Advances in Social Sciences 社会科学前沿, 2025, 14(11), 206-213
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2025.1411988

Hans汉斯

人工智能可解释性的技术困境与法律消解

王奕翔

暨南大学法学院/知识产权学院，广东 广州

收稿日期：2025年9月22日；录用日期：2025年10月28日；发布日期：2025年11月7日

摘　要

人工智能技术的持续突破，正在以颠覆性的力量重塑全球科技创新与产业经济发展格局。然而，人工智能的迅速发展也带来了诸多风险，其中人工智能的“算法黑箱”现象和不确定性导致人们无法理解人工智能的决策和输出，从而引发信任危机和归责困境，很大程度上阻碍了人工智能产业的发展。本文从人工智能的“算法黑箱”本质出发分析其可解释性问题，分别从技术角度、用户信任和法律价值三个角度论证对人工智能可解释性进行法律规制的必要性和可行性。在明确可解释性和透明度的定义和区别的基础上，将可解释性分为面向用户和专业人员两方面的可解释性，从可追溯、可视化、反事实解释三个角度构建面向用户的可解释性要求，提出根据领域分类的“硬法”“软法”可解释性要求，以法律引导、促进人工智能技术的安全、可靠、可信任的发展。

关键词

人工智能，可解释性，算法黑箱

The Technical Dilemma of Artificial Intelligence Explainability and Legal Resolution

Yixiang Wang

Law School & Intellectual Property School, Jinan University, Guangzhou Guangdong

Received: September 22, 2025; accepted: October 28, 2025; published: November 7, 2025

Abstract

The continuous breakthroughs in artificial intelligence technology are reshaping the global pattern of technological innovation and industrial economic development with disruptive power. However, the rapid development of artificial intelligence has also brought many risks, among which the

文章引用: 王奕翔. 人工智能可解释性的技术困境与法律消解[J]. 社会科学前沿, 2025, 14(11): 206-213.
DOI: 10.12677/ass.2025.1411988

phenomenon of the "algorithmic black box" and uncertainty lead to people's inability to understand the decisions and outputs of artificial intelligence, thus triggering a crisis of trust and accountability dilemma, which largely hinders the development of the artificial intelligence industry. This paper analyzes the interpretability issues starting from the essence of the "algorithmic black box" of artificial intelligence, demonstrating the necessity and feasibility of legal regulation of artificial intelligence interpretability from three perspectives: technological perspective, user trust, and legal value. Based on clarifying the definitions and distinctions of interpretability and transparency, interpretability is divided into user-facing and professional interpretability. From the three angles of traceability, visualization, and counterfactual explanation, user-facing interpretability requirements are constructed, proposing "hard law" and "soft law" interpretability requirements categorized by fields to guide and promote the safe, reliable, and trustworthy development of artificial intelligence technology through law.

Keywords

Artificial Intelligence, Explainability, Algorithmic Black Box

1. 引言

随着我国人工智能产业的发展，DeepSeek、豆包等人工智能产品如雨后春笋般冒出来，但是随之而来的是人工智能带来的各种问题。在各种人工智能产品的使用过程中，人们发现，人工智能很容易胡说八道，由于训练数据中的虚假信息的影响，以及人工智能算法本身的特性，会导致人工智能"幻觉"的出现，从而引发各种风险。可解释性最早由人工智能之父约翰·麦卡锡(John McCarthy)提出，其在 1956 年达特茅斯会议上首次提出了"人工智能"这一术语，并提出了著名的"人工智能的四个基本要求"，即可理解性、可表达性、可学习性和可适应性。麦卡锡在 1965 年提出了"公平合理化"的概念，认为人工智能系统应该能够以允许人类理解和评估它们的方式解释它们的决定。这一早期发展为人工智能算法的理解和解释的研究奠定了基础。自 20 世纪 50 年代早期开始，对人工智能的可解释性研究已经走过了漫长的道路。各国也在寻求通过规范性文件对人工智能的可解释性作出规定。欧盟《人工智能法案》(EU AI Act)第 13 条提出透明度义务和信息提供义务，规定高风险人工智能系统的设计和开发应当使用户能够解释系统的输出并加以适当适用，确保用户理解系统如何得出结果；美国国家标准与技术研究所(National Institute of Standards and Technology, NIST)发布的《可解释人工智能四原则》(Four Principles of Explainable Artificial Intelligence)，建议打算或需要解释的人工智能系统遵循四个原则：解释性，即对于输出和/或过程而言，系统提供或者包含伴随的证据或理由；有效性，即系统所提供的解释对于预期的接收者来说可以理解；解释的准确性，即解释能够正确反映输出的原因和/或准确反映系统的输出过程；知识局限性，即系统仅在其设定的条件(环境)下运行，并且仅在其对输出达到足够的信心时运行。

可解释性无疑对人工智能的发展有着重要的影响，目前虽然学界对可解释性讨论颇多，并且也有国际上对可解释性的法律规制实践，但可解释性的问题仍然处于未解决的状态。首先，可解释性一定程度上是技术方面的问题，对人工智能的解释是否有技术上的可行性？如果技术上无法对人工智能的行为进行解释，那么法律的规定就成为无稽之谈；当人工智能的解释具有技术上的可行性，它能够解释到什么程度？而是否需要通过法律去规定可解释性要求仍然存在疑问，同时法律与技术存在的壁垒导致很难去

确定解释的程度。因此，可解释性的法律问题需要从技术和法律两个层面去分析。

2. 人工智能可解释性的技术困境

2.1. 深度神经网络的算法黑箱与算法歧视

在大数据时代的环境下，人工智能技术正在迅速发展。与传统的计算机模型这种具有明确数量关系的算法解不同，如今的人工智能技术通过对数据间相关特征的拟合来完成复杂推理的输出，并通过对人脑活动的模拟进行深入，从而导致模型越来越复杂[1]。机器学习是人工智能的一个子领域，而深度学习是一种机器学习的分支，使用多层神经网络模型来进行学习和推理，以海量的数据为基础运行。深度学习在计算机视觉、自然语言处理等领域获得了巨大的成功，但可解释性一直是深度学习的一个弱点，深度学习复杂性高、参数多，人们一般无法解释这种“端到端”模型做出决策的原理，也无法判断决策的可靠性[2]。深度神经网络模型基于海量的数据以及由海量参数构成的复杂非线性结构构成。大模型是深度学习中的大型神经网络模型，它们通常包含数亿甚至数十亿个参数，可以处理海量的数据，具有强大的特征表达和推理能力。人工智能大模型处理的数据规模巨大，数据维度复杂多变，其模型本身也高度复杂。实际上，在计算机领域，一个计算机程序涉及的代码就已经是以“数量级”来形容，传统的算法程序，其代码如果未加注释，即使是相同领域的专业人员想要理解该算法的内容也已经非常困难，人工智能大模型由于数据规模庞大以及其模型本身的高度复杂，使得即使是专业人员也难以理解其模型运转原理和行为逻辑。并且，对于复杂模型，无法通过简单的局部线性逼近而是需要关注其模型结构和高维非线性关系[3]。

基于人工智能大模型的复杂结构所导致的输入与输出之间映射过程高度不透明、难以追溯，如同一个无法窥视的“黑箱”(Black Box)。算法黑箱导致人工智能的决策或预测行为的逻辑无法被人类所直接理解，人们只能观察到模型的输入和输出，对于模型为什么对这个输入产生特定的输出，完全无从得知。

2.2. 复杂系统导致的不确定性

复杂系统(complex system)指难以用还原论方法处理的各种系统的总括，泛指规模大、变量多、各部分关系和运动规律复杂或具有较大不确定性的系统[4]。即使有一个非常简单的系统，要想对其所有的细节进行完整的描述也是不可能的[5]，对于复杂系统来说更甚。复杂系统通常由大量相互作用的组成部分构成，同时这些组成部分之间也存在着复杂的相互关系和动态变化，其相互作用是非线性的，这将导致系统的突然变化和不可预测行为，从而加剧了解释的难度。现实中的人工智能系统通常并非非孤立模型，而是由多模块、多模型、动态数据流构成的复杂系统(如自动驾驶感知－决策－控制链)。系统级行为源于各组件间的非线性、实时交互。这引发的问题是，一是组合爆炸与混沌效应，组件交互可能产生设计者未预期的“涌现行为”(Emergent Behavior)，其因果关系链极其复杂甚至不可追踪；二是环境依赖性与动态演化，系统行为高度依赖实时输入数据和环境状态，且模型自身可能在线更新(Online Learning)，例如Chat GPT，其模型在与用户交互的过程中也一直在实时更新和学习。这使得对单一决策的解释难以复制和泛化，系统行为呈现本质不确定性(Inherent Uncertainty)。

对于这种人工智能系统的解释好比对人脑活动与决策的解释，譬如让一个人回答今天晚上吃什么的问题，他可能会立即得出一个吃火锅的结论，但是为何会产生这样的结论，其原因是完全不可解释的，对其进行追根溯源最终会回到神经元的电信号之中，而人脑的复杂结构让我们也无法得知为什么针对这个问题人脑会给出这样一个结论，神经元的电信号是如何产生又是为何会产生的，完全不可知也不可解释。并且，这种回答是完全随机的，如果你在第二天的同一时间同一地点问同一个人相同的问题，很可能又会得出完全不同的答案。

2.3. 可解释性技术的专业性和可视化难题

即使通过技术手段(如显著性图、反事实解释、概念激活向量等)生成解释信息，其有效传达与用户理解仍是巨大挑战。人工智能的专业性使得专业人员与用户之间存在认知鸿沟，对于人工智能输出的技术性解释(如特征权重、梯度图)对普通用户而言晦涩难懂，由于目前人工智能产品的用户范围过于广泛，其应用涵盖社会生活的各个方面，面向的用户绝大多数并没有计算机的专业背景，对于技术性解释无法理解。然而，对人工智能输出的简化解释也有一定风险。

解释也存在可视化瓶颈，将高维、抽象的模型决策过程转化为直观、准确的、普通用户能够理解的信息。既要避免过度简化导致信息失真，又要防止过于复杂引发认知负荷。尤其在处理图像、文本等多模态数据时，可视化难度陡增。

解释还存在解释保真度与效用权衡的问题。如果追求解释的完全准确(高保真度)可能导致解释本身过于复杂，如冗长的逻辑树等，不仅会降低用户理解度，还极大增加了解释成本。反之，追求用户友好与效能，常常需损耗部分准确性(低保真度)。

技术困境的根源在于，AI 系统(尤其是深度学习)的运作本质上是基于统计模式识别而非符号逻辑推理。其优势在于处理海量模糊信息，劣势恰在于其决策过程缺乏人类可直观理解的清晰逻辑链条。强求达到人类决策式的、完全透明的“白箱”解释，在当前及可预见的技术路径下，面临根本性困难。

面对 AI 系统的可解释性问题，目前有两种类型的应对方法，可解释性总体上可以分为 2 类，一种是事前(ante-hoc)可解释性和事后(post-hoc)可解释性[6]。事前可解释性指模型本身具有可解释性，其模型通常为结构简单、易于理解的如决策树、线性回归等简单模型，这些结构简单的模型具有高可解释性和透明度，但其模型拟合能力有较大的限制，准确性与深度神经网络等复杂模型相比有较大差距。

复杂模型的高准确性往往伴随着严重的黑箱问题，而事后可解释性大多针对的是复杂模型。事后可解释性通过利用解释方法或解释模型对给定的已经训练完毕的模型的工作机制、决策行为和决策依据进行解释[6]。LIME 模型与 SHAP 模型是两种常用的事后可解释模型。LIME 模型是一种局部代理模型，其核心思想是在复杂模型的一个局部区域内，通过采样和权重分配，构建一个简单模型来近似复杂模型的行为。LIME 模型具有易于理解、能够适用于任何机器学习模型的优势，然而这种可解释模型会忽略特征之间的相关性，从而导致解释具有不稳定性，两个相近的样本点可能出现完全不同的解释[7]。SHAP 模型的核心思想是通过考虑每一个特征量对模型输出结果的贡献，将预测结果分解为每个特征的影响，提供全局和局部的可解释性。SHAP 模型与 LIME 模型一样具有广泛的应用范围，能够解释几乎任何类型的机器学习模型，并能同时提供针对单次预测的局部解释与针对整个模型全局解释，并有可视化功能较强的优点，但在面对大量数据与较多特征数量时，精确计算将耗费大量时间，同时特征数量较多将导致可视化结果极多而难以解读，面对高度相关的特征量时，将会导致无法直观和准确地解读其结果[8]。目前可解释性工具的发展虽然呈现出百花齐放的态势，并未能完全解决可解释性的问题。

3. 可解释性的法律规制理论基础和现实考量

3.1. 可解释性是可信任的人工智能的基础

随着人工智能技术的快速发展，其带来的机遇与引发的风险并存，而对于人工智能领域的治理围绕着确保人工智能安全、可靠、可控进行。如何增强人工智能使用者的信心，是关系到人工智能长远发展的重要议题。面对人工智能引发的信任焦虑，发展可信任的人工智能早已成为全球共识。早在 2018 年，国际计算机学会 ACM (Association for Computing Machinery)就发起会议 ACM FACCT (ACM Conference on Fairness, Accountability, and Transparency)，围绕机器学习的公平性、可问责和透明性展开探讨；而 2019

年中国科技部发布《新一代人工智能治理原则——发展负责任的人工智能》，突出发展负责任的人工智能的主题；而美国也在 2022 年推出《算法问责法案》。人工智能作为一把双刃剑，不能任由其野蛮生长，而是需要推动和引导，而可解释性就是可信任的人工智能的核心问题之一。

诚然，用户对于一类产品的信任取决于许多因素，但最重要的因素还是这个产品能否完全实现其目的，对于人工智能产品的信任同理，当 DeepSeek 出现“AI 幻觉”、输出虚假信息的次数逐渐增多时，人们对其信任度就已经大幅下降了。对于人工智能产品的信任实际上只需要让其在回答一百个问题时都能够给出可接受的答案就可以了，但目前从技术上无法避免“AI 幻觉”的发生，其输出内容通常会出现各种各样的错误，这是人工智能技术的本质造成的，其归根到底是基于文字之间的关联和概率进行输出，即靠“猜”。以大语言模型为例，人工智能其实并不能理解文字中的含义，它能够输出一篇文章、或作一首诗，亦或是回答用户的问题，实际上只是模型中根据用户的输入，“猜”这个输出的可能性更大而已。因此，对于人工智能产品的信任，当无法做到百分之百的正确的时候，就需要对其输出作出解释，从而让用户能够判断其输出的正确性，才能使用户更加信任人工智能。这其实与医学判断极为类似，医生对于病因的判断并不能做到一定正确，但通过对其判断的解释，通常也能够说服患者从而让患者对医生的判断信服。人工智能的输出同样需要解释才能让用户信任。

另一方面，人工智能的“算法黑箱”特性使其运行逻辑、输出过程不能让用户完全理解，其输出的不确定性，特别是其出乎意料的输出内容也容易极大降低用户对人工智能产品的信任度[9]。

3.2. 法律介入以寻求价值平衡和构建行为框架

针对人工智能的“算法黑箱”以及不确定性问题，学界早就有关于其法律规制的探讨，其引发的各种风险也很早就被各界所意识到，在国外也已经有关于人工智能可解释性的各种规范性文件，但其反对声音主要在于法律的事前规制需要保持一定谦抑性，在面对算法规制的过程中可能引发对市场自由竞争的限制、对技术革新进步的损害[9]。但法律并非试图替代技术解决所有解释难题，而是能够通过设定功能目标，基于保护法益(权利、安全、公平)的需求，明确解释应达到何种功能效果(如“使用户理解决策主因并质疑”、“使监管者验证合规”)，而非强行规定具体技术路径，为技术创新留出空间。同时，法律可以通过建立行为规范与责任预期的方式，通过设定义务(如“提供有意义解释”)、标准(如“解释应清晰、及时、相关”)和违反后果，法律能有效引导开发者、部署者投入资源研发和应用解释技术，形成持续改进的动力机制。法律通过制度设计为技术发展留下足够空间并引导发展方向，同样也是对创新的一种制度性鼓励[10]。

客观而言，人类也不具有完全理解算法内部运作机理的能力。对可解释性进行规制并非要求开发者将人工智能的决策过程完全公开，也并非要求向用户完全解释人工智能的运行机理，而是针对人工智能的单个决策进行解释，向用户提供决策的原由和改变的方法。法律从此角度介入，则将“打破黑箱”转向“决策解释”，一是从技术角度而言，避免一味苛求“打破黑箱”而导致法律规定与技术实践脱节；二是明确法律以人为本的价值理念，平衡用户权益保障和技术发展。

4. 可解释性的定义明确与法律规制路径

4.1. 法律语境下可解释性的定义与关键区分

学界对于“可解释性”的定义通常与“透明度”进行混淆。苏宇认为可解释性部分关系到算法设计本身，透明度则主要关系到后续的解释与说明[11]。在他的观点中，可解释性是指算法满足既是决策者的精确代理，又能为人类所理解；而透明度则是要求以一定的方式和程度向用户或公众说明自动化决策的内在逻辑，尤其是解释用户或公众所关心的特定因素对算法决策的具体影响。也有学者指出，从人工智

能诞生之初，可解释性便作为人工智能透明度的一项重要指标，将可解释性作为人工智能透明度的一部分[5]。周辉通过比对联合国教科文组织、美国国家技术标准研究院(NIST)、中国国家人工智能标准化总体组和全国信标委人工智能分委会对可解释性的定义，认为人工智能可解释旨在阐明人工智能系统的决策过程和结果，从而建立信任和促进负责任的使用，体现了社会对人工智能系统透明度和可控性的期待[12]。孙晋则是不对算法透明度和可解释性作区分，而是直接结合两者讨论算法治理的制度构建，认为当前算法透明度和可解释性的起点在于公开，并围绕公开构建两条内外依托、协同治理的算法优化路径[10]。《澳大利亚的人工智能伦理框架》(Artificial Intelligence Australia's Ethics Framework A Discussion Paper)中也将透明度和可解释性放到一起，定义为向公众告知何时使用算法、算法会怎样影响其权利。可以看出，学界对于算法透明度和可解释性的定义是模糊的，透明度与可解释性这两个概念很容易混淆到一起。

确定可解释性与透明度的区别，可以聚焦于可解释人工智能的定义与国际上对于透明度的定义。可解释人工智能的典型例子是以决策树为基础的人工智能产品，决策树属于典型的线性回归模型，能够为人类所直接认识和理解。而关于透明度的规定，加州的透明度法案规定的是企业使用人工智能产品与用户进行交互时必须清晰明显地披露出人工智能的存在和使用。因此，可解释性解决的是模型算法为什么作出如此决策和预测的问题，而透明度解决的是模型算法相关的信息披露的问题。透明度意味着使用户知道在什么情况下和什么程序中使用这些建议，可解释性反映的是输入端的规则以及输入数据与输出决策上的决定之间能够形成清晰易懂的联系[13]。法律上应当对此进行明确规定。

在法律语境下，透明度应当聚焦于大模型算法整体的开放性，旨在使人工智能大模型的运行算法的公开透明，为人所知。算法模型的透明度应当指向系统整体信息的获取，包括系统的运行逻辑、开发设计原则、训练数据概况、性能局限、运维方信息等，旨在人工智能算法的算法运行逻辑和运行过程公开透明。

可解释性指的是为机器的输出寻找合理解释，因此其包含这样的期望：即决策或建议应说明管理系统的规则并展示背后参考的事件[13]。在法律语境下，可解释性应当聚焦于“面向用户的功能性解释”，让用户或受影响方理解单个决策如何产生，针对具体决策的逻辑与原因的可理解性。AI 系统的提供者有义务向受其输出影响的用户或主体，提供关于该系统特定决策或行为之关键原因、主要依据或运行逻辑的清晰、简洁、及时的信息，旨在帮助其理解该结果如何产生，并支持其据此做出后续判断或行动。可解释性所解决的问题就是人工智能决策或预测无法理解的问题，其并不在于打开黑箱揭示其内在运行原理，而是为其特定输出探寻输出的依据和解释，并且使得使用者能够理解。

4.2. 明确人工智能的决策应当解释到何种程度

国外对“可解释性”的定义虽然并没有形成通说，但也都趋向于两个侧重点，即侧重强调面向公众的可解释性和侧重技术上实现算法的可解释[14]。出现这种现象的原因在于，人工智能技术具有相当的专业性，而产品却是面向社会公众，用户和开发者对于人工智能技术的认知必然会有较大的差距，自然人工智能的可解释性要求无法做到折中，而需要从两个维度去进行解析。因此，对于可解释性的要求应当分为用户层面和专业人员层面。对于人工智能的可解释性要求应当更加关注于用户层面，用户不具备专业知识，在使用人工智能产品时承担着更大的风险，法律对于可解释性的要求应当聚焦于使用户能够理解人工智能的输出或决策，专业人员具有人工智能的相关知识，对于普通用户都能理解的解释当然能够更容易去理解，而对于更加深层的算法运行逻辑和机制，由于可能涉及到知识产权和商业秘密，也不宜增加对普通用户和开发者以外的专业技术人员的公开和解释义务。

在用户层面，可解释性应当在于人工智能的决策应当有解释能够使用户明白该决策的依据和基础，从而进行判断自己是否应当信任该判断和依据。面向用户的解释应当是符合一般用户的个体认知、理解

能力、知识储备的，不能要求用户拥有能够看懂算法代码和专业术语的能力。笔者认为，可解释性要求应当分为以下三点：一是可追溯性，即人工智能的决策应当有明确具体的原因；二是可视化，即人工智能决策的原因应当以用户可理解的方式呈现；三是提供反事实解释，即提供“如何改变决策”的方式。从而使用户能够知悉“为何得出该结果”“为何作出此决策”，并且能够对不满意的决策采取有效的措施。例如，在银行 AI 拒绝信贷申请之后，应当向用户解释拒绝的原因——当前负债收入比为 65%，而阈值为小于 50%；近六个月新开 4 张信用卡。实际上，人们很多时候并不在乎原因是如何得出结果，而仅仅是需要原因的信息。医生在给病人进行诊断时也不可能向病人完全解释为何出现如此症状则表明是该种疾病，反而仅仅是向患者表明之所以得出该诊断结果是因为出现如此症状。即使医生向患者解释症状和疾病之间关联性的原理，患者由于不具备专业的医学知识，也无法理解这个解释。因此，面向用户的可解释性的关键要求在于向用户提供可视化的“因”，并在合理的范围内提供“因”为何得出“果”。

2025 年 9 月，汉堡数据保护和信息自由委员会(Hamburgische Beauftragte für Datenschutz und Informationsfreiheit, HmbBfDI)在其发布的中期报告中，发布了 2025 年以来的 15 起行政违法诉讼案件，其中对一家金融行业公司处以 492,000 欧元的罚款，原因为该公司在个别案件的自动决策中侵犯了受影响客户的权利[15]。该公司的客户在申请信用卡时被使用自动决策拒绝，尽管其信用评级良好。当受影响客户随后要求被拒绝的申请提供理由时，该公司并没有充分履行其法定信息和披露义务。在这个情况下，用户当然并不是要求公司对其使用的自动决策算法的运行逻辑与输出过程完全披露和解释，况且用户也并不一定能够看懂这些专业性较高的内容；反而，用户对被拒绝申请的诉求仅仅是该算法在其信用评级良好的情况下根据其他何种指标作出的拒绝申请，其最终目的是能够根据这些指标采取行动进而获得申请信用卡的资格。

4.3. 基于领域的可解释性原则

欧盟《人工智能法案》将人工智能以风险等级作为区分从而对人工智能的相关主体施加一系列的义务和要求，其中将人工智能分为不可接受的风险、高风险、有限风险和轻微风险四种类型，针对不同类型施加了不同的监管措施。其中将高风险人工智能系统以领域列明，分别是生物识别相关、关键基础设施相关、教育和职业培训相关、就业相关、私人服务和公共福利相关、执法相关、边境控制相关、司法和民主进程管理相关的人工智能系统，对这些类型的人工智能系统规定了从入市前到入市后的全流程风险管理措施，以预防并控制这类人工智能系统对人类安全产生负面影响。

实际上，对于高风险人工智能系统的规制本质上还是针对特定领域的人工智能系统提出针对性要求。对于人工智能的可解释性要求，我国可不必效仿欧盟通过风险等级进行分类，反而可以对于不同领域，以“硬法”“软法”结合，基于不同领域的不同要求提出针对性的、不同约束力的可解释性要求。对于与公民的基本权利和利益有密切关系的领域，例如教育、医疗、司法、就业、社会福利等关键领域，应当规定强行性的可解释性要求，明确在这些领域中投入使用的人工智能系统必须具备上文所述的可解释性，即能够提供决策的原因、决策原因可被一般用户理解、提供反事实解释，从而能够让用户了解人工智能的决策为何作出，并能够采取措施防止自身利益受到损害。而对于其他人工智能系统如聊天机器人、文字和图片识别以及生成软件等，则可通过指南性质的“软法”进行规定，可以借鉴美国的《可解释人工智能四原则》，不对这些类型的企业设定强行性要求，而是允许企业自愿遵守可解释性要求，并鼓励行业自愿建立各行业的行为准则和标准。

5. 结语

人工智能技术的飞速发展带来的问题已经不容小觑，从技术角度来看，短时间内“算法黑箱”无法

打破，人类也不可能完全认识和理解人工智能系统的运行机理，同时法律也不可能要求每一个人工智能产品的用户都具有人工智能的相关专业知识。因此在法律层面上对开发者施加人工智能系统运行机理的解释义务无疑是强人所难。反而法律应当从用户层面着手，明确用户所需要的解释是何种解释，将透明度与可解释性区分开来，根据用户对人工智能产品决策的解释需求作为可解释性要求的着力点，并以“硬法”对公众利益密切相关的领域设立可解释性要求，同时对其他领域以“软法”进行规制，构建可解释性的规制体系，为人工智能的发展保驾护航，实现人工智能的安全可靠发展。

参考文献

[1] 谢佛荣, 景海龙, 邓菲菲. 人工智能的可解释性问题探析——基于因果关系视角[J]. 南华大学学报(社会科学版), 2022, 23(4): 40-45.

[2] 张吉祥, 张祥森, 武长旭, 赵增顺. 知识图谱构建技术综述[J]. 计算机工程, 2022, 48(3): 23-37.

[3] 尹艳霞. 大数据环境下机器学习模型的可解释性研究[J]. 电脑知识与技术, 2025, 21(5): 58-60+63.

[4] 《数学辞海》编辑委员会. 数学辞海[M]. 太原: 山西教育出版社, 2002.

[5] Siegenfeld, A.F. and Bar-Yam, Y. (2020) An Introduction to Complex Systems Science and Its Applications. *Complexity*, **2020**, 1-16. https://doi.org/10.1155/2020/6105872

[6] 纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096.

[7] taoKingRead. 模型可解释性-LIME [EB/OL]. https://blog.csdn.net/iqdutao/article/details/108397239, 2025-10-09.

[8] RessMatthew. 一篇入门: 彻底搞懂模型解释神器 SHAP 的核心原理与实践[EB/OL]. https://zhuanlan.zhihu.com/p/1926315437081228126, 2025-10-09.

[9] 金龙君. 生成式 AI 的不可解释性及其法治应对[J]. 法治研究, 2025(2): 42-53.

[10] 孙晋, 顾瑞琪. 基于算法透明度和可解释性优化的算法治理制度构建[J]. 数字法治, 2023(2): 64-75.

[11] 苏宇. 优化算法可解释性及透明度义务之诠释与展开[J]. 法律科学(西北政法大学学报), 2022, 40(1): 133-141.

[12] 周辉. 人工智能可解释的制度建构[J]. 山东师范大学学报(社会科学版), 2025, 70(1): 94-106.

[13] Ződi, Z. (2022) Algorithmic Explainability and Legal Reasoning. *The Theory and Practice of Legislation*, **10**, 67-92. https://doi.org/10.1080/20508840.2022.2033945

[14] 郑飞, 朱溯蓉. 人工智能“可解释性”的两个维度及其适用[J]. 大连理工大学学报(社会科学版), 2025, 46(2): 80-87.

[15] Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit (2025) Zwischenbilanz 2025: HmbBfDI verhängt Bußgelder von insgesamt 775.000 Euro. https://datenschutz-hamburg.de/news/zwischenbilanz-2025-hmbbfdi-verhaengt-bussgelder-von-insgesamt-775000-euro