人工智能算法的认识不透明性困境及其治理研究

李小芬

摘要：人工智能技术的快速发展与多方面突破，正在成为推动人类社会发展的新引擎和“拓荒者”，也成为网络空间安全发展的一个新的关键变量。智能技术的触角已经延伸到世界的各个“神经末梢”，同时也带来了算法困在“算法”里的认识不透明性问题。人工智能算法的认识不透明性是人类无法理解、还原和审查计算机模型内部运作和做出决策的全部过程。治理风险表现为：技术实现上的不透明性导致解释困难、数据不透明引发个人信息安全和不透明算法偏见加剧社会不公平现象。通过增加算法透明度，着力提升可解释性；推进算法公开，提升隐私保护技术；构建教智人类命运共同体，实现人工智能协同共治，以此提升人工智能治理新效能。

关键词：人工智能；DeepSeek；算法；认识不透明性；治理

中图分类号：TP18 **文献标识码：**A **文章编号：**1673-5706(2025)01-0050-08

随着人工智能技术的不断发展，尤其是以DeepSeek为代表的大模型横空出世，强大的数据处理能力、数据分析能力以及学习能力让世人感到惊叹。人工智能技术的惊人能力正在深刻改变人类社会，成为推动人类社会发展的新引擎和“拓荒者”，也成为网络空间安全发展的一个新的关键变量。一方面，人工智能技术在各个领域的广泛应用给人类社会带来了极大的社会和经济价值；另一方面，人工智能技术的发展趋势对技术治理的转型效率与时效提出了更高的要求，特别是人工智能算法的认识不透明性更是增加了治理的难度。基于深度学习模型的人工智能系统被广泛应用于各个领域，却由于不透明性导致了信任问题^[1]。人工智能的透

明度议题在人工智能技术治理中占据重要位置，成为目前人工智能治理的焦点。人工智能算法的认识不透明性“黑箱”模型和先进工业发展不可阻挡的“机化”，两者的结合直接引发了智能技术可靠性和可控性的诘难。对人工智能的神奇功能，众多学科领域正在进行如火如荼的研究，基于人工智能算法的认识不透明性对人工智能治理路径加以考察，无疑是一个不可忽视的新视角。同时，坚持发展与治理并举是全面提升人工智能治理水平的有力抓手。

一、认识不透明性的概念界定

认识不透明性不是一个二元属性，而是具有程度之分的，大体可以分为认识不透明性（EO）和本质认识不透明性（EEO）^[2]。不透明对应的英

文是 opacity, 意为不透明的质量、状态或不透明的东西, 也有“不透明度”的意思, 可追溯到拉丁语 opacitas 一词, 意为“黑暗”^[3]。“认识不透明”一词指的是查询者可能无法获得某一过程实现其应该实现的任务的全部细节和内部运行机制。认识不透明“将认识的主体、认识的客体、认识的过程和方法等各种因素与各个环节纳入所考虑问题的范围之内”^[4]。通常指认识论主体在解释系统, 如计算方法和智能机器不透明时存在局限性, 无法获得有关其运作的细节。

保罗·汉弗莱斯 (Paul Humphreys), 在 2004 年将认识不透明性应用到计算机模拟。虽然汉弗莱斯的工作主要涉及高能物理和分子生物学等非科学学科的计算机模拟, 但他对“不透明”的分析也成为讨论人工智能黑盒问题的一个有用起点。2009 年汉弗莱斯又基于区分认识不透明性的强弱程度提出“部分认识不透明性 (epistemically opacity)”和“完全认识不透明性 (fully epistemically opacity)”两个类型。部分认识不透明性就是如果代理主体 X 在 t 时不知道该过程在认识上的所有相关要素, 那么相对于认知代理 X 而言, 该过程在 t 时就是认识不透明的; 完全认识不透明性就是一个过程对 X 来说本质上是认识不透明的, 当且仅当考虑到 X 的性质, X 不可能访问并能审查被证明的所有相关要素, 即本质认识不透明性^[5]。从汉弗莱斯这种科学方法的计算机模拟不难看出, 对于大部分的计算模拟而言人工智能具有认识的不透明性, 因为机器学习模型和研究底层系统都是复杂和难以理解的。面对庞大的计算量, 认识不透明性主要关涉计算系统的复杂性, 这决定了人类不可能审查全部的计算过程。作为一个统计系统, 复杂性产生的原因有很多, 主要包括非线性和不连续性。然而, 在机器学习模型中更重要的突出特征是它们包含了比典型的人工设计模型更多的输入变量。许多人工智能系统依赖于模拟神经网络中的机器学习技术。利用这些技术, 机器“学习”可以捕获数据中的模式并采用对系统做出的决策有用的方式来标记这些模式, 而程序员实际上并不知道哪些模式是系统所用数据中的模式。事实上, 随着程序在不断发展,

当有新数据输入或给出新反馈“这是正确的”“这是错误的”时, 学习系统使用的模式也会随机发生变化。这意味着结果对用户或程序员来说存在着不可避免的认识盲区, 即认识不透明。

根据汉弗莱斯的观点, 这种知识涉及系统中与认识论相关的元素。虽然汉弗莱斯没有进一步分析认识论相关元素的概念, 但可以用一种易于理解的方式来呈现。例如, 一个要素可以被理解为输入到输出转换过程中的一个步骤, 也可以被理解为系统随时间整体演化过程中的一个瞬间状态转换。与认识论相关的要素是指不仅代理人知道, 而且可以被他用来说明其他要素的发生或系统的整体输出的要素。此外, 就像这类过程对某些人不透明而对其他人也不透明一样, 这类过程也可能在不同的时间对不同的人也不透明。换句话说, 这种关于认识不透明性的解释是普遍适用的。它是对认识不透明的一般和相对解释, 因为它是相对于特定代理人在特定时间的情景而言的。认识不透明是人类无法理解、还原和审查计算机模型内部运作和做出决策的全部过程, 是无法解释对计算机模型为什么这样做以及为何这样做的认识上的未知与空白。

人工智能每秒钟可能需要分析数十亿或数万亿个数据示例以及数千或数万个数据属性。算法的内部决策逻辑处理海量数据, 尤其是异构数据会增加处理难度、机器学习技术在扩展过程中面临计算资源的限制, 程序员往往用写入代码的技术来解决这个问题, 这就无形之中增加了代码的不透明性。不透明代表着信息的一种品质, 即信息的不可见性以及其对相关参与者的意义或可操作性的否定, 这种否定给算法偏见留下了肆意作为的缝隙。因为程序的质量在很大程度上取决于所提供数据的质量。因此, 如果数据已经存在偏见却不被人所知, 那么程序就会不断重现和演化这种偏见。随着人工智能工作变得越来越复杂, 决策过程也变得越来越不透明, 最终达到人类无法理解的程度。值得关注的是, 不了解人工智能如何运作和决策的这种不透明会产生道德、伦理、社会和法律问题并形成连锁反应。

二、人工智能算法的认识不透明性困境

人工智能算法的认识不透明性导致人类无法

审查、分析、解释和理解人工智能的全部计算过程。不透明性和偏见是现在被称为“数据伦理”的核心问题^[6]。用于自动决策支持和预测分析的人工智能系统引发了对缺乏正当程序、问责制、社区参与和审计的重大担忧^[7]。同时，受影响的人通常不知道系统是如何得出这一结果的，即系统对该人来说是不透明的。如果系统涉及机器学习，它甚至对专家来说也是不透明的，他们不知道如何识别特定模式，甚至不知道模式是什么，决策系统和数据集中的偏见因这种不透明性而加重。人工智能的可信度、可靠度、可亲度受到严重质疑，人工智能认识不透明治理面临着严峻的挑战。

（一）技术实现上的不透明性导致解释困难

就人工智能自身而言，因具有技术实现上的不透明性，从而出现解释困难问题。对于人工智能的计算过程，人们很难知道它为什么要做他们所做事情或者它是如何执行的，在这一问题上存在认识上的盲区和不透明性。

一方面是深度学习导致的本质不透明性问题。深度学习模型一般由多层神经网络组成，通过大量数据的训练与抓取，它们可以自动识别数据中的特征，最终实现对各种各样指令的决策功能。如在图像识别中，深度学习模型能够精准地分辨出图片中的事物，在自然语言处理中，深度学习模型能够理解并生成类人的语言。具体而言，深度学习的本质不透明性主要体现在模型内部复杂的工作原理方面。虽然人们清楚它是基于对数据的深度分析做出的决策，但关于其中的详细步骤与环节我们却无从知晓。事实上，神经网络作为深度学习的基础，是由多个层级组成，每一层都由多个神经元（或称为节点）构成。这种复杂的结构使得深度学习模型内部的参数形成了非线性的映射关系，致使人们很难理解模型。深度学习被认为是“黑箱”的本质，当信息系统呈现的信息不够详细或者信息优先顺序对于技术用户来说是无用的，这使得解释、理解模型变得更加棘手。例如，Google Photos 的图像分类器将一些黑人标记为“大猩猩”就是这样一个例子。代理生成了自己的分类规则产生了不合理的机制。然而，这个问题难以解决，以至于谷歌只能将“大猩猩”类

别全部移除^[8]。很多人工智能模型包含复杂的非线性关系。比如，激活函数的应用，以便模型能够学习复杂的语言、数据模式，以此增强人工智能的性能。但这使得模型的行为无法通过简单的数学方程或图形来还原，这必将加重认识不透明性。

另一方面是缺乏解释技术工具。解释技术工具的欠缺直接关系到人们对人工智能的信任度。在一些领域，如医疗、金融等，大家对决策结果的精确度与可解释性有着较高的要求。若是大家无法理解智能机器是怎样做出决策的，一般会有所顾虑并不会选择“言听计从”。例如，医生通常不会随意采纳无法解释的诊断建议。人工智能技术很大一部分内容主要体现为某种编程语言，由于编程语言与人类语言完全不同，因此大多数人无法直接使用它。目前仍然缺乏可解释性工具来弥合两种语言之间的鸿沟。舒尔茨（Schulz）等人以自动驾驶汽车为例说明了这一点，自动驾驶汽车将自我执行可能限制或替代人类行为的规则。考虑到算法可能需要体现在伦理困境中做何种选择，伦理困境可以从生与死的情况到不太严重的情况，这意味着法律以前留给正当性问题的决定通过算法规则变成系统化编码^[9]。如此种种，都是由于解释技术工具的缺乏。人工智能发展日新月异，特别是 DeepSeek 模型的应用，人工智能发展速度可谓一日千里，而可解释工具开发却相对滞后。

（二）数据不透明引发个人信息安全危机

相较于个人而言，数据不透明会带来个人信息安全危机。为了提高人工智能的性能，往往需要大量的个人数据，越多数据的“投喂”，人工智能性能就越高。在如此巨大的个人数据被人工智能吸纳后，信息安全问题就成为人们最为担心的一个方面，也成为人工智能可信性的一个重要考量维度。

一方面是个人的数据和隐私安全。不透明的人工智能系统可能在处理个人数据时难以被掌握与把控。这增加了个人数据被他人滥用的可能性。众所周知，数字时代，数据成为一种极其重要的资产，但数据不透明现象却引发了严重的个人信息安全危机，对个人数据隐私安全构成了威胁。数据不透明首先体现在数据收集环节。在收集用户数据时，没有明确知晓用户数据实际用途与收

集范围的情况时有发生。数据存储环节也存在不透明问题,给个人数据隐私安全同样带来了潜在风险。倘若数据存储系统遭到黑客攻击或出现数据泄露事件,用户的个人信息将面临极大的风险。当然,在数据处理环节中同样存在数据不透明的问题。数据不透明所引发的个人信息安全危机产生的影响是多维度深层次的。表面上看它造成了个人经济损失、精神困扰等困境,然而,影响的深度绝不仅限于个人层面。当人们频繁听闻个人数据被泄露的事件时,就会对人工智能应用和服务的安全性产生深深的怀疑。这种怀疑情绪一旦蔓延开来,将直接影响数字经济的发展。用户在使用人工智能的过程中,模型可能未经个人授权就搜集、存储、传播、扩散个人数据。个人对自身数据的流动或被他人使用毫不知情,这种现状为不法分子滥用他人数据信息提供了“温床”。不透明的算法决策过程可能导致对个体隐私的潜在风险。认识的不透明性造成个人未知的监控风险,人工智能的监控系统成为悬在人们头上的“上帝之眼”,人们被迫接受它的督导或规训,甚至发生个人隐私被侵犯的现象,导致未经授权的个人生活和行为被曝光。过度共享也是个人隐私被侵害的另一个“元凶”,个人数据被分享给第三方而个体并不清楚这种共享的范围。可能导致个人信息流向未经授权的平台,增加了个人隐私侵犯的可能。

另一方面是个人信息泄露。不透明的数据处理实践会引发数据泄露,使得个人敏感信息暴露给未经授权的个人、机构或者平台。个人信息作为一种宝贵的信息资源,人工智能算法的认识不透明性随时可能引发个人信息泄露,给个人带来严重的危害。不透明的数据处理实践在数据收集阶段就埋下了隐患。许多智能平台收集信息时,用户在不经意间就将自己的姓名、年龄、联系方式、浏览历史等大量个人信息拱手交出。这些信息一旦被不法分子获取,就可能成为实施诈骗、骚扰、转卖等违法行为的工具。在数据存储环节不透明的数据处理同样令人担忧,数据处理过程的不透明性也增加了个人信息泄露的风险,不透明的过程可能通过输出结果泄露信息。即使模型本身的参数是保密的,但通过模型输出的结果可能仍然

包含对输入数据的隐私敏感信息。加之,不透明的算法无法实现逆向解释,即使用户要求解释算法的决策过程,也可能无法得到逻辑严密的解释。因此,难以杜绝潜在的信息泄露风险。

(三) 不透明算法偏见加剧社会不公平

对于社会来说,不透明的算法偏见会加剧社会不公平问题。不透明的模型可能因为训练数据中的偏见而做出不公平的决策。如果模型使用具有偏见的数据进行训练,它可能在不同群体之间产生不平等的结果。

一方面是歧视性决策导致的社会不公平问题。不透明的算法有可能在训练数据中就带有偏见,最终导致智能系统在做出决策时出现歧视性的决策。人们在收集与选择数据时,各种偏见经常会随着人类自身的介入而渗入数据。如果数据带有原始性偏见,用带有偏见性的数据再去训练算法,又会产生更深的算法偏见与算法歧视等一系列问题。毫无疑问,不透明性会导致用户对系统决策过程的参与度降低,以致用户无法有效地行使个人数据的控制权。用户对系统如何使用个人数据缺乏信心,久而久之,就会降低对新技术的接受度。实际上,控制不透明算法的权力通常集中在少数人(或机构)手里。决策依靠数据,数据创造垄断,这种决策的垄断化趋势以及决策权的让渡显然不符合人民群众的利益。歧视性决策,犹如社会肌体上的毒瘤,对社会的各个层面都产生深远而恶劣的影响。在经济方面,歧视性决策严重阻碍了人力资源的合理配置。以就业市场为例,当雇主依据性别、种族或年龄等不合理因素做出招聘决策时,诸多有能力的求职者被拒之门外。女性可能由于性别因素,在传统男性主导的岗位上遭到不公平的待遇,面临更高的就业门槛。像这种歧视性的、不合理的决策难道不会导致人才的淹没吗?造成企业人才的流失吗?当然,企业会因为没有招到优秀的人才,进而影响经济的效益。同时,被歧视的人群想要获得稳定的收入也会遭到重重阻碍,这无疑将拉大贫富差距,引发社会经济结构失衡失序。在教育方面,歧视性决策破坏了教育公平的基石。学校如果根据学生的家庭背景、地域差异或社会阶层分配教育资源,就会使贫困

地区的孩子、弱势群体的子女处于劣势，优质的教育资源集中在少数地区与特定群体，导致教育机会的不平等，制约了个体的发展。从长远来看，必将削弱整个社会的创造力。试想如果社会无法给予全部孩子平等的教育机会，那么，实现社会的可持续发展也只能是一种幻想。

另一方面是造成的机会不平等现象。由于算法决策的不透明，个体不清楚智能系统做出决策的现实依据，进而出现信息不对称现象。数字化时代，数据成为重要的生产要素，具有巨大的经济价值。然而，不同的人在交互中拥有的信息数量却不尽相同，进而出现数据不公平的结果。算法决策系统借助复杂的算法与大量的数据，对大家的个人事务做出决策。但由于系统内部运作机制不透明，个体时常处于信息劣势。从公平角度看，一部分人没有办法在平等的信息基础上参与竞争，剥夺了他们本该享有的平等权利。从效率角度看，有些有能力的人群被排除在机会之外，可能导致人才浪费与社会资源的低效利用，成为社会的文明进步的桎梏。如果不透明的算法系统更偏向于服务某些特定群体，而排斥其他群体，将会导致社会中的一部分人群被数据边缘化，进一步衍生出更多更层次的社会不公平现象。

三、人工智能算法的认识不透明性治理路径

人工智能算法的认识不透明性引发人类社会深刻忧思，如何既保持机器学习的变革性潜力，同时减轻认识不透明的负面影响成为当下不得不回答的时代课题。比尔斯泰克（Biersteker）认为需要采取一种综合的方法，融入法律、组织和技术要素并需要具备相应的能力支撑。治理包含了一个定义、约束和塑造管理者期望的规则系统。从全球治理角度看，将人工智能的治理标准化为那些主体间认可的规则。这些规则定义、约束和塑造了对人工智能主体基本属性的期望^[10]。具体而言，可以通过增加算法透明度，提升隐私保护技术以及构建数智人类命运共同体，实现人工智能协同共治。

（一）增加算法透明度，着力提升可解释性

人工智能的发展重点除了强化性能以外也应重视可解释性，人工智能应用的解释程度应该与决策的重要性相匹配，算法设计应该是可理解的。

提升人工智能的可解释性是治本之策，以使用户能够理解系统的运作过程、运作方式和最终决策。一旦技能短缺问题得到解决后，透明度问题将开始变得更加容易治理。

第一，设计可解释性模型。引导研究人员朝着设计可解释性的人工智能模型方向发展。在设计阶段将解释性需求内嵌其中，可多使用简单的模型，减少复杂模型的使用，以便更容易理解和解释“黑盒”的“内部工作”。另外，可以考虑设计透明的模型结构。在模型结构上下功夫，选择使用更为透明的模型结构，决策树、线性模型或规则引擎。这些模型更加通俗易懂，便于人类理解，直观的决策规则从源头上减少不透明性。

第二，引入可解释性模块。可解释性模块能够为人工智能系统的决策提供清晰的解释。以在医疗领域为例，一个基于深度学习的疾病诊断智能系统，如果引入可解释性模块，医生就能了解到系统是基于哪些症状、体征以及医学知识得出的诊断结果。这不仅能增强医生对诊断结果的信任，还能帮助医生更好地理解疾病的诊断依据。从技术角度来看，可解释性模块的引入可以促进人工智能技术的持续性发展。研究人员可以通过分析可解释性模块提供的信息，搞清楚模型内部的运行机制，实现对模型更好地优化。如果发现模型对特定特征过度关注出现了决策偏差，就可以有针对性地加以调整，将庞大的模块细分为若干小的模块，以便增强人工智能系统的可解释性。通过采用局部解释性方法，开发可视化工具和仪表盘，帮助用户理解模型的决策过程。这些工具可以以图形化的方式展示模型的内部结构和重要环节，还应在解释工具的创造上发力，为人工智能认识不透明性治理夯实技术根基。

第三，确保透明的数据收集。确保透明的数据收集至关重要，这不但关乎个人隐私权，而且还关系到数据的质量与社会的公正性。透明的数据收集要做到明确告知数据主体收集的目的。只有当用户清楚了解数据的用途，他们才能做出是否提供数据的理性决定。同时，数据收集的范围也必须透明。提供者应详细列出所收集的数据类型，包括个人信息和使用过程中产生的数据。在数据收集方式上，也应保持透明。这样用户可以

更好地了解自己的数据是如何被获取的，并且可以根据自己的需求对某些数据收集方式进行设置。数据收集者还应建立有效的反馈机制。当大家对数据收集过程有疑问或担忧时，可以及时得到相应的回应，进而增强对数据收集过程的信任度。在运用过程中做到数据收集是透明、公开、公平的，摒弃隐形的偏见。重塑用户选择的主体性，明确告知用户其个人数据将用于何种用途，确保透明的数据收集、存储与使用。在算法中嵌入伦理与人本逻辑，坚持“人是算法的尺度”，从设计开发到应用全过程着力提升人工智能的可解释性。

（二）推进算法公开，提升隐私保护技术

始终坚持以人为本的技术观确保个人信息安全，推进算法公开要积极向数据主体披露数据控制方持有的个人数据情况。人工智能问题及其对透明度提出的新挑战往往被描述为对公共价值观的挑战^[11]。对公共价值观的践行，如欧盟在《通用数据保护条例》中引入了相关的人工智能透明度要求^[12]。推进算法公开透明，提升隐私保护技术以规范人工智能系统在处理个人信息时的行为。根据拉图尔（Latour）的说法，黑盒算法本质上是包含通过算法的编程过程创建人际关系的动态融合。如果这个概念化是一个起点，那么彻底打开这些黑匣子的探索是一种误解^[13]。只有在算法训练和应用阶段注重隐私保护，减少个人信息安全问题的发生才是治标之策。

第一，采用差分隐私技术。差分隐私技术能够做到在描述数据集中的群体特征的同时又保护个人的数据信息，核心特点在于通过添加适当的噪声来模糊个体数据的特征，使得在分析数据时无法准确识别出单个个体的信息。这种对个体隐私的严格保护，人们在提供数据时就会更加主动与安心，从而有利于数据的收集和存储。从数据利用的角度来看，差分隐私技术并没有因保护隐私而完全牺牲数据的价值。它允许研究人员和数据分析人员从经过处理的数据中获取有价值的统计信息，如平均值、比例、趋势等。

第二，引入加密技术。使用加密技术对个人数据进行加密，通过特定的算法将原始数据转换为密文形式，以此增强数据的安全性。这就好比

给个人数据加上了一把坚固的锁，只有持有钥匙的人才能打开。只有拥有正确解密密钥的授权用户才能将密文还原为明文。在数据存储环节，加密技术同样发挥着重要作用。无论是本地存储设备还是云端服务器，对个人数据进行加密可以防止数据因存储介质被盗取或服务器被攻击而泄露。即使存储设备落入不法分子手中，他们所得的也只是一堆无法解读的密文，无法达成非法获利的目的，从而有效保护了个人数据的安全。从访问控制的角度来看，加密技术确保只有经过授权的用户才能访问原始个人数据。这样可以避免内部人员的违规操作导致数据泄露，同时也能防止外部黑客的入侵。质言之，使用同态加密算法，引入加密技术能够实现个人数据加密，提升个人数据隐私保护效果。

第三，采用分布式学习。分布式学习通过让每个设备处理部分数据并计算模型更新，充分利用了各个设备的计算资源。在一个拥有众多移动设备的网络中，每个移动设备都可以根据自身所拥有的数据进行局部模型的计算。这避免了将大量数据集中到一个中心位置进行处理，不但减少了数据传输的压力，而且在一定程度上保护了数据隐私。由于每个设备只处理部分数据，数据的分散性为个体数据集中获取增加了难度。以全球科研合作项目为例，不同国家的研究团队可以利用各自的设备进行分布式学习，然后共享更新信息，共同推动项目所涉及的模型不断向前发展。通过这种方式可以大大减少个人数据泄露的风险，因为每个设备持有和转载的个人信息都是分解后的，能够起到保护个人数据的作用。总之，坚持以人为本，多层次多渠道多方位刻画数据全样本，以人文驱动技术，对信息安全进行管控是解决人工智能认识不透明性带来隐患的有利举措，能够避免更多的个人数据、个人隐私及个人信息流向平台和社会。

（三）构建数智人类命运共同体，实现人工智能协同共治

数据具有极强的流动性，能够轻易地跃出国界，因此国际合作也至关重要。要着力构建数智人类命运共同体，实现人工智能协同共治，促进人工智能系统在决策中做到公正、安全、可靠。

第一，弘扬全人类共同价值，凝聚国际共识。

从价值维度看,人工智能本身作为人的智力延伸而生产出来的技术人工物,存在价值体现为人类服务和满足社会生活的需要,应保持先进本色,不应成为侵害他人隐私、盗取他人数据的工具。作为技术服务于人类的人工智能本身不具备价值观,但研发和应用的人却有自己的价值观,这难免会将自身的价值观内嵌于算法之中,导致作为人工智能的中枢神经的算法带有明显的价值倾向。加强对人工智能系统的研发和应用的监管,必须弘扬全人类共同价值,凝聚国际共识,引领人工智能研发、应用朝着有利于人类解放的方向发展。以价值观“软”措施引领人工智能健康发展,有人提议在数据表中对数据集进行标准化操作,这将使识别这种偏见更加可行^[14]。

第二,制定和完善制度建设。建立定期审查制度,对人工智能研发、应用进行定期审查,确保在任何阶段都符合监管和道德要求。系统的更新也需要经过审查,以确保符合安全和保密要求,进行伦理和道德审查,确保模型的使用符合道德和社会共同价值观,确保模型的输出与社会共同价值观对齐,促使模型做出符合人类预期的决策,避免模型做出对人类有害或不符合道德标准的行为。同时也要定期审查潜在的社会影响、公平性和隐私问题。例如,一辆自动驾驶汽车可以根据决策标准来优化安全方面的相关性能指标,而另一辆则优化与速度相关的性能指标。无论是谁设计的都要执行安全检测。

第三,加强法治监管。人工智能监管仅凭价值观引领和定期审查是远远不够的,还需要有相关法律规范行业行为。安娜尼(Ananny)等人认为,如果我们认识到单靠透明度无法创建负责任的制度并深挖这种窘境背后的原因,我们也许能够使用透明度的极限作为概念来理解算法如何可能会被追究责任^[15]。建立健全国际《人工智能法案》,创设数据隐私等方面的国际标准,加大对不法份子滥用他人数据、侵犯他人隐私、嵌入歧视算法等行为的监管“硬”措施力度。通过法律手段打击因人工智能认识不透明性产生的违法犯罪行为。只有通过全球合作,实现人工智能协同共治,才能更有效地规范人工智能技术的使用。

当前,人工智能的认识透明性问题是人工智能治理亟需应对的挑战之一。通过提升人工智能的可解释性到采用隐私保护技术的标本兼治,从价值观引领到法律管制内外联动,“软”“硬”兼施来根治认识不透明带来的歧视性决策、隐私安全和社会不公平等问题。我们有望建立透明、可信、负责的人工智能系统,推动人工智能在科技发展和社会进步中发挥更为积极的作用。值得一提的还有,对于人工智能技术的依赖,需要我们重新审视人的主体性。人工智能技术作为现代社会科技的“领头雁”,不仅是一种强大的物质力量,也是一种巨大的精神力量。对于这一点,我们必须时刻保持清醒的头脑。在人工智能技术发展过程中厚植人文关怀精神,尊重人的价值,确保人的主体性不被机械物所吞噬,推动人工智能走向更安全、更高效和可持续发展的未来。

参考文献:

- [1] 贾玮晗,董春雨.事后解释可以消除认识的不透明性吗[J].探索与争鸣,2024,(6).
- [2] 梁丽,苏湛.计算机模拟的认识不透明性[J].自然辩证法通讯,2021,43(11).
- [3] The Free Dictionary.Opacity[EB/OL].<https://www.thefreedictionary.com/opacity>, 2025-02-25.
- [4] 董春雨.从机器认识的不透明性看人工智能的本质及其限度[J].中国社会科学,2023,(5).
- [5] Humphreys, P. The Philosophical Novelty of Computer Simulation Methods[J]. Synthese, 2009, 169(3): 615-626.
- [6] Floridi L, Taddeo M. What is Data Ethics[J]. Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences, 2016, 374(2083).
- [7] Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V.,... Schultz, J. (2018). AI Now Report 2018[EB/OL]. New York: AI Now Institute, New York University.https://ainowinstitute.org/AI_Now_2018_Report.pdf.

(下转第86页)