• 专题：智能医疗的伦理与治理 •

编者按：

人工智能技术在医疗领域的广泛应用催生智能医疗这一新兴医学诊疗模式。作为以人为直接对象的技术应用，医疗技术的安全性被置于有效性之前，这致使智能医疗的风险研究伴生于技术应用扩展，从宏观的技术伦理向具体的应用领域与核心问题聚焦，成为当前人工智能应用伦理领域的热议话题。

本期专题收集了三篇智能医疗伦理与治理的研究文章，都具有较高的前沿性与现实针对性。王昱洲、廖新媛的文章从算法黑箱对医疗信任的挑战出发，基于对事后解释与“第二意见”两种解决策略的反驳，提出通过人机耦合建立医疗信任关系这一可行模式。孙丹阳、李侠的文章基于对智能心理治疗理论基础与运行机制的分析，揭示人工智能的技术性风险因心理治疗的数据敏感、对象脆弱、目标特殊等因素而存在增溢效应，提出贯穿技术研发－使用－监管全过程的风险应对方案。袁洁铃、陈海丹的文章澄清了对话智能体在抑郁症诊治应用中所面临的准确性、数据安全与隐私、固有偏见、情感鸿沟和患者依赖五大伦理难题，并从审查与评估、数据伦理、角色定位、人机交互以及监管与信任的维度构建治理策略。希望本期专题可以引发学界同仁对智能医疗这一现实问题进行更多的伦理省思与治理探索。

（专题策划：孙丹阳）

人机耦合：一种解决医疗人工智能信任问题的方案

Human-Machine Coupling: A Solution to the Problem of Trust in Medical Artificial Intelligence

王昱洲 /WANG Yuzhou[1, 2]　廖新媛 /LIAO Xinyuan[3]

（1. 北京大学外国哲学研究所，北京，100871；2. 北京大学哲学系，北京，100871；
3. 武汉大学马克思主义学院，湖北武汉，430072）
(1. Institute of Foreign Philosophy, Peking University, Beijing, 100871;
2. Department of Philosophy, Peking University, Beijing, 100871;
3. School of Marxism, Wuhan University, Wuhan, Hubei, 430072)

摘　要：信任关系在医疗临床实践中有着至关重要的作用。用于医疗诊断的人工智能，由于其算法缺乏足够的透明性，在临床应用中面临着如何建立信任的问题。这一问题可以进一步分为“医生－机器”与

基金项目：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”（项目编号：22JJD720007）；国家社会科学基金青年项目“智能算法下网络价值引导的现实困境及其破解路径研究”（项目编号：24CKS088）；北京市习近平新时代中国特色社会主义思想研究中心重大项目“构建支持全面创新体制机制研究”（项目编号：24LLZXA096）。
收稿日期：2024 年 9 月 9 日
作者简介：王昱洲（1994–）男，河南南阳人，北京大学外国哲学研究所、北京大学哲学系助理教授，研究方向为科学技术伦理、生命伦理、道德知识论。Email: yuzhouw@pku.edu.cn
廖新媛（1995–）女，福建建瓯人，武汉大学马克思主义学院讲师，研究方向为科学技术哲学。Email: liaoxinyuan@whu.edu.cn

“患者–机器”两个方面。相比于“破除算法黑箱”进路以及将人工智能视为“第二意见”的进路，“人机耦合”进路能够更好地刻画人工智能与人类的关系。根据这一进路，理想的医疗诊断人工智能应当能够作为医生认知过程的一部分，与医生共同组成一个紧密耦合的认知系统。这使得该系统，而非机器，成为需要被信任的对象。这一进路为潜在的医疗人工智能信任问题提供了一种可能的解决方案。

关键词：医疗人工智能　算法黑箱　人机信任　人机耦合

Abstract: Trust plays a crucial role in clinical medical practice. However, artificial intelligence (AI) used for medical diagnosis faces the challenge of establishing trust in clinical applications due to the lack of sufficient transparency in its algorithms. This challenge can be further divided into two aspects: the “doctor-machine” trust and the “patient-machine” trust. Compared to approaches such as “breaking the black box” and viewing AI as a “second opinion”, the “human-machine coupling” approach better characterizes the relationship between AI and humans. According to this approach, an ideal AI for medical diagnosis should become part of the doctor’s cognitive process, forming a tightly coupled cognitive system with the doctor. This makes the system, rather than the machine itself, the entity that needs to be trusted. This approach provides a possible solution to the problem of trust in medical AI.

Key Words: Medical artificial intelligence; Black box algorithms; Human-machine trust; Human-machine coupling

中图分类号：R197; TP18　DOI: 10.15994/j.1000–0763.2025.09.001　CSTR: 32281.14.jdn.2025.09.001

疾病诊断是当前医疗人工智能发展的一个重要方向。通过对照片、X光片，或CT影像进行分析，人工智能系统可以判断患者是否患有疾病，并根据其诊断结果提供治疗建议。目前，在糖尿病视网膜病变、乳腺癌等疾病筛查中，人工智能诊断的准确率已经相当于甚至超过了人类医生。而且，相较于人类医生，人工智能还能更快，更早地发现疾病。由于这些优势，许多学者认为人工智能在未来至少能够部分地取代人类医生。[1]，[2]人工智能准确且高效的分析水平，也让优质医疗有望变得更经济、更惠民。

然而，医疗人工智能在临床应用中面临着一个显著的困难——算法黑箱问题。算法黑箱的不透明性也引发了人们对人工智能的担忧和不信任。本文将在第一节简述算法黑箱问题，并从“医生–机器”与“患者–机器”两个方面论述该问题对医疗信任的挑战，分析医生与患者因何难以信任医疗人工智能。第二节将讨论两种针对信任问题的解决方案——“破除算法黑箱”进路以及将人工智能视为“第二意见”的进路，并分析二者存在的缺陷。本文第三节将基于延展认知理论（the extended cognition thesis）提出一种新的解决方案，即“人机耦合”进路。这一进路避免了上述两种进路的缺陷，因此更具发展前景。第四节为总结与展望。

一、算法黑箱问题对医疗信任的挑战

当前用于医疗诊断的人工智能主要依赖于机器学习与大量数据的训练，其内部算法通常过于复杂且不透明。这种不透明性又可以进一步分为方法论不透明性（methodological opacity）与认识论不透明性（epistemological opacity）。方法论不透明性指的是编写人工智能程序的过程本身使得算法变得不透明。由于一个人工智能系统涉及到许多算法模块，每个算法模块都需要编写大量代码，算法模块和子程序之间又需要相互关联，这个过程会不可避免地将可理解的信息掩藏在代码细节之中。因此，即便我们得到了一个人工智能系统的完整代码，方法论不透明性也会使得信息的提取与解读变得十分困难。认识论不透明性指的是算法的推理过程是不透明的。在使用医疗人工智能时，人类用户得到的仅仅是一个诊断结果。用户并不清楚人工智能生成诊断的原因，而人

工智能也无法为自己的诊断提供理由。认识论不透明性导致人工智能无法实现报告证据、理解和解释结果等主要认知功能。[3] 虽然算法黑箱问题广泛存在于各类人工智能中，但是由于医疗领域对信任关系的依赖性，算法黑箱问题对医疗人工智能的应用造成的挑战尤为严重。这两种不透明性都影响了医疗信任的建立与保持。接下来，本节将分别讨论不透明的算法黑箱对“医生–机器”与“患者–机器”信任关系带来的挑战，并回应对该挑战的反驳。

1. 算法黑箱问题与“医生–机器”信任关系

医疗人工智能的应用首先会涉及到人工智能与医生之间的合作。[4] 而这种合作成立的前提是医生需要对人工智能有一定程度的信任。如果人类医生不信任人工智能，在人工智能做出的诊断结果与自身的不一致时总是不相信机器而不去采纳人工智能的结果，那么即便人工智能的诊断准确率高于人类医生，这一优势也无法在医疗中体现出来。

当然，从整体上来看，如果人工智能诊断的整体正确率显著高于人类医生，那么医生就应当更倾向于相信人工智能而非自身的判断。但是这种统计上的差异并不足以为单次的决策提供理由。[5] 只要人工智能不是百分之百正确，那么医生总是需要去考虑其犯错的可能。而人工智能的算法黑箱问题，尤其是方法论不透明性使得人们无法判断算法是否在某个具体问题上犯错。马科·里贝罗（Marco Ribeiro）等学者给出了一个高度准确的人工智能也有可能不值得信任的例子。在这个例子中，虽然人工智能可以根据图片非常准确地区分狼和哈士奇，但是这一区分并不是由二者的生物学特征，而是依据图像背景中是否有雪来作出的。[6] 医疗人工智能很可能在影像分析中也存在类似的问题，即使用与疾病无关的因素来对其进行判断。而这些无关的因素在环境更为复杂的临床实践中并不总是与疾病同时出现的。一旦出现这种系统性的错误，医疗人工智能的应用将会导致巨大的生命健康风险。因此，只有当医疗人工智能在整体准确率较高并且推理过程合理时才应当被医生信任。而方法论不透明性使得医疗人工智能的推理过程无法得到评估，导致“医生–机器”信任关系无法被建立。

致使“医生–机器”信任关系无法被建立的另一个原因体现在诊断分歧的解决方面。当医生与医生之间对于诊断结果存在分歧时，这些分歧大多可以通过交流与沟通得以解决。在这个过程当中，医生可以表达自身判断的依据，分析分歧的原因，从而找出自己或对方判断中的问题，并对原先的决策进行调整。[7] 但是当医生与人工智能发生分歧时则无法使用同样的方式找出分歧的原因。由于算法的认识论不透明性，医生并不清楚人工智能诊断结果背后的原因，也无法通过进一步的沟通发现哪一方的判断更准确。在这种情况下，医生仅仅能够理解自己的推理过程与判断依据，而无法获知人工智能的推理过程与判断依据。根据这些信息，医生合理的决策是相信自身的判断而非人工智能的判断。这使得作为理性决策者的医生在面对诊断分歧时无法信任医疗人工智能。

2. 算法黑箱问题与“患者–机器”信任关系

在传统的医疗行为中，患者对医生的信任关系是保证医疗效率的重要因素。约书亚·哈瑟利（Joshua Hatherley）指出，信任在医患关系中具有内在的与工具性的双重意义。首先，信任是内在于医患关系中的，因为患者的脆弱性（vulnerability）使其不得不信任医生来得到治疗。其次，信任也具有工具性的价值。如果患者与医生有信任关系，则更有可能接受并遵从医生的判断，从而获得更好的治疗效果。因此，患者与医生之间信任关系的建立是医疗成功的关键。[8] 而为了建立这样的信任关系，医生需要站在患者的角度，让医疗决策符合患者的价值和优先项，让患者理解为什么要这么做，而不能是大家长式的独裁。在医疗过程中，不理解医疗决策，患者便无法做出真正的知情同意，这也意味着患者没有得到充分的尊重。因此，争取患者对医疗决策的理解与认同，不仅有助于治疗效果的提升，也是对患者自主性的尊重。

如果使用医疗人工智能代替人类医生进行决策，这意味着患者需要将部分信任从人类医

生转移到人工智能之上。但是这一转移很可能是无法达成的。当医生做出医疗诊断时，患者可以询问医生诊断的理由与依据，并向医生提出任何相关的疑问。通过对这些问题的回答，医生可以帮助患者理解其自身的病情。而算法黑箱问题的存在使得患者无法与医疗人工智能进行有效沟通，只能处于被动接受的状态。患者无法向人工智能询问其决策依据，也无法求助于医生解释其诊断结果。[9] 虽然大语言模型这类人工智能可以回应患者的提问，但是它们给予的回应仅仅是基于大数据生成的高概率回答，而没有解释得出诊断结果的真正原因。在不理解诊断结果背后的原因的情况下，患者很可能选择不去相信人工智能作出的诊断，并拒绝接受进一步的治疗。此时，患者则需要继续向人类医生问诊，导致医疗人工智能无法发挥分担医疗压力的作用。

综上所述，不论是在“医生–机器”层面还是“患者–机器”层面，人工智能的算法黑箱问题都对医疗信任的建立造成了严重挑战。

3. 基于计算可靠论的反驳及其缺陷

一些学者可能会依据计算可靠论（computational reliabilism）对上述论证提出反驳。计算可靠论认为，如果一个结果是由可靠的计算过程得到的，那么这个结果就是应当被信任的。根据计算可靠论，算法黑箱并不是一个需要被解决的问题。只要我们有理由认为这个算法是可靠的，那么我们是否知道或理解计算过程就是无关紧要的了。而算法的可靠性的建立也不需要我们理解其计算过程。胡安·杜兰（Juan Durán）和尼科·福尔马内克（Nico Formanek）认为，算法的可靠性存在四个来源：首先，算法的准确性可以通过将其与已知解决方案以及实验数据的比较来进行评估。其次，稳健性分析（robustness analysis）可以用于检验算法的结果是否表征了核心特征。第三，我们可以从历史上成功或失败的案例中推测算法的可靠性。最后，专家的知识也可以为算法的可靠性提供辩护。[10]

如果我们接受计算可靠论，那么我们就有理由在不理解医疗人工智能计算过程的情况下仍然相信其给出的诊断结果。但需要注意的是，相信人工智能生成的结果并不等同于信任人工智能。即便医生或患者相信医疗人工智能能够做出准确的医疗判断，这也不足以建立人工智能与医生或患者之间的信任关系。首先，根据罗素·哈丁（Russell Hardin）的观点，信任不仅需要信任者对被信任者的行为有所预期，还需要信任者相信被信任者具有正确的行动动机。比如，我们会预期太阳的升起和落下，但这并不意味着我们信任太阳，因为我们并不认为太阳会出于我们的利益而行动。只有当一方认为另一方具有动机做出对其有利的行为时，信任关系才会成立。[11] 其次，哈瑟利指出，信任关系与依赖关系最大的区别在于信任会为被信任方赋予责任与义务。当患者信任医生治疗其疾病时，医生便有义务提供力所能及的医疗服务，不去辜负患者的信任。因此，只有能够承担责任与义务的道德主体才能够作为被信任的对象。[8] 即便我们有理由相信人工智能得出的结果是准确的，我们也不一定有理由信任人工智能，因为这还取决于人工智能是否具有动机以及是否能够作为道德主体。

然而，上述这类问题目前仍有很大争议。这些争议不仅仅存在于哲学层面，同样也存在于科学层面。即便哲学上的争论得到了解决，对判断标准达成了共识，我们也需要结合具体的人工智能系统来判断其是否真正具有动机与主体性。而算法黑箱问题则会在很大程度上阻碍我们的判断，因为只有知道了人工智能系统的内在结构与逻辑，我们才能判断其是否符合动机与主体性的标准。因此，在解决算法黑箱问题之前，我们无法对某个人工智能系统是否值得信任的这一问题做出判断。

二、已有的两种进路及其各自的缺陷

1. 破除医疗人工智能的算法黑箱

面对算法黑箱问题，最直接的一种解决方案是通过技术手段直接对算法进行优化，从而破除黑箱。例如使用事后解释（post-hoc explanations）分析算法的输出来增强模型的可

解释性。这类解释方法可以进一步分为全局解释和局部解释。[12]

对人工智能进行全局解释旨在让人理解模型的整体运行情况，明白机器的运作逻辑。比较常见的方法有部分依赖图（partial dependence plots）、全局代理模型（global surrogate models）等。部分依赖图可以对黑箱模型输出的特征分布进行边缘化处理（marginalizing），[13]然后通过视觉化的方式，描述某类特征如何影响黑箱模型的输出。全局代理模型则是先选定一个数据集，然后输入这个数据集来收集黑箱模型的输出，接着基于这个数据集和黑箱模型的输出训练一种可解释的模型（线性模型、决策树等）。（[14]，p.139）另一种讨论得更多的进路是局部解释。局部解释不关注模型整体，而只关心特定的输入输出，因此也比全局解释更加准确。（[14]，p.18）目前，模型无关的局部可解释性方法（local interpretable model-agnostic explanations，LIME）已可用来解释目标算法的单个预测。通过对局部范围内的目标样本进行拟合，LIME可以判断出人工智能的预测结果如何受到不同因素的影响。例如，当一个医疗人工智能预测某名患者患有流感时，医生可以使用LIME分析出患者病史中的哪些症状导致了这一预测。比如，LIME可能指出“打喷嚏”和“头疼”可能是促成该预测的因素，而“未感到疲劳”可能是反对该预测的因素。[6]通过这些信息，医生可以更好地理解人工智能的预测，并判断其可靠性。

然而，上述两种事后解释方法至少存在三个共同的缺陷。首先，使用解释模型对人工智能进行成功解释的前提是用户需要信任这些解释模型。如果用户不信任这些模型，那么这些模型后续给出的解释结果也就不会被接受。[15]然而，解释模型本身也是算法黑箱。用户只能看到其解释结果，但是无法得知其做出解释的理由。而如果我们又借助于其他解释模型来解释这些解释模型，则会导致无限回溯的问题。因此，解释模型的运用并没有真正地破除算法黑箱，而只是使用一个新的算法黑箱取代了原本的黑箱而已。

其次，使用解释模型生成的解释可能不是目标人工智能产生结果的真正原因。因为它只是为不透明的模型增加了一个解释层，而不透明的模型已经完成了它的预测。人们不清楚事后解释是否充分把握了目标算法的运作特征。如果人们可以通过事后解释预测目标算法的行为，那么为什么不用事后解释替代那个目标算法呢？如果事后解释做不到这点，那就说明事后解释仍然不够完备。[16]

最后，即便我们通过事后解释达成了算法的可解释性，在临床实践中让医生和患者去理解算法仍是不切实际的。[17]解释人工智能可能需要用户具有相关理论知识，并能够使用编程软件与其他数学工具。考虑到大部分的医生与患者并非人工智能领域的专业人士，甚至可能对人工智能的基本概念都并不清楚，为了理解算法，双方都需要花费大量的时间精力。这会导致医疗效率的下降，与医疗人工智能的目的背道而驰。[18]

2. 作为“第二意见”的人工智能

还有一些学者提出，即便算法黑箱问题无法解决，我们仍可以通过恰当的医疗决策程序来避免该问题对医疗信任的影响。比如，亨德里克·肯普特（Hendrik Kempt）和萨斯基亚·纳格尔（Saskia Nagel）认为，我们不应让疾病诊断人工智能独立地做出医疗决定，而是应将其作为“第二意见”（second opinion）辅助医生的判断。在传统医疗过程中，当患者得到主治医生的诊断之后，可能会去咨询另一位医生以获取“第二意见”。如果“第二意见”与原诊断结果一致，那么“第二意见”就可以作为原诊断的佐证，增强原诊断的置信度。如果“第二意见”与原诊断结果相左，那么就需要进行进一步的讨论与检查，降低误诊的可能性。使用人工智能代替人类医生提供“第二意见”可以减少医生的工作量，但是并不会取代医生作为最终决策者的地位。主治医生依旧有最终决策权，并承担相应的责任。因此，这一方案也可以避免需要人工智能承担责任的情况。[19]

人工智能的“第二意见”与主治医生的初

步诊断相同时，该意见并不需要进一步的讨论，可以直接作为支持原诊断的依据。而当人工智能的“第二意见”与主治医生的初步诊断相矛盾时，由于医生无法通过与人工智能沟通来发现分歧的原因并解决分歧，肯普特和纳格尔提出了“分歧规则”（rule of disagreement）来处理这一问题。根据“分歧规则”，主治医生不能无视人工智能的诊断结果而坚持己见，也不能盲目听信人工智能的判断，而是需要另一位人类医生的参与来提供额外意见。在两位医生的讨论之后，主治医生才能判断是否对原诊断结果做出改变。

“第二意见”进路很好地避免了有人工智能参与的医疗过程中的信任问题：医生不需要信任人工智能，而只需要仅仅将其作为意见的提供者来看待。并且由于最后做出判断与承担责任的依旧是人类医生，病人也不需要信任人工智能。然而，这一进路同样存在缺陷。首先，将人工智能视为“第二意见”这一做法本身的合理性存疑。需要明确的是，并不是所有的意见都具有同样的认知价值（epistemic value）。比如在疾病诊断中，完全没有医学知识的人的意见则不具有任何认知价值。而这类不具有认知价值的意见即便与原诊断相左，也不需要被纳入考虑范围之内。要使得“第二意见”具有认知价值，其提出者则需要在该领域中与主治医生的认知能力相当才行。只有这样，在“第二意见”与原诊断出现分歧时，该分歧才能够被认为是需要被进一步处理的“同侪分歧”（peer disagreement）。[20] 传统医疗过程中，由于医生之间的认知能力与知识水平通常较为接近，产生的分歧也主要是同侪分歧。

然而，人类与人工智能在认知上却存在着巨大差异。人工智能在诊断疾病时，并不像医生那样，使用生理学与病理学的相关知识与原则进行诊断，而是使用数学工具对疾病进行建模，并根据输入的信息计算出最终结果。比如，当前的许多医疗人工智能系统是遵循“模糊逻辑”（fussy logic）进行机器学习的。这一过程在对疾病建模后，需要对数据进行模糊化处理（fuzzification process），使得原先清晰数据转化为模糊数值。在通过模型计算之后，再把结果中的模糊数值进行去模糊化（defuzzification），从而得出一个明确的诊断结果。[2] 这类显著的认知差异使得我们没有理由认为人工智能的判断与医生的判断具有相似的认知价值。因此，我们也没有理由认为当医疗人工智能提供的“第二意见”与医生意见相左时，该分歧是一种“同侪分歧”。

其次，这种依赖额外的人类医生来解决分歧的方式不仅没有使医疗变得更简洁高效，反而增加了其过程的复杂度。当两位医生的诊断发生分歧时，该分歧通常能够在二者之间得到解决。而根据“分歧规则”，医生与人工智能的分歧则需要另一位医生的参与，使得参与方变成了三个。在解决分歧的过程中，人工智能的参与也无疑使分歧的处理变得更加困难。而且，由于人工智能自身仍旧没有得到解释，我们也没有理由认为额外参与进来的医生有能力很好地解决主治医生与人工智能之间的分歧。因此，虽然人工智能在“第二意见”与主治医生一致时可以替代人类医生，减少人力需求，但是发生分歧时的额外困难却又抵消了这一优势。

综上所述，“破除算法黑箱”进路与“第二意见”进路都存在明显的理论与实践缺陷。

三、通过人机耦合建立医疗信任关系

上述两种对于人工智能信任问题的解决方案，都仅仅将人工智能当作一个外在于人的个体去看待。但是这一观点本身是可以被质疑的。根据延展认知理论，人们的认知并不完全发生在大脑之中，外在因素有时也可以构成人们认知过程的一部分。[21] 医疗诊断是医生的一种认知活动。当人工智能参与其中并发挥作用时，它也就成为了医生认知过程的一部分。在医疗情景中，医疗人工智能和使用它的医生存在着双向的互动。人工智能把他处理过的信息作为输入传递给与之合作的医生，医生依据这个输入进行判断产生输出，这个输出又作为整个系统（医生–机器）的输入，使得这个系

统产生输出。这种动态过程也被称为“耦合”（coupling）。在不断的交互过程中，医疗人工智能和医生将会构成一个耦合系统（a coupled system），而在这个认知系统中的人工智能发挥着不可或缺的作用，没有它会导致整个系统认知水平的降低，或者是改变整个系统的行为。

当然，不同的耦合系统之间也存在着耦合程度上的区分。[22] 理查德·赫斯明克（Richard Heersmink）提供了一个可以用于判断系统的耦合程度的框架，该框架包含八个维度：信息流（information flow）、可靠性（reliability）、持久性（durability）、信任（trust）、程序透明度（procedural transparency）、信息透明度（informational transparency）、个性化（individualization）、变换（transformation）。在这些维度上的得分越高，人机耦合的程度越深。[23] 根据这一框架，高度耦合的医生与医疗人工智能系统将具有以下特征：（1）医疗人工智能将能够让医生更快地注意到一些细节，让认知任务变得容易。医生的输出也会影响医疗人工智能下一步要处理什么输入；（2）医疗人工智能会与医生保持长期互动，且不会被轻易篡改，使医生能够可靠地获得医疗人工智能提供的信息；（3）医疗人工智能与医生的耦合关系是持久的，而非一次性的或暂时的；（4）医生会默认医疗人工智能提供的信息是真的；（5）医生并不需要有意识地注意（conscious attention）就能使用医疗人工智能，医生对医疗人工智能的感知和行动是一个程序化的、透明的过程；（6）医生可以解释并理解医疗人工智能提供的信息；（7）医疗人工智能与医生个人的行为方式相契合，无法被轻易替换；（8）医生会内化医疗人工智能的表征方式，使自身的表达和认知能力被人工智能影响。如果满足以上这八个特征，那么我们便可以认为医生与人工智能在认知上达成了紧密耦合。

当医生和人工智能紧密耦合时，“医生–机器”的信任问题便不复存在了。因为这时人工智能已然成为了医生认知的一部分，是属于医生的。两者的关系就像一个人和自己手臂的关系那样。一个人可能不了解自己的手臂的生理机制，但他可以对手臂有一种内在感知，比如即使闭着眼也知道手臂处于什么姿态。一般人不会面临是否信任自己手臂的问题，类似地，紧密耦合状态中的医生与人工智能作为一个整体，同样也不会面临信任问题。人机耦合进路与前面提到的“破除算法黑箱”进路的主要区别在于，医生在不了解医疗人工智能的底层逻辑的情况下依旧可以与之建立耦合系统。可解释的人工智能还需要具有相关知识的人类去解释它，而人机耦合系统则只依赖于医生与人工智能之间存在长期稳定的互动关系，并不会为医生带来太多的额外负担。

针对“患者–机器”的信任关系则更麻烦一些，毕竟患者与医疗人工智能并没有长期稳定的耦合关系，患者与人工智能也没有直接的沟通方式。然而，这种“患者–机器”的信任关系在医生与医疗人工智能达到高度耦合后就不再必要了。取而代之的是患者与整个耦合系统之间的信任关系。而为了使患者信任医生和医疗人工智能组成的耦合系统，处于该系统中的医生则需要成为医疗人工智能的代言人，承担起向患者解释人工智能的行为，并使整个系统取得患者信任的角色。当医疗人工智能针对患者产生输出时，处于耦合系统中的医生需要及时向患者解释输出的依据，陈述整个耦合系统对患者的诊断结论。

与“第二意见”进路类似，“人机耦合”进路也是通过将“人–机”信任转换为“人–人”信任的方式来解决患者对机器的信任问题的。但是，比起“第二意见”进路，人机耦合进路并没有将人工智能看作单独的认知主体，从而避免了认知价值判断的问题。此外，虽然人机耦合进路需要人类解释算法，但是这并没有显著增加人力需求，因为解释者本就是认知过程的一个组成部分。而且，人机耦合系统中的医生相比于另一个独立的人类医生，在解决分歧方面更具权威性。这是因为耦合系统中的医生与医疗人工智能之间的联系更为紧密，对人工智能有更深的使用体会和理解，能够以更全面的视角看待人工智能的结论。因此，对于疾病诊断人工智能的信任问题，“人机耦合”进路

是一种相比于“第二意见”进路更好的潜在解决方案。

四、结论与展望

可以预见，人工智能将在未来更紧密地参与到人类的医疗事业中，在疾病诊断方面扮演越发重要的角色。无论人机合作以何种方式出现，医疗信任的建立都是需要考虑的问题。然而，人工智能的不透明性加大了建立医疗信任的难度。面对算法黑箱这一损害信任的问题，我们不仅需要理解建立信任的难点，以及已有进路的不足；同时，我们也需要探索建立信任的新方式。目前“破除算法黑箱”进路与“第二意见”进路存在诸多认识论与实践上的困难。本文依据延展认知理论，采取了“人机耦合”进路，论证了通过医生与医疗人工智能形成高度耦合的认知系统建立医疗信任的可能性。在该耦合系统中，人工智能是医生认知的一部分，是属于医生的，因此不存在“医生–机器”的信任问题；同时，患者不需要信任人工智能，而只需要信任作为代言人的医生即可。这使得患者在得到个性化、人性化的解释之外，还可以明确自己的追责对象，从而增加对整个系统的信心。

虽然“人机耦合”在理论上存在显著优势，但是其真正的难点在于实践。这些实践难题的解决将是该进路后续研究的主要方向，这里简要列举三点。首先，建立医生与医疗人工智能紧密耦合的认知系统将是一个实践难题。因为这一方面要求医疗人工智能在疾病诊断之外还要具有特定的功能（如易于携带与操作、可进行个性化设置、难以篡改等），另一方面也需要医生积极参与使用医疗人工智能。其次，在耦合系统建立之后，如何保持人类医生在该系统中的认知主导地位同样也是一个难题。人类在耦合系统中可能会逐渐懈怠，把本该承担的认知负担转移到人工智能上，过于依赖人工智能的判断。这种担忧其实提醒着人们在提升耦合水平之外，还有必要提升耦合的质量，从而发挥耦合系统中各个部分的综合优势。例如，可以让耦合系统中的人类接受必要的技能培训，以保持对个案的敏感度，防止人类过度依赖机器，造成自身业务能力生疏，最终失去实际主导决策的能力。最后，如何让患者对医生与医疗人工智能紧密耦合的系统产生正确的认知同样是一个实践难题。这不仅需要通过教育和宣传等方式提高人们的人工智能素养，也需要医院建立合理的医患沟通渠道，方便患者更快更准确地了解与理解医疗人工智能。

［参 考 文 献］

[1] Grzybowski, A., Brona, P., Lim, G., et al. 'Artificial Intelligence for Diabetic Retinopathy Screening: A Review'[J]. *Eye*, 2020, 34(3): 451–460.

[2] Kaur, S., Singla, J., Nkenyereye, L., et al. 'Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives'[J]. *IEEE Access*, 2020, 8: 228049–228069.

[3] 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]. *Journal of Medical Ethics*, 2021, 47(5): 329–335.

[4] 侯应龙、朱明琪、宋启元 等. ChatGPT之于临床医生是助手而非替手[J]. 医学与哲学, 2024, 45（14）: 1–5.

[5] Grote, T. 'Trustworthy Medical AI Systems Need to Know When They Don’t Know'[J]. *Journal of Medical Ethics*, 2021, 47(5): 337–338.

[6] Ribeiro, M. T., Singh, S., Guestrin, C. "“Why Should I Trust You?”: Explaining the Predictions of Any Classifier'[A], *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*[C], New York: Association for Computing Machinery, 2016, 1135–1144.

[7] Kempt, H., Heilinger, J. C., Nagel, S. K. "“I’m Afraid I Can’t Let You Do That, Doctor”: Meaningful Disagreements with AI in Medical Contexts'[J]. *AI & SOCIETY*, 2023, 38(4): 1407–1414.

[8] Hatherley, J. J. 'Limits of Trust in Medical AI'[J]. *Journal of Medical Ethics*, 2020, 46(7): 478–481.

[9] 彭运朋、徐毅华. 医疗人工智能对知情同意的挑战与应对[J]. 医学与哲学, 2023, 44（10）: 25–29.

[10] Durán, J. M., Formanek, N. 'Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism'[J]. *Minds and Machines*, 2018, 28(4): 645–666.

[11] Hardin, R. *Trust and Trustworthiness*[M]. New York:

Russell Sage Foundation, 2002.

[12] Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., et al. 'From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies'[J]. *Mobile Information Systems*, 2022, 2022(1): 8167821.

[13] Confalonieri, R., Coba, L., Wagner, B., et al. 'A Historical Perspective of Explainable Artificial Intelligence'[J]. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2021, 11(1): e1391.

[14] 克里斯托弗·莫尔纳. 可解释机器学习——黑盒模型可解释性理解指南[M]. 朱明超 译, 北京: 电子工业出版社, 2021.

[15] Rudin, C. 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead'[J]. *Nature Machine Intelligence*, 2019, 1(5): 206–215.

[16] Liao, S. M. *Ethics of Artificial Intelligence*[M]. New York: Oxford University Press, 2020.

[17] 陈子瑜、程国斌. 医疗人工智能中的算法黑箱及其核心伦理问题[J]. 医学与哲学, 2024, 45（12）: 6–10.

[18] 庄昱、周程. 从政策推动到研究产出——浅析医院主导人工智能研究的技术性挑战[J]. 中国科学院院刊, 2023, 38（4）: 643–653.

[19] Kempt, H., Nagel, S. K. 'Responsibility, Second Opinions and Peer-Disagreement: Ethical and Epistemological Challenges of Using AI in Clinical Diagnostic Contexts'[J]. *Journal of Medical Ethics*, 2022, 48(4): 222–229.

[20] Christensen, D. 'Epistemology of Disagreement: The Good News'[J]. *The Philosophical Review*, 2007, 116(2): 187–217.

[21] Clark, A., Chalmers, D. 'The Extended Mind'[J]. *Analysis*, 1998, 58(1): 7–19.

[22] Vold, K., Hernández-Orallo, J. 'AI Extenders and the Ethics of Mental Health'[A], Ienca, M., Jotterand, F. (Eds.) *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*[C], Cham: Springer International Publishing, 2022, 177–202.

[23] Heersmink, R. 'Dimensions of Integration in Embedded and Extended Cognitive Systems'[J]. *Phenomenology and the Cognitive Sciences*, 2015, 14(3): 577–598.

[责任编辑 李斌]