科学通报 2023 年 第 68 卷 第 13 期: 1604 ~ 1610
科技伦理前沿谈专题 观 点
《中国科学》杂志社 SCIENCE CHINA PRESS

医学人工智能的算法黑箱问题: 伦理挑战与化解进路

杨军洁[1], 周程[1,2*]

1. 北京大学哲学系, 北京 100871;
2. 北京大学医学人文学院, 北京 100191
* 联系人, E-mail: zhoucheng@pku.edu.cn

人工智能技术正逐渐成为推动生物医学发展的重要力量. 与此同时, 医学人工智能带来的伦理与治理问题也日益凸显. 笔者以“medical artificial intelligence ethics”为主题在Web of Science数据库中进行检索, 共获得相关文献554篇(1975~2022年). 基于这些文献, 用VOSviewer1.6.18软件绘制关键词共现图谱(图1), 可见隐私(privacy)、自主(autonomy)和责任(responsibility)等问题已成为医学人工智能伦理研究中的重点议题; 而算法的透明性(transparency)问题则为连结“伦理原则”(黄色)、“算法模型”(红色)和“医疗实践”(蓝色)等主题聚类的重要节点, 表明透明性问题已成为医学人工智能领域中的关键伦理挑战.

本文拟聚焦医学人工智能的透明性问题, 从考察医学人工智能算法黑箱的含义与特点出发, 剖析其中蕴含的伦理难题, 结合当下全球伦理研究中的道德多元主义和语境主义思潮, 参考近年医学人工智能研究与应用中的具体案例, 探讨应对算法黑箱问题的道德设计进路与方法, 并为混合式道德设计进路进行辩护.

1 医学人工智能的算法黑箱与伦理挑战

随着人工智能的飞速发展, 人类正逐渐进入“算法社会”(algorithmic society)[1]. 在算法社会中, 人工智能依托其强大的机器学习能力而具有一定的自主性, 但这也导致了算法缺乏透明性的问题. 具体而言, 模型经过大量训练后, 其内部状态变得相当复杂, 输入与输出之间的运算是自动进行的, 这也导致人们既难以准确预测算法的行为, 也不易理解算法输出特定结果的机制. 由此, 人们就把这一现象称为“算法黑箱”(algorithmic black-box)问题[2].

目前, 在机器学习领域中, 已有部分研究者通过研发若干可解释性工具, 来提升算法的透明度和可解释性, 以此试图解决算法黑箱问题. 例如, 模型无关的局部解释(local interpretable model-agnostic explanations, LIME)技术可以帮助人类理解图像识别模型中的分类依据; 夏普利值(Shapley value)可以用来描述各个特征值对模型预测结果的贡献度, 并以此提高算法的可解释性等. 然而, 在医学人工智能领域中, 具有可解释性的模型却并不普及, 其中一个原因在于, 这些模型

杨军洁 北京大学哲学系博士研究生. 研究领域为生物医学伦理、生命科学史.

周程 北京大学哲学系教授、博士生导师, 北京大学医学人文学院院长, 中国科协-北京大学科学文化研究院副院长, 国务院学位委员会科学技术史学科评议组成员. 研究领域为科学社会史、科学技术与社会、创新管理与科技政策.

无法在临床实践中提供真正令人满意的解释[3], 而算法专家、医疗专家、患者和公众对可理解性和透明性的不同要求, 使得算法黑箱在医学人工智能的研发与应用过程中仍然是一个严峻的伦理挑战. 从根本上讲, 这些伦理挑战乃上述不同道德主体在面临各种道德风险时可能产生的价值冲突, 表1即为采用伦理矩阵[4]对医学人工智能算法黑箱引发的伦理问题进行的梳理.

在数据维度上, 机器学习的表现高度依赖于数据集的数量与质量, 但是算法黑箱使不同主体对数据的掌握、理解和利用水平存在差异, 这就导致了一定的伦理风险, 如医学人工智能企业未将患者数据脱敏而产生数据滥用问题, 以及患者隐私权被侵犯的问题. 据统计, 截至2021年, 美国食品药品监督管理局已经批准了118种用于临床实践的机器学习算法, 但是其中84%的算法缺乏足够的用于评估模型效果的验证

© 2023《中国科学》杂志社

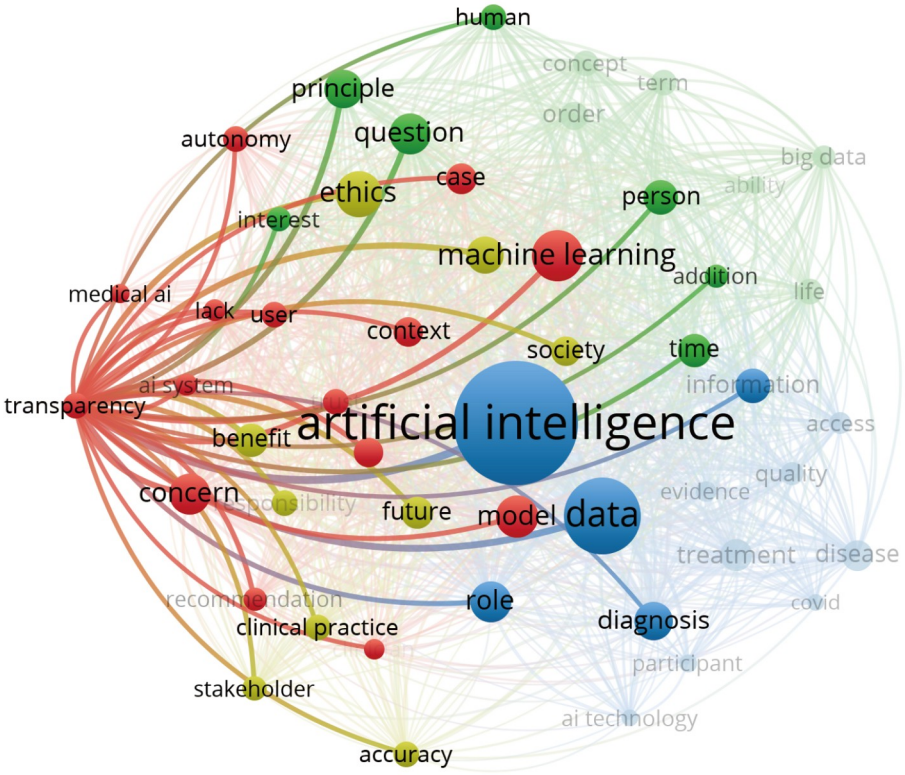

图 1 “医学人工智能伦理”主题文献关键词知识图谱. Web of Science数据库/VOSviewer, 检索日期: 2023-03-10
Figure 1 Knowledge graph of keywords of “medical artificial intelligence ethics” topic literature. Web of Science database/VOSviewer, retrieval date: 2023-03-10

表 1 医学AI算法黑箱的伦理矩阵分析
Table 1 Ethical matrix analysis of algorithmic black-box of medical AI

道德主体	数据伦理	算法伦理	社会伦理
内部层次(算法专家)	数据滥用问题	知识产权问题	歧视偏见问题
内外交互层次(医疗专家)	数据误用问题	自主性问题	责任归属问题
外部层次(患者、公众)	侵犯隐私问题	安全性问题	信息茧房问题

集, 或者未披露任何关于患者评估的验证集信息[5], 这可能会导致以“垃圾输入, 垃圾输出”(garbage in, garbage out)为特点的数据误用问题, 影响医疗决策的准确性与安全性.

在算法维度上, 专家开发的程序应该受到知识产权的保护, 甚至有些算法出于安全的考虑被定为机密, 这又与透明性产生了价值冲突. 当患者无法获得或理解数据与模型的原始信息时, 难免也会对产品的安全性质疑. 算法黑箱还会威胁到医生的诊疗自主权, 即本应该作为一种辅助手段的医学人工智能可能会导致医生对算法提供的诊疗决策过度依赖. 例如, 有许多医生报告IBM沃森诊疗系统常常会给出一些令人费解的用药建议, 甚至一些治疗方案对于特定患者来说是相当危险的, 但医生无法向算法黑箱“询问”它为何会做出这类看起来相当不合理的决策[6], 而一些经验不足的医生就有可能盲目接受医学人工智能给出的建议而造成误诊.

在社会维度上, 算法企业利用黑箱使得公众在不知不觉中接受了算法控制, 又由于算法专家和公众之间存在技术素养的差异而形成了信息茧房, 致使可理解的医学人工智能更难以实现. 由于算法黑箱具有不透明性, 人们难以察觉并修正模型中可能包含的偏见与歧视, 这进而会对患者的健康带来威胁[7]. 此外, 医生与医学人工智能共同参与医疗决策过程, 会使相关的责任归属问题变得尤为复杂: 谁该为诊疗中出现的差错负责? 是算法开发者还是听取算法建议的医生? 人工智能自身又能否作为责任主体? 一项针对美国民众的调查显示, 当医学人工智能造成医疗事故时, 公众(66.0%)更倾向于认为需要承担主要责任的是医生一方, 而医生(43.8%)则更倾向于认为是人工智能产品的提供方应该承担主要责

任[8]. 对于医学人工智能自身能否承担道德责任的问题, 当前学界的主流意见是, 由于医学人工智能没有自由意志, 因此无法作为积极的道德主体承担责任[9]. 但是, 随着人工智能技术的发展, 我们仍有必要思考如何构建符合人类道德要求的医学人工智能, 以应对未来具有高自主性的医学人工智能可能带来的社会风险与伦理挑战.

2 应对算法黑箱的道德设计进路

由前文可见, 算法黑箱问题成为当前医学人工智能领域中的关键伦理挑战, 而发展具有自主道德推理能力的人工智能, 即构建人工智能道德体(artificial moral agents, AMAs), 被认为是化解相关伦理问题的一种可行方案. 这一方案试图通过将伦理上“制度性封闭”[10]的范式纳入研发过程中, 以规避算法黑箱可能带来的伦理风险. 目前, 有学者区分了三条实现AMAs的进路——自上而下进路、自下而上进路与混合式进路[11]. 笔者认为: 自上而下和自下而上进路都存在一定局限性; 要想化解医学人工智能算法黑箱导致的伦理挑战, 应该采取“自上而下-自下而上”相结合的混合式道德设计进路.

2.1 自上而下与自下而上进路的局限性

“自上而下进路”指的是基于特定的伦理原则设计医学人工智能道德体, 以实现透明且可解释的医学人工智能. 许多学者和机构都曾对医学人工智能提出过各种关于道德设计的规范与倡议. 例如, 在2021年世界卫生组织发布的《卫生领域人工智能的伦理与治理》指南中, 将“确保透明度、可解释性和可理解性”作为医学人工智能所应遵循的基本伦理原则之一, 并对技术限制、操作记录、数据性质和算法模型等相关信息的透明性提出了要求[12]. 然而, 这种自上而下进路所能取得的效果是相当有限的[13], 问题在于这些预设的伦理原则有时难以对复杂的伦理情境作出适当的反应.

首先, 医学人工智能作为一个新兴的学科领域, 在专家内部本就缺乏一定程度的伦理共识——仅在数据科学家中就存在道德原子主义与整体主义两大阵营的分歧[14]. 前者认为事实与价值相互独立, 且采取技术中立论的立场, 而后者则认为事实与价值二者不可分割, 并认为应该对人工智能技术的伦理风险进行防控. 而在伦理学家中, 对应该采取何种伦理原则的问题也难以形成一致的意见. 在不同原则之间存在冲突的情况下, 采用某套既定的伦理框架指导医学人工智能设计的合理性会受到质疑. 例如, 透明性原则要求医学人工智能可以接受审查, 但是保护隐私的原则又要求其对数据信息进行保密, 自上而下进路引发的这类道德两难问题, 将会给设计者带来更多挑战.

其次, 医学人工智能的伦理原则与实用目标之间可能存在矛盾, 透明性的实现常常需要以牺牲一定程度的准确性为代价. 例如, 在手术中麻醉医师需要监测许多生理指标来对麻醉深度进行调节, 这些生理指标之间往往存在线性关系, 因此有学者开发了一套机器学习算法, 通过梯度下降法搭建回归模型以实现麻醉的自动调控, 这是一种具有高度可理解性的算法; 但是在临床实践中, 这种算法并不能提供最佳的剂量建议, 而基于透明性较低的神经网络所输出的决策, 则能更好地实现对麻醉深度的控制, 可是这又会带来一定的安全风险[15]. 是要准确性和安全性还是要可理解性和透明性? 自上而下进路并不能提供直接的回答.

最后, 当我们在具体实践中面临道德困境时, 伦理原则常常过于抽象, 且缺乏一定的灵活性, 无法为我们提供具体的指导意见. 如果设计者基于透明性原则为医学人工智能设定逻辑单调的道德推理算法, 一个难以解决的问题就是这种算法能否与不同社群和文化中多元的道德直觉相契合[16]. 例如, 制造商、医生和普通用户在评估用于健康管理的可穿戴设备算法时, 因所处情境及认知目标的不同, 对算法可理解性程度的评价也不同[17]. 而自上而下进路难以对不同主体间差异化的可理解性标准进行充分的考量, 因此这一进路仍无法化解算法黑箱的伦理风险.

另外, 还有一些学者支持“自下而上进路”. 这一进路无需工程师为人工智能设置一套既定的伦理原则, 而是在一系列基于具体案例的强化学习场景中, 使其自主进化出一套符合人类道德直觉的运作方式. 但是, 仅仅依赖这一进路同样无法解决算法黑箱带来的伦理问题. 首先, 自下而上进路作为一种事后的规制和调整手段, 其试错过程中产生的消极道德后果无法规避. 其次, 不同人的道德直觉存在一定差异, 而机器对人类道德行为的模仿未必能收敛到一个统一的道德推理框架中. 例如, 诊疗机器人在训练中可能无法识别出精神障碍患者话语中对其病史的隐瞒和欺骗, 而自下而上进路由于缺乏伦理原则层面的指导, 可能会让医学人工智能习得一些违反道德直觉的行为模式[18]. 此外, 从人工智能系统内部发展出的道德“功能”需要长期的实验探索, 在目前相关技术还非常不成熟的情况下, 人类主体无法预知人工智能未来的行为是否还会符合人类社会的伦理要求, 这反而增加了算法的不透明性. 因此, 自下而上进路无法成功构建符合我们要求的医学人工智能道德体, 算法黑箱产生的伦理难题仍然无法得到化解.

2.2 混合式道德设计进路之展望

鉴于当前人工智能的发展现状, 有必要将更符合伦理的道德设计嵌入医学人工智能研发的全流程中[19]. 笔者认为, 自上而下与自下而上相结合的“混合式进路”才能更好地应对医学人工智能算法黑箱所带来的伦理挑战. 混合式道德设计进路要求, 既通过自上而下的路径, 让工程师为医学人工智能设定具有一定弹性的伦理框架, 通过对语境进行充分考量以有效提升算法的透明性程度; 又采取自下而上的路径, 让算法学习人类的道德行为模式, 发挥医学人工智能在处理多情境信息中的优势, 发展出多元的道德推理模型.

2005年科林·艾伦(Colin Allen)等人[20]提出混合式进路时，学界并不清楚应该怎样通过这一进路使人工智能实现联结主义美德，即如何依托于神经网络使人工智能的道德推理与现实情境相契合. 早期混合式进路对人工智能道德地位的理解仍然采取的是计算主义的立场，即认为人工智能只有机械性地进行道德计算(moral computation)的能力[21]，而不能形成真正的道德推理，因此无法通过道德图灵测试，进而根本不具备成为道德主体的可能[22]. 随着人工智能伦理研究的不断推进，混合式的道德设计进路取得了理论上的澄清与突破，例如，目前有学者阐述了将人类价值观嵌入人工智能系统中的具体流程[23]. 而在实践层面，混合式进路也从早期实验室环境下非常有限的机器训练，逐渐走向了让人工智能处理真实世界中道德困境的尝试[24].

然而，仍有部分学者认为，混合式进路及其所希望达成的“人工智能道德体”根本是冗余的设定，我们需要的仅仅是更加安全准确的人工智能，而非更加道德的人工智能[25]. 但笔者认为，鉴于当今各种价值体系和文化形态在全球范围内流动、冲突与交融，人工智能所面临的伦理挑战日益复杂，需要我们从混合式进路对人工智能的伦理问题进行反思；而在生命医学领域中，人们对人工智能的特殊要求，也为混合式进路的合理性提供了辩护. 首先，与传统的道德一元论相反，道德多元主义认为医学中的道德决策是复杂且多元的，不同主体在不同情境下对人工智能的道德行为有着不同的期望，而混合式进路能更好地回应现实世界中的道德分歧；其次，道德语境主义反对规范伦理学框架下关于道德的绝对主义立场，强调可理解性和透明性的判定标准与人们所处的语境是强相关的，并不存在单一的判定标准，而混合式进路为化解算法黑箱伦理难题提供了价值对齐(value alignment)的路径，能够使医学人工智能的道德行为与利益相关各方的价值判断相契合；最后，这一进路遵循了道德哲学中的反思平衡(reflective equilibrium)方法，在确立医学人工智能的实用性目标基础上，识别出其中蕴涵的伦理挑战，在拟定的道德原则基础上不断依据现实情况进行调整，使医学人工智能得以更好地应对复杂的道德问题，也能更好地满足人们对安全性和准确性的需求.

从内部层面看，混合式道德设计需要算法工程师在设计之初就要考虑算法的透明性问题. 一方面，通过开发与底层机器学习模型相分离的可解释性工具，规避算法黑箱可能带来的伦理难题. 例如，算法工程师在建立神经成像的机器学习模型时，加入一个支持预测、验证和解释的互补性程序，评估噪音和干扰对模型的影响，提示算法黑箱中可能蕴涵的歧视与偏见[26]. 对于分析医学图像的深度神经网络来说，提升算法可解释性的方法包括概念学习模型、反事实解释、内部网络表示等[27]. 另一方面，混合式进路要求工程师构建具有自我解释能力的算法，从模型内部打开黑箱. 有学者提出，可以通过卷积神经网络等方法实现患者数据的可视化，让医学人工智能具有更直观的人机交互界面，提升算法的可理解性；也可以通过增强医学人工智能决策流程的可追踪性与决策规则的稳健性[28]，提升算法的透明性.

从外部层面和内外交互层面看，混合式进路需要利益相关各方积极参与算法设计过程，以化解算法黑箱难题. 目前，已经有一些医学人工智能开发者将医生和患者纳入算法设计环节之中，通过融入多元视角来提升医学人工智能的可理解性. 有团队开发了一个评估双相情感障碍患者治疗方案的算法，将患者纳入算法设计的研究咨询小组中，向其征求透明性和可理解性方面的意见，以此来保证患者的自主权[29]. 此外，医疗专家除了可以帮助算法工程师处理标签、监督学习外，还可以在模型验证环节发挥更积极的作用[30]. 有研究者利用一种名为“Delphi”的道德推理神经网络，来处理医学专家的意见与共识，这种算法能够根据医生的临床实践及时调节其可理解性[31]. 在医学人工智能的算法设计阶段实行开放的、参与式的研发模式，其意义不仅仅在于通过各主体间的合作提升算法透明性，还在于构建“人本主义”的医学人工智能道德体，使医学人工智能在拥有道德体地位的同时，能够保护人类的尊严与主体性，同时让开发者承担相应的道德责任[32].

综上，采取混合式的道德设计进路，能让医学人工智能在现实情境中学习道德推理，发展出内嵌于自身算法中的道德模式，这使得医学人工智能可以有效应对不同伦理原则间的冲突，适应实际应用中的道德需求，规避自上而下进路中原则的僵化而带来的困难. 另外，混合式进路倡导多元主体的广泛参与，协调不同主体间的道德选择与利益冲突，这能帮助医学人工智能更好地处理复杂的道德决策场景，使其行为与不同文化中的道德直觉相契合，以解决自下而上进路中的不确定性问题.

需要指出的是，就人工智能的发展现状来看，混合式进路目前只能达到对人工智能内部相当有限的可理解性，在短期内算法黑箱不可能被完全“打破”. 另外，人工智能系统究竟能否具有意识、知觉与自由意志，对于这些形而上学问题，人们仍存在诸多争议，这使它们无法获得完全的道德主体地位. 基于此，具有局部可解释性的“算法灰箱”设计，由于能够在准确的“黑箱”与透明的“白箱”之间做出较好的平衡，符合医学领域对人工智能道德体的诉求，因此可以作为当前医学人工智能道德体设计的可行目标. 例如，有研究从卷积神经网络中提取脑癌图像的分类信息，再从病史中提取脑癌的位置、大小等特征信息，将二者结合以提升脑癌诊断模型的可解释性[33]. 在发展医学人工智能道德体的过程中，同样可以将依据道德推理法则的白箱，与基于大量道德直觉和道德判断数据而训练出的黑箱相结合，通过这种混合式进路应对医学中更复杂的伦理挑战.

3 结语

人工智能技术正在引领着医疗健康领域的变革，而算法

黑箱则为医学人工智能的发展带来了不小的伦理挑战. 化解算法黑箱的伦理难题, 构建医学人工智能道德体, 不仅需要算法工程师、公司、政府、医生、患者、伦理学家等多方的参与, 为医学人工智能的设计提供具有一定共识的伦理原则, 同时也需要为医学人工智能设定功能性的道德体地位, 使其行为和价值框架与人类的可理解性和自主性相契合. 只有通过这种混合式的道德设计进路, 才能更好地回应当前社会对道德多元主义和语境主义的诉求, 让人工智能依据多元价值体系和具体道德情境进行道德推理与行动, 才能更有效地助推医学人工智能的创新与发展.

致谢 感谢教育部人文社会科学重点研究基地重大项目“当代认知哲学基础理论问题研究”(22JJD720007)资助.

推荐阅读文献

1 Schuilenburg M, Peeters R. The Algorithmic Society. New York: Routledge, 2021
2 Durán J M, Jongsma K R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics, 2021, doi: 10.1136/medethics-2020-106820
3 Antoniadi A M, Du Y, Guendouz Y, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Appl Sci, 2021, 11: 5088
4 Forsberg E M. The ethical matrix—A tool for ethical assessments of biotechnology. Glob Bioethics, 2004, 17: 167–172
5 Ebrahimian S, Kalra M K, Agarwal S, et al. FDA-regulated AI algorithms: Trends, strengths, and gaps of validation studies. Acad Radiol, 2022, 29: 559–566
6 Smith H. Clinical AI: Opacity, accountability, responsibility and liability. AI Soc, 2021, 36: 535–545
7 DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inf Assoc, 2020, 27: 2020–2023
8 Khullar D, Casalino L P, Qian Y, et al. Public vs physician views of liability for artificial intelligence in health care. J Am Med Inf Assoc, 2021, 28: 1574–1577
9 Chen A T, Zhang X Q. Discussion on the medical ethical responsibility problems caused by artificial intelligence assisted diagnosis and treatment (in Chinese). Chin Med Ethics, 2020, 33: 803–808 [陈安天, 张新庆. 医学人工智能辅助诊疗引发的伦理责任问题探讨. 中国医学伦理学, 2020, 33: 803–808]
10 Chen X P. Artificial intelligence: Applicability, risk analysis, and innovation mode upgrading (in Chinese). Sci Soc, 2021, 11: 1–14 [陈小平. 人工智能: 技术条件、风险分析和创新模式升级. 科学与社会, 2021, 11: 1–14]
11 Wallach W, Allen C. Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press, 2008
12 World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization, 2021
13 Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell, 2019, 1: 501–507
14 Greene T, Dhurandhar A, Shmueli G. Atomist or holist? A diagnosis and vision for more productive interdisciplinary AI ethics dialogue. Patterns, 2023, 4: 100652
15 Connor C W. Artificial intelligence and machine learning in anesthesiology. Anesthesiology, 2019, 131: 1346–1359
16 Constantinescu M, Voinea C, Uszkai R, et al. Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. Ethics Inf Technol, 2021, 23: 803–814
17 Li R C, Muthu N, Hernandez-Boussard T, et al. Explainability in medical AI. In: Cohen T A, Patel V L, Shortliffe E H, eds. Intelligent Systems in Medicine and Health. Cham: Springer, 2022. 252–253
18 Rzepka R, Araki K. ELIZA fifty years later: An automatic therapist using bottom-up and top-down approaches. In: Van Rysewyk S, Pontier M, eds. Machine Medical Ethics. Cham: Springer, 2015. 257–272
19 Teng Y, Wang G Y, Wang Y C. Ethics and governance of general models: Challenges and countermeasures (in Chinese). Bull Chin Acad Sci, 2022, 37: 1290–1299 [滕妍, 王国豫, 王迎春. 通用模型的伦理与治理: 挑战及对策. 中国科学院院刊, 2022, 37: 1290–1299]
20 Allen C, Smit I, Wallach W. Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics Inf Technol, 2005, 7: 149–155
21 Crook N, Nugent S, Rolf M, et al. Computing morality: Synthetic ethical decision making and behaviour. Cogn Comp Syst, 2021, 3: 79–82
22 Véliz C. Moral zombies: Why algorithms are not moral agents. AI Soc, 2021, 36: 487–497
23 van de Poel I. Embedding values in artificial intelligence (AI) systems. Minds Mach, 2020, 30: 385–409
24 Winfield A F, Michael K, Pitt J, et al. Machine ethics: The design and governance of ethical AI and autonomous systems. Proc IEEE, 2019, 107: 509–517
25 van Wynsberghe A, Robbins S. Critiquing the reasons for making artificial moral agents. Sci Eng Ethics, 2019, 25: 719–735
26 Kohoutová L, Heo J, Cha S, et al. Toward a unified framework for interpreting machine-learning models in neuroimaging. Nat Protoc, 2020, 15:

1399–1435
27 Salahuddin Z, Woodruff H C, Chatterjee A, et al. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med, 2022, 140: 105111
28 Yao J, Liu Y, Li B, et al. Visualization of Deep Models on Nursing Notes and Physiological Data for Predicting Health Outcomes Through Temporal Sliding Windows. In: Shaban-Nejad A, Michalowski M, Buckeridge D L, eds. Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability. Cham: Springer, 2020. 115–129
29 Banerjee S, Alsop P, Jones L, et al. Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. Patterns, 2022, 3: 100506
30 Vellido A. Societal issues concerning the application of artificial intelligence in medicine. Kidney Dis, 2019, 5: 11–17
31 Lukas J M, Hein A, Diepold K, et al. Clinical ethics: To compute, or not to compute? Am J Bioethics, 2022, 22: W1
32 Coeckelbergh M. Narrative responsibility and artificial intelligence. AI Soc, 2021, doi: 10.1007/s00146-021-01375-x
33 Loyola-Gonzalez O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access, 2019, 7: 154096–154113

Summary for “医学人工智能的算法黑箱问题: 伦理挑战与化解进路”

Algorithmic black-box problem in medical artificial intelligence: Ethical challenges and solution approach

Junjie Yang[1] & Cheng Zhou[1,2*]

[1] *Department of Philosophy, Peking University, Beijing 100871, China;*
[2] *School of Health Humanities, Peking University, Beijing 100191, China*
* Corresponding author, E-mail: zhoucheng@pku.edu.cn

Artificial intelligence (AI) is gradually becoming an important force driving the development of biomedicine. However, as AI increasingly relies on complex and opaque machine learning algorithms, a critical ethical issue known as the “algorithmic black-box” problem has emerged. Despite the development of several explainability tools, they have not been widely adopted in the medical field due to their inability to provide satisfactory explanations in clinical practice. Furthermore, different stakeholders including algorithm experts, medical professionals, patients, and the general public have varying requirements for explainability and transparency. As a result, this has created a series of internal, internal-external interaction, and external-level ethical issues in data, algorithmic, and social dimensions.

In the ethical challenges associated with the increasing complexity and opacity of medical artificial intelligence, constructing medical artificial moral agents has been proposed as a viable solution. To implement ethical frameworks in this domain, three approaches have been identified: The top-down approach, the bottom-up approach, and the hybrid approach. The top-down approach prioritizes moral design based on specific ethical principles. However, this approach faces difficulties in responding appropriately to complex ethical situations due to the lack of consensus among ethical experts, contradictions between ethical principles and practical goals, and the abstract nature of moral principles. The bottom-up approach, on the other hand, requires medical artificial intelligence to develop a set of operating methods that align with human moral intuition in a series of case-based reinforcement learning scenarios. Nonetheless, this approach is only effective in retrospective regulation, and converging moral reasoning to a certain pattern remains challenging.

In light of the current state of artificial intelligence development, it is imperative to adopt a “hybrid approach” that integrates both top-down and bottom-up approaches throughout the process of developing medical artificial intelligence. This involves establishing a flexible ethical framework via the top-down approach that takes into account contextual factors to enhance algorithm transparency, and leveraging the strengths of medical artificial intelligence to develop diverse models of moral reasoning through the bottom-up approach that incorporates multiple contextual information. While some scholars may argue that the hybrid approach is redundant, given the contemporary demand for moral pluralism and contextualism, this path toward reflective equilibrium can better address moral disagreements in the real world, ensuring that the ethical behavior of medical artificial intelligence aligns with the value judgments of relevant stakeholders.

From an internal perspective, the hybrid approach involves algorithm engineers developing tools for explainability that are independent of the underlying machine learning models, assessing ethical risks, or constructing algorithmic models with self-explanatory capabilities to enhance the explainability of the algorithm. From external and internal-external interaction perspectives, the hybrid approach entails stakeholders actively participating in the algorithm design process, incorporating diverse viewpoints through an open and participatory research and develop model to enhance the interpretability of medical artificial intelligence, and establishing a “humanistic” ethical framework for medical artificial intelligence. While the hybrid approach cannot entirely eliminate the algorithmic black-box phenomenon at present, the development of “algorithmic gray-box” with local transparency represents a feasible goal for designing current medical artificial moral agents.

artificial intelligence, algorithmic black-box, transparency, moral design, ethics of science and technology

doi: 10.1360/TB-2022-1320