装

订

线

北方工业大学

研究生课程考核论文（作业）

课 程 名 称 《工程伦理》

论文（作业）题目 AI 决策系统的“黑箱”困境与公众知情权：工程伦理的责任边界

:

学院名称	人工智能与计算机学院	专 业 年 级	智能专研 25
姓　　名	王耀彬	学　　号	2025354100103
任课教师	陈 首 珠	提 交 时 间	2025.11.15

论文（作业）基本内容、写作格式要求和成绩评定

(一)基本内容要求（教师编写）

请同学们根据《工程伦理》课程教学内容，结合自身所学的专业或感兴趣的研究主题，完成一篇 3000 字左右的有关工程与伦理方面的学术论文，含题目、中英文摘要、关键词、论文内容、参考文献等。

严禁抄袭、剽窃等学术不端行为！

（该页所有内容不要删除， 论文作业 A4 纸正反双面打印，左侧装订）

(二)写作格式要求

1、论文（作业）题目。用 3 号宋体，加粗，居中；

2、论文摘要与关键词。摘要用小 4 号宋体，加粗，居中；关键词用小 4 号宋体，加粗，左顶格，

3、论文分级标题。统一用 1，1.1，1.1.1 层次编写，左顶格。

4、论文开本。论文开本大小为 A4 纸；正文采用小 4 号宋体。

5、注释。引用他人的成果必须标明出处。所有引用过的文献，应按引用的顺序以脚注的格式每页独立顺序编号排列。

6、参考文献。参考文献格式参阅北方工业大学学报或学位论文要求。

(三) 成绩评定指标和评分(教师自定)

指标体系	最高分值	评分
选题价值及创新程度如何，分析是否透彻	40	
逻辑结构是否严密，研究方法是否得当	30	
摘要、关键词等要素是否都具备，是否存在语句错误	30	
总成绩	100	

评阅人签字：

AI 决策系统的“黑箱”困境与公众知情权：工程伦理的责任边界

王耀彬[1)]

[1)](北方工业大学人工智能与计算机学院，北京市石景山区晋元庄路 5 号，100144)

摘要

随着人工智能（AI）决策系统在司法、医疗、金融等高风险公共领域的广泛应用，其内部决策逻辑不透明的“黑箱”问题日益凸显。AI 系统的“黑箱”不仅是一个技术挑战，更是一个严峻的工程伦理问题。其直接挑战了公众在切身利益受影响时所享有的基本知情权。本文首先剖析了“黑箱”问题的技术与伦理双重维度，并以美国司法领域广泛使用的 COMPAS 算法为例，深度分析了其在实践中暴露的算法偏见与公平性争议。并探讨了医疗 AI 领域中可解释性的缺失如何侵蚀医患信任与医疗责任。在此基础上，本文提出一个面向工程伦理的“责任透明度”三维框架，该框架整合了技术可解释性（如 LIME、SHAP 等 XAI 工具）、流程规范性（如算法影响评估 AIA）与责任可追溯性（如“伦理黑匣子”与 IEEE 伦理设计准则），旨在为工程师提供一套可操作的伦理实践指南。本研究预期，通过实施此框架，能够增强 AI 系统的公信力，保障公众的合法权益，并最终将透明度确立为工程师在 AI 时代不可推卸的核心伦理责任。

关键词：工程伦理；人工智能；黑箱问题；算法偏见；可解释性；

1 引言

1.1 研究背景：算法决策时代的伦理困境

随着技术的进步与时代的发展，当今社会决策问题正逐渐全面转化为以算法为核心支撑。人工智能（AI）与机器学习模型正以前所未有的深度和广度渗透到社会运行的方方面面。在此基础上，各类基于数据驱动的模型在司法裁判、医疗诊断、金融信贷等高风险公共领域得到广泛应用，其输出结果直接影响个体的自由、健康与财产安全。从理论层面看，这类系统被寄予提升效率、增强客观性与保持一致性的期望。然而，与其功能扩展相伴随的，是日益突出的伦理难题：其内部决策机制往往呈现出高度不透明的“黑箱”特征。

此类不透明性主要源于深度学习等复杂模型内部庞大参数结构所决定的决策路径难以追溯，即便是系统的设计者也难以对其运行逻辑做出完整解释[1]。由此引发的风险包括：固化并放大训练数据中潜藏的历史性偏见，进而导致歧视性结果；在系统出现错误时，责任归属模糊，削弱了有效的问责机制；公众对自动化决策系统的信任受到严重冲击[2]。

由此可见，以复杂性换取准确性的技术范式，与公共领域所要求的透明性和可问责性之间，存在着根本性且系统性的张力。这一矛盾已超越单纯的技术问题，演变为亟待回应的工程伦理挑战。

1.2 案例引入与核心问题

为进一步凸显人工智能决策系统“黑箱”特性所引发的伦理困境，本文选取刑事司法与医疗健康两个典型应用场景作为案例分析。这些案例集中体现了算法不透明性对公共利益的直接冲击。

（1）刑事司法领域的 COMPAS(Correctional Offender Management Profiling for Alternative Sanctions)算法已在美国多个州的司法体系中被用于评估被告的再犯风险，其结果直接影响保释、量刑及假释等关键司法决策。然而，2016 年 ProPublica 的调查报道显示，该算法在预测中存在显著的种族偏差，对黑人被告的误判率明显高于白人被告，由此引发了关于算法公平性与司法正义的广泛争论[3]。

（2）在医疗健康领域，人工智能驱动的诊断工具在医学影像识别（如 X 光片、MRI）中展现出较高的应用潜力，但其“黑箱”特性同样引发伦理关切。当系统给出诊断或治疗建议时，若医生无法理解其背后的推理逻辑，将难以对结果进行验证或向患者做出合理解释。这不仅削弱了患者的知情同意权，也对医疗安全与责任划分提出了挑战。

基于上述案例，本文提出核心研究问题：当具有“黑箱”特征的人工智能决策系统对公共利益产生重大影响时，其不透明性是否构成对公众基本知情权的侵犯？工程师应如何将这一冲突重新界定，并将其纳入工程伦理责任的核心议题加以回应？

2 相关工作

2.1 算法偏见的实证研究：COMPAS 案例深度剖析

COMPAS 算法的争议为“黑箱”问题及其伦理后果提供了迄今为止最深刻和最具影响力的实证案例。

2016 年，ProPublica 发表了题为“机器偏见（Machine Bias）”的调查报告，通过分析佛罗里达州布劳沃德县超过 7000 名被告的 COMPAS 风险评分及其后两年的犯罪记录，得出了惊人的结论：该算法在预测错误上表现出显著的种族差异[3]。

报告的核心统计数据显示：

- 错误的“高风险”标签：黑人被告被错误地标记为“高风险”的比例（44.9%）几乎是白人被告（23.5%）的两倍。
- 错误的“低风险”标签：白人被告被错误地标记为“低风险”的比例（47.7%）远高于黑人被告（28.0%）[3]。

这些数据揭示了一个系统性问题：COMPAS 算法的错误倾向于对黑人被告更为不利，而对白人被告更为“宽容”。

随后的学术研究进一步揭示了一个更深层次的问题：当两个群体的基础再犯率（base rates）本身就存在差异时，任何算法在数学上都不可能同时满足“预测均等性”和“均等化赔率”这两个公平标准。这意味着，在一个反映了社会不平等的数据库中，“算法公平”并非一个可以被完美求解的客观技术问题，而是一个必须在不同类型的伤害之间做出权衡的伦理选择。工程师无法通过技术手段“消除”偏见，而只能选择让哪个群体来承担哪种类型的统计误差。这一发现从根本上将算法设计问题从技术领域推向了伦理与价值判断的前沿。“预测均等性”和“均等化赔率”的描述见下表：

表 2-1 “预测均等性”和“均等化赔率”

公平性指标	定义	支持方论点	关键影响/后果
预测均等性	对于任何给定的风险分数，黑人与白人被告的实际再犯概率是相同的。	Northpointe 公司：分数对所有种族都具有同等的预测准确性，是无偏的预测工具[4]。	接受了不平等的错误率，意味着某个群体将不成比例地承担更多被错误标记为“高风险”的代价[4]。
均等化赔率	不同种族群体在被错误预测为“高风险”和被错误预测为“低风险”上的比例应该是相等的。	ProPublica：防止任何一个群体因算法的错误而受到不同的惩罚或伤害，保障程序公平[3]。	导致一个给定的风险分数对不同群体意味着不同的实际再犯风险，牺牲了预测的一致性[5]

2.2 算法时代的知情权：从法律原则到实践挑战

面对算法决策的不透明性，保障公民的知情权已成为全球法律与政策关注的核心议题。其中最具代表性的体现是“解释权”。

欧盟《通用数据保护条例》（GDPR）在此方面率先提出相关要求，其条款暗示用户有权了解自动化决策的逻辑。近期通过的《人工智能法案》则进一步明确：对于列高风险人工智能系统（涵盖司法、就业、公共服务等领域）所做出的、对个人产生法律效力或类似重大影响的决策，受影响者有权从系统部署方获得“关于 AI 系统在决策程序中的作用及主要决策内容的清晰且有意义的解释”[6]。

然而，法律条文与工程实践之间存在显著落差。法律上的“解释权”并不能直接化解技术上的“黑箱”难题。实践中仍面临多重挑战：复杂神经网络的解释应如何界定为“有意义”；是提供庞大的参数细节，还是给出可能过度简化的类比；在涉及商业秘密时，又如何在保护知识产权的同时确保企业履行解释义务[7]。这种法律权利与技术能力之间的脱节，凸显了仅依赖法律框架不足以应对问题，亟需通过工程实践与伦理规范加以弥补。

3 本文贡献

3.1 工程伦理新范式：将透明度视为核心安全与责任要素

鉴于“黑箱”AI 在公共领域应用的深刻影响，本文主张，必须将“透明度”从一项技术期望提升为工程伦理的核心原则，其重要性应与传统的公共健康、安全和福祉等基本准则等同[2]。

这一主张与《工程伦理》课程大纲的核心精神高度契合。根据通论第二章“工程中的风险、安全与责任”，不透明的“黑箱”系统本身即构成重大风险，因其可能掩盖偏见、错误与失效模式。透明性因此成为工程师保障安全、履行责任的基本要求。分论第十章“信息与大数据伦理问题”进一步指出，在利用大数据训练并应用于决策时，提供透明与可解释性是开发者和使用者对公众应尽的核心伦理义务[8]。

本文所提出观点，主要思路来源于笔者近期的科研项目“AMR-SELOR：一种基于一种用于生成语义逻辑解释的神经符号框架”。属于人工智能领域，人工智能可解释性方向的工作。笔者正是从当中认识到了深度学习“可解释性”的重要性与困难点。由此为出发点，通过文献查阅，总和前人工作，得到本文提出的“责任透明度”三维框架。

3.1 “责任透明度”三维框架

为了将上述伦理原则转化为可操作的工程实践，本文提出一个包含三个层次的“责任透明度”框架。该框架并非单一解决方案，而是一个整合了技术工具、组织流程和治理保障的综合体系。

3.2.1 技术可解释性：工程师的工具箱

这一层次关注工程师可直接采用的技术手段，以揭示“黑箱”模型的内部运作机制。其核心是可解释性人工智能（Explainable AI, XAI），即一系列旨在提升模型决策过程可理解性的技术[9]。对于已建成的复杂模型，常用的“事后”解释方法主要包括两类工具：

• LIME (Local Interpretable Model-agnostic Explanations)：通过在单个预测实例周围生成扰动数据，并以简单可解释模型（如线性回归）拟合局部行为，从而识别该实例中各特征对预测结果的正负影响[10]。

• SHAP (SHapley Additive exPlanations)：基于博弈论的夏普利值，计算各特征在所有可能组合中的边际贡献，量化展示其对预测结果的具体推动作用[9]。

在工程实践中，XAI 工具应被视为高风险人工智能系统验证、审计与调试的必要组成部分。它们既有助于工程师发现和纠正模型偏差，也能为受影响个体提供事后解释，从而在一定程度上回应公众的知情权诉求[11]。

3.2.2 技术可解释性：工程师的工具箱

技术工具本身不足以确保伦理落实，必须嵌入规范化的组织流程之中。本层次强调，透明度应贯穿人工智能项目的全生命周期，而非事后补救。实现这一目标的关键机制是算法影响评估（Algorithmic Impact Assessment, AIA）[11]。

AIA 是一种结构化的风险评估流程，旨在系统识别、评估并记录人工智能系统可能带来的社会与伦理影响。AI Now 研究所提出的框架包括三个核心环节：机构内部自评以识别潜在风险，引入外部研究人员进行独立审查，以及在系统采购或部署之前向公众发布公告，说明用途、能力与已知风险，并征求公众意见[11]。

AIA 将伦理考量从技术人员的个体责任转化为组织层面的制度化义务，使风险管理和公众参与在潜在危害发生前得以落实。

3.2.3 责任可追溯性：治理的最后防线

考虑到商业秘密等因素可能限制完全的技术透明性，仍需建立能够在系统故障或争议发生后确保责任可追溯的治理机制。本层次引入了“伦理黑匣子”（Ethical Black Box）的概念[2]。

该概念借鉴了航空领域的飞行数据记录仪，其核心是嵌入人工智能系统的标准化数据记录模块，持续保存运行过程中的关键信息，包括输入数据、关键中间状态、输出结果及相关决策依据或置信度。这些数据在常规情况下保持加密与不可访问，仅在发生重大事故或争议时，由具备授权的独立审计或监管机构调取，用于事后调查与责任认定[2]。

“伦理黑匣子”为电气与电子工程师协会（IEEE）《伦理协同设计》所提出的核心原则提供了具体实现路径。其中，原则五（透明性）要求“人工智能系统特定决策的基础应始终可被发现”，原则六（问责制）要求“为所有决策提供明确的基本原理”。“伦理黑匣子”正是落实这一“可发现性”

与“可问责性”的关键机制，在保护企业知识产权与维护公共利益之间实现了务实平衡。

4 预期结果与展望

采纳并实施本文提出的“责任透明度”三维框架，有望在多个层面产生深远影响。

• 对工程师及工程行业而言：该框架将抽象的伦理原则转化为可操作的实践路径。通过工具（XAI）、流程（AIA）与标准（伦理黑匣子及 IEEE 准则）的结合，工程师能够在日常工作中系统评估与缓解风险，从而强化其在公共安全中的专业责任。

• 对公众与受 AI 决策影响的个体而言，该框架为基本权利提供制度化保障。强制性的信息公开、解释机制与问责渠道，有助于建立公众对 AI 系统的信任，并确保知情权、公平权与救济权的有效实现。

• 对社会整体而言，该框架的推广有助于引导 AI 技术走向公正与负责任的发展方向。通过制度化设计抑制算法偏见与错误风险，推动以人为本的创新文化，确保技术发展服务于增进人类福祉的目标。

然而，实施过程中仍面临挑战，包括企业对商业秘密公开的抵触、AIA 与伦理黑匣子带来的额外成本、法律与行业标准的滞后，以及工程师跨学科伦理培训的不足。

参考文献

[1] LONDON A J. Artificial intelligence and black-box medical decisions: accuracy versus explainability[J]. Hastings Center Report, 2019, 49(1):15-21.

[2] IEEE. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition[S]. 2019.

[3] ANGWIN J, LARSON J, MATTU S, KIRCHNER L. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks[EB/OL]. ProPublica, 2016.

[4] CHOULDECHOVA A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments[J]. Big Data, 2017, 5(2):153-163.

[5] KLEINBERG J, MULLAINATHAN S, RAGHAVAN M. Inherent trade-offs in the fair determination of risk scores[EB/OL]. arXiv:1609.05807, 2016.

[6] EUROPEAN COMMISSION. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)[S]. 2021.

[7] GOODMAN B, FLAXMAN S. European Union regulations on algorithmic decision-making and a "right to explanation"[J]. AI Magazine, 2017, 38(3):50-57.

[8] ZWITTER A. Big data ethics[J]. Big Data & Society, 2014, 1(2):2053951714559253.

[9] 9LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//Advances in Neural Information Processing Systems. 2017:4765-4774.

[10] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?": Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1135-1144.

[11] REISMAN D, SCHULTZ J, CRAWFORD K, WHITTAKER M. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability[R]. AI Now Institute, 2018.