具理性往往占据上风，XAI被视为一种风险管理或用户体验优化的成本。而在公共政策和伦理讨论中，价值理性的呼声则更为强烈。如何平衡这两种理性，决定了XAI究竟是沦为一种粉饰门面的“伦理洗白”（Ethics Washing）工具，还是真正成为捍卫数字时代公民权利的利器。

3 突破路径：走向语境化与生态化的解释框架

要真正突破透明性悖论，我们需要一场范式转换，从追求一种普适的、静态的、技术中心的“透明”，转向构建一种语境化的、动态的、社会-技术协同的“解释生态”。

3.1 拥抱“语境化解释”

如前所述，不存在放之四海而皆准的“好解释”。未来的XAI设计必须明确其解释契约（Explanation Contract）——即在特定应用场景下，为特定的利益相关者，提供满足其特定需求的解释。这要求：（1）受众建模：在设计解释系统前，深入理解目标用户的认知水平、知识背景、信息需求和情感状态。（2）目的驱动：清晰界定解释的目标是用于调试、审计、建立信任还是赋权申诉，并据此选择合适的解释类型（如反事实解释、示例解释、自然语言理由等）。（3）交互式解释：将解释视为一个对话过程，而非单向的信息输出。允许用户提问、质疑并引导解释的方向，使解释更具针对性和参与感。例如，在医疗诊断场景中，给放射科医生的解释可能是突出病灶区域的热力图和关键影像特征；而给患者的解释，则应是一段通俗易懂的、说明诊断依据和治疗建议的自然语言，并辅以相关的健康知识。在司法审判场景中，解释同样不应是单向的宣判，而是一个面向不同受众、允许对话与反馈的沟通过程。对于审判人员（法官），解释的核心在于可检验性与法律逻辑的严密性。此时，解释应像一份拆解精密的“思维导图”，允许法官随时“放大”查看细节，或对某一环节提出质疑。面向当事人的交互

式解释，他们的核心诉求往往不是“理解法律逻辑”，而是“我的处境会怎样”“为什么这样对我”“我还能做什么”。

3.2 超越“解释”，构建“问责”与“治理”生态

过分聚焦于技术层面的“解释”，可能会让我们忽视一个更重要的目标——问责（Accountability）。解释只是实现问责的众多手段之一。一个健全的AI治理体系，应当包含多层次的保障机制：（1）事前治理：通过严格的算法影响评估（AIA）、数据集审计和模型卡（Model Cards）制度，确保AI系统的设计符合伦理与法律规范。（2）事中监控：部署持续的性能监控和偏见检测系统，及时发现并干预模型的异常行为。（3）事后救济：建立清晰、便捷的申诉渠道和人工复核机制，确保当AI决策出错时，受影响者能够获得有效的救济。在这种生态中，XAI的角色被重新定位。它不再是唯一的“救世主”，而是整个问责链条中的一个关键环节，与其他治理工具协同工作[4]。例如，一个简洁的反事实解释（“如果你的收入再高5%，你的贷款申请就会被批准”）可以作为用户启动申诉程序的有力依据，而最终的裁决则依赖于人工审核和既定的规则。

3.3 人机协同：从“理解机器”到“与机器共事”

最后，我们必须调整对人机关系的期待。与其执着于彻底“理解”机器，不如致力于构建高效的人机协同（Human-AI Collaboration）模式。在这种模式下，AI被视为一个强大的、但有其局限性的合作伙伴。人类无需洞悉其全部奥秘，只需掌握足够的信息来判断何时信任它、何时质疑它、以及如何有效地与之互动。这要求XAI的设计不仅要提供解释，更要提供行动指南（Actionable Insights）。例如，在自动驾驶中，系统不仅应告知“我检测到前方有障碍物”，还应明确指示“我将采取减速并变道的措施，请您注意”。这种面向行动的、情境化的信息传递，比展示其卷积层的激活图更能有效支持人类的监督与接管。

4 结语

算法黑箱与透明性悖论，是人工智能时代给人类文明出的深刻难题，表面是技术问题，实则是哲学思辨。本文剖析XAI在认识论、本体论与价值论的困境，指出仅靠技术无法根除悖论。未来出路不是找打开所有黑箱的万能钥匙，而是学会与黑箱共处。要放弃绝对透明幻想，拥抱务实多元主义，承认解释的语境依赖性，构建技术、制度、法律和社会规范协同的综合性治理生态，重塑人机伙伴关系。如此，我们才能在享受AI红利时，守护人的主体性、尊严与社会公平正义，让算法成为服务人类福祉的可靠伙伴，引领我们走向更负责、更具韧性的智能未来。

参考文献

[1]袁曾.算法黑箱的治理迷思与破解[J].中国海商法研究,2025,36(04):22-31.

[2]陈磊.算法黑箱下大模型使用者刑事注意义务的冲突与重构——基于风险治理视角的刑法调适[J].政法学刊,2025,42(06):51-58.

[3]刘建,吴理财.算法治理的黑箱及规制：基于治理界面的视角[J].学习与实践,2025,(11):34-45.

[4]张卫涛.算法黑箱的成因、危机及治理路径选择[C]//中国智慧工程研究会.2025数字时代的社会结构变迁与治理创新学术交流会论文集（下）.中国人民公安大学;,2025:374-377.