家长主义视域下的人工智能伦理工具问题探析

An Analysis of Ethical Tools of Artificial Intelligence in the Perspective of Paternalism

孙晓宇 /SUN Xiaoyu　夏保华 /XIA Baohua

（东南大学哲学与科学系，江苏南京，211189）
(Department of Philosophy and Science, Southeast University, Nanjing, Jiangsu, 211189)

摘　要：人工智能伦理问题成为各界普遍关注的热点话题之一。各界从制定一系列的人工智能原则开始转向开发各种人工智能伦理工具。人工智能伦理工具可用于识别、处理、消除或缓解人工智能面临的伦理风险，应用于金融和医疗等多个领域，逐渐成为判断和评估人工智能是否符合伦理的标准，影响或代替人类决策。然而，从家长主义的视角来看，人工智能伦理工具正逐渐呈现出家长主义的现象，并与医疗家长主义相似，可能会影响用户的自主性，引发信任危机。由此，本文借鉴医疗家长主义的改进方式，提出遵循尊重的原则，促进和尊重用户的自主性，构建共享决策制定的模型等建议，以促进人工智能伦理工具更好地发展。

关键词：人工智能　伦理工具　家长主义　医疗家长主义

Abstract: The ethical issues of artificial intelligence have become one of the widely discussed and prominent topics of concern across various sectors. Various fields have transitioned from the establishment of a series of artificial intelligence principles to the development of diverse AI ethical tools. AI ethical tools can be used to identify, deal with, eliminate or mitigate ethical risks faced by artificial intelligence, applied in many fields such as finance and medical care, and gradually become the standard for judging and evaluating whether artificial intelligence is ethical, affecting or replacing human decision-making. However, from the perspective of paternalism, AI ethical tools are gradually taking on the phenomenon of paternalism and are similar to medical paternalism, which may affect users' autonomy and cause a crisis of trust. Therefore, we draw on the improvement of medical paternalism, and put forward suggestions such as following the principle of respect, promoting and respecting the autonomy of users, and building a model of shared decision making, so as to promote the better development of AI ethical tools.

Key Words: Artificial intelligence; Ethical tool; Paternalism; Medical paternalism

中图分类号：TP18; B82　DOI: 10.15994/j.1000–0763.2025.12.010　CSTR: 32281.14.jdn.2025.12.010

随着科技的创新发展，人工智能逐渐应用到各个领域中，渗透到人类社会的方方面面。人工智能在给人们带来便利的同时，也产生了诸多伦理风险，例如，用户隐私问题、算法公平性和可解释性等。社会各界高度关注人工智能伦理和治理，从制定大量的伦理原则和指导方针逐渐转向技术实践，[1]开发了许多人工智能伦理工具，从技术层面治理人工智能伦理问

基金项目：国家社会科学基金重大项目“技术创新哲学与中国自主创新的实践逻辑研究”（项目编号：19ZDA040）。
收稿日期：2024年9月19日；**返修日期：**2025年4月9日
作者简介：孙晓宇（1997–）女，安徽阜阳人，东南大学哲学与科学系博士研究生，研究方向为科技哲学。Email: sxiaoyu03@126.com
夏保华（1969–）男，河南商城人，东南大学哲学与科学系教授，研究方向为技术哲学与技术社会学。Email: xiabaohua111@sina.com

题。比如说，微软开发的SamrtNoise，用于防止隐私信息泄露；[2]谷歌提出了联邦学习，以保护用户的个人隐私，[3]等等。

学界也对人工智能伦理工具展开了广泛的探讨。例如，乔治·凯西斯（Georgios Kaissis）等学者通过分析联邦学习在医疗成像中的应用并与其他隐私保护方法进行对比发现，联邦学习不直接使用用户的数据，可以避免泄露用户的隐私，保护用户的数据隐私安全。[4]一些学者介绍了一种新的分布式计算工具——联邦分析（federated analytics），它和联邦学习的不同在于，联邦学习强调的是训练模型，而联邦分析注重从数据中得出结论。[5]还有学者认为联邦学习虽然可以保护用户的隐私，但在实践中存在一些瓶颈，影响其性能，由此他们提出了编码联邦学习（coded federated learning），降低联邦学习的计算延迟，从而最大限度地降低训练延迟，提高隐私保护性能。[6]还有部分学者通过研究发现，InterpretML在实践中也存在一定的局限，它对表格数据（tabular data）有更好的解释，但不适用于解释图像数据（Image data）等。[7]

然而，本文认为人工智能伦理工具不仅仅是存在技术层面的瓶颈，更重要的是人工智能伦理工具正逐渐呈现出家长主义的倾向。基于此，首先本文简要介绍了人工智能伦理工具的发展概况，并阐述了不同类型的伦理工具及其运作方式，指出目前伦理工具可以实现检测、评估和消解人工智能伦理问题的全过程，正在影响或代替人类的决策，产生家长主义的风险。在第二部分，从一般家长主义概念的层面验证人工智能伦理工具家长主义是否有意义，即使人工智能伦理工具不是严格意义上的家长主义，但它正在呈现出家长主义现象。最后，人工智能伦理工具可能会逐渐演变为伦理标准的实际掌握者，代替人类进行判断和决策，影响到人类的自主性，导致信任危机，与医疗家长主义的局限相似，因而本文借鉴医疗家长主义的改进方式，提出遵循尊重原则和构建共享决策制定的模型等措施，促进以用户为中心的人工智能伦理工具的发展。

一、人工智能伦理工具的演化及分类

1. 人工智能伦理工具的演化及特征

人工智能的发展带来了巨大的社会和经济机遇，但其产生的伦理问题也不容小觑。[8]针对人工智能伦理问题，各国政府相关部门、研究机构和科技企业等纷纷出台了一系列的人工智能伦理原则。据相关数据统计，全球发布的人工智能伦理原则或指南已有百余份。[9]有学者通过分析发现，人工智能伦理原则文件中出现了五项全球趋同的伦理原则：可解释性（或透明性）、公平性、责任、隐私和安全。[10]

然而，人工智能伦理原则主要侧重于高层次的伦理理念，[11]导致理论和实践之间存在一定的隔阂。在此背景下，人工智能伦理治理出现了新浪潮，即从伦理原则转向技术实践，开发人工智能伦理工具，用技术工具来治理特定的人工智能伦理问题，以促进人工智能伦理原则落地，实现从“是什么”（what）到“如何做”（how）的转化。人工智能伦理工具是指可以应用在人工智能系统研发、部署和应用过程中的技术，旨在将抽象的伦理原则转化为具体的技术实践，落实人工智能伦理原则，使人工智能系统更公平、更强大和更透明。[12]具体而言，人工智能伦理工具遵循特定的人工智能伦理原则，如公平、透明性、隐私和安全等，用于检测、评估和解决人工智能引发的歧视、偏见和隐私等伦理问题，以确保其行为符合人工智能伦理标准和原则，并且它们能够覆盖人工智能系统的全生命周期，包括设计、开发、部署和运行等阶段，目的是通过具体的方法将人工智能伦理原则从理论转化为实践，确保人工智能的负责任和向善发展。[13]

人工智能伦理工具被广泛应用于医疗、监督和审计等多个领域。例如，在医疗领域，伦理工具可以评估与患者相关的数据集是否符合公平标准，识别疾病特定的特征，并选择具有较少偏见问题的数据集，避免混合高度异质的数据集。[14]在数据科学和在人工智能系统的研发过程中，可以通过伦理工具预处理相关的数

据集，识别和去除偏见数据，提高模型的公平性。[15]

由此可见，人工智能伦理工具有如下特征：（1）创新性和前瞻性，人工智能伦理工具是一种新兴的人工智能伦理治理方法，将人工智能伦理理念应用于实践，并伴随着新技术的开发和对已有人工智能伦理治理技术的改进和优化；（2）应用性和实用性，能够解决具体的人工智能伦理问题，满足用户的伦理需求。如，使用可解释性工具来解释人工智能的决策过程，提高和增强系统的透明度和可解释性；使用公平性工具减少或消除算法偏见和歧视，实现公平性，等等；（3）理论指导性，人工智能伦理原则是其开发的坚实基础，为其开发和具体应用提供了指导方向，例如，根据隐私保护原则开发的隐私保护工具，用于防止数据隐私泄露等；（4）伦理性，承诺遵循了一定的伦理标准，能够解决人工智能产生的各种伦理问题，如数据隐私泄露、算法歧视和偏见、算法“黑箱”和系统透明性不足等，以落实隐私、公平性、可解释性（或透明性）、责任和安全的原则；（5）经济性，除了政府相关部门外，科技企业也是主要的开发机构，如，谷歌开发了Fairness Indicators和Federated Learning等，微软开发了Fairlearn和InterpreteML等，由此人工智能伦理工具具有一定的商业价值。

2. 人工智能伦理工具的分类及其运行方式

根据人工智能伦理工具处理的伦理问题和功能可以将其分为以下两大类（见下表1）。

（1）针对不同伦理问题的伦理工具

从表1可得，根据伦理工具解决的不同问题，可分为公平性工具、可解释性工具、隐私工具和多用途工具等。

公平性/可解释性/隐私/安全性工具。顾名思义，此类伦理工具是根据特定的伦理问题所开发的，即公平性工具是为了治理人工智能的歧视和偏见问题，例如，AI Fairness 360和Fairlearn等。以此类推，可解释性、隐私和安全性伦理工具分别是为了应对人工智能的“黑箱”问题、隐私泄露和安全性风险。

多用途伦理工具。这类伦理工具不只应用于单一的伦理问题，比如说，What-If Tool（WIT）可应用于可解释性和公平性两个方面。具体而言，在可解释性方面，WIT能够自动为用户提供加载数据集中所有特征分布的汇总统计和图表，以帮助用户解释局部和全局模型。[16]在公平性方面，WIT可以在模型上直接计算公平性指标，[16]探索不同的公平性优化策略，即用户可以根据所需的特征对已有的数据集进行处理，并使用不同的优化策略来检查每片阈值被调整后的变化，根据阈值的变化，判断和实现不同的公平。[16]

（2）不同功能的伦理工具

按照伦理工具的功能不同，可分为检测型伦理工具、消除或缓解型伦理工具和多功能型伦理工具。

检测型伦理工具。此类伦理工具只提供检测服务，用于识别和评估模型或系统是否存在伦理风险，但不具备减轻伦理风险的功能。例如，WinoMT是一个最新的自动测试套件，用于检测机器翻译系统中的性别偏见。它主要是通过将测试数据集中的英语句子（每个句子包含两个实体，一个是关于职业，另一个是基于句子上下文的代词）翻译成另一种语言的句子，然后对比翻译后的性别与原始英语句子中的性别，从而识别和评估翻译模型是否包含性别偏

表1 人工智能伦理工具的分类

	不同的伦理问题	不同的功能
伦理工具	公平性工具	检测型伦理工具
	可解释性工具	消除或减轻型伦理工具
	隐私工具	多功能型伦理工具
	安全性工具	
	多用途伦理工具	

见。[17]

消除或减轻型伦理工具是指能够提供消解或减轻人工智能伦理问题的技术。比如说，隐私保护工具PySyft，它主要是提供联邦学习、差分隐私和加密技术来达到隐私保护的目的，即允许在不拥有原始数据集和数据加密的情况下，对远程所需的数据进行分布式计算，以达到保护元数据安全的目的，[18]简而言之，通过“数据的可用不可见”以保障数据隐私的安全。

多功能型伦理工具是指既能检测和识别伦理问题，又能消除或缓解伦理问题的伦理工具。例如，Themis-ML提供了一些测量公平和缓解偏见的方法。在测量方面，Themis-ML能够衡量数据或预测潜在歧视的程度，包括群体歧视和个体歧视，分别通过平均差、一致性和情境测试分数的方法来实现；[19]在缓解偏见方面，Themis-ML利用重新标记、重新加权、采样等数据预处理方法，修改数据集，以减少模型训练过程中的歧视性预测。

综上所述，第一，人工智能伦理工具的类型繁多，但多用途和通用型伦理工具的数量较少。[20]第二，人工智能伦理工具主要是通过量化人工智能行为，来检测、识别和评估人工智能是否符合伦理，然后大多通过对数据和模型进行不同方式和程度的处理以减轻其可能带来的伦理风险。

虽然人工智能伦理工具在解决人工智能伦理问题时具有一定的客观性，但伦理工具逐渐成为判定人工智能伦理问题的手段和标准，影响或代替人类对人工智能伦理问题的决策。从家长主义的视角来看，人工智能伦理工具的操作可能是家长式地强加伦理标准，以其对伦理问题的认知方式进行实践，可能会忽视用户、不同的社会背景以及程序规则。[21]

二、家长主义视角下的人工智能伦理工具

家长主义行为通常出现在法学、政治学、哲学等领域，随着人工智能伦理工具的发展和应用，人工智能伦理工具可能也会引发家长主义风险，下面将从概念维度来分析伦理工具的家长主义现象。

1. 人工智能伦理工具家长主义的概念是否有意义

家长主义（paternalism）基于拉丁语父亲（pater）一词，最早源于罗马法，家庭中的每个成员必须服从于父亲，即家长的权力。[22]家长主义概念在使用中会产生许多争议，本文采用的是一般家长主义的定义，即一个国家或个人对另一个人的干涉，并声称是为了被干涉者的利益或免受伤害，但可能违背了被干涉人的意愿。[23]由此家长主义可被理解为，出于他人利益最大化的目的，为他人做决定，却不征求他人的意见。[24]杰拉尔德·德沃金（Dworkin Gerald）又提出了家长主义的三个条件，即干涉条件、同意条件和福利（benefit）条件，[25]具体而言是指：1）决策者A（家长式）的决策干扰了被决策者B的自由或自主性；2）A未征得B的同意而做出决定；3）A认为自己的行为会对B产生有益的作用或影响。

以上体现出一般家长主义概念的两个主要特征：第一，行为本身具有一定的强制性，家长式的干预行为会对被干预者的自主性或自由造成不同程度的限制或影响；第二，决策者的意图是善意的，家长式干预措施的最初目的是为了保障被干预者的福利或利益。

根据一般家长主义的定义，如果想要人工智能伦理工具家长主义的概念有意义，需要满足四个条件：第一，有意图地与用户进行互动；第二，干涉用户的自由或自主性；第三，未征得用户的同意；第四，对用户有益。[26]

首先，一般家长主义的行为主体是人，而人工智能伦理工具家长主义的行为主体是人工智能伦理工具，没有意图，因而人工智能伦理工具家长主义的概念必然不能严格按照一般家长主义的界定方式进行定义。但伦理工具的目标是为了解决人工智能带来的伦理挑战，增进人类福祉，具有目标导向行为，因而需在一般家长主义概念的基础上进行修改和调整，才能满足第一个条件。

其次，第二个条件并不是意味着人工智能伦理工具能够干涉用户的行动自由，它是指会

影响用户的选择行为。在伦理工具的设计过程中，工程师通过设计一些技术功能选项为用户提供服务，这些预设的功能选项可能会对用户的选择行为施加限制，在某种程度上会影响用户的自主性。[27]此外，企业在研发产品时通常会考虑自身的竞争力和盈利情况，因而伦理工具也会受到企业偏好的影响，忽视用户的价值取向和需求。[28]由此，将一般家长主义概念应用于人工智能伦理工具家长主义现象的第二个条件是满足的。

再次，关于第三个条件，虽然用户已经明确同意使用人工智能伦理工具，但这不能表示用户同意伦理工具影响自身的自主性，尤其是用户可能根本不会意识到会对自己造成影响，特别是排除特定选项后。[26]比如说，Fairlearn既不能解决由刻板印象造成的歧视和不公平，也不能实现更广泛的社会层面的公平性（正当程序、正义等），[29]即排除了这两个选项，但有的用户可能并不知道或根本意识不到伦理工具会排除哪些特定的问题，由此，可以认为伦理工具在特定的情况下，未征得用户的同意，满足第三个条件。

最后，第四个条件是福利（benefit）条件。虽然人工智能伦理工具没有意识，即没有关于“善”的概念和对人类有益的意向。但它的目的是为了落实人工智能伦理原则，治理人工智能伦理问题，促进人工智能技术向善发展，造福于人类。

综上所述，虽然目前人工智能伦理工具家长主义并不完全契合一般家长主义的四个条件，但人工智能伦理工具家长主义的概念是有一定的意义，即人工智能伦理工具正在逐渐呈现出家长主义现象，这是因为：第一，人工智能伦理工具可能会干扰或影响用户的自主性或自由，第二，在特定的情况下，未取得用户的同意或用户没有意识到；第三，伦理工具的行为是为了促进或帮助用户对人工智能伦理和治理方面的意识和规划。

2. 人工智能伦理工具家长主义的类型

一般家长主义具有不同类型，如，硬和软家长主义、强和弱家长主义、狭义和广义家长主义。不同类型的家长主义存在不同程度的区别，下面将简要介绍几种具有代表性的家长主义类型，并讨论人工智能伦理工具家长主义的类型。

狭义和广义家长主义。狭义家长主义主要是指国家家长主义，即法律层面的强制性措施，广义家长主义不仅包括国家家长主义，还包括其他领域的家长主义，如，医疗家长主义、个人家长主义等。[30]

硬和软家长主义两者具有显著的区别。硬家长主义是指对主体行为自由的干涉，不管被干预者是否能自主地做出决策，违背了主体的意愿。而软家长式干预是为了检验或确定在给定情况下主体是否具有决策自主性，或是为了提高主体的自主性。[31]德沃金引用密尔的例子来阐释，当一个人在不知情的情况下，走过一座受损的桥，我们无法告知其危险性（因为语言不通），及时阻止的行为属于软家长主义，因为没有违背此人的真实意愿；但如果此人是为了自杀才走过这座危险的桥，那么软家长主义就会允许此行为，而硬家长主义则不管此人是过危险的桥，还是为了自杀，都会进行阻止。[23]软家长主义还有一种形式为助推（nudge）家长主义，“助推”是指既不用激励机制，也不采取直接的禁止措施，而是由某种微妙的（操纵性）影响或引导人们的决策。[32]，[33]

强和弱家长主义的不同在于被干预者的目的和手段。强家长主义认为被干预者的目标或结果是错误的，必须加以干预。而弱家长主义是为了帮助被干预者实现其目的，认为被干预者为追求目的而采取的手段不足以实现目标，但不怀疑其目标的合理性。[23]比如说，当X想要乘坐公共汽车去离家较远的游乐场，但Y阻止其乘坐公共汽车，开车载X去游乐场，Y的行为属于弱家长主义，因为Y认为乘坐公共汽车花费时间较长，但如果Y阻止X是因为去游乐场玩浪费时间，否定了X目的合理性，那么Y的行为就是强家长主义。

总体而言，不同类型的家长主义虽略有区别，但共同点是：第一，干预者的动机是善意的，是为了被干预者的利益；第二，干预者采取的

措施会对被干预者的自主性或自由产生一定的影响或限制。

从一般家长主义的类型划分来看，人工智能伦理工具家长主义属于广义的家长主义，但目前不具有强制性，也没有违背用户的真实意愿，不能归属于强家长主义。此外，用户想要解决人工智能伦理问题的目的是合理的，但可能采取的方法无法达到目的，由此人工智能伦理工具家长主义可认为是一种弱家长主义。值得注意的是，人工智能伦理工具对于人工智能伦理问题（量化的）的界定（或判定标准）是预先设定的（或编程的），即人工智能伦理工具目标导向行为的外部预定义目标，简而言之，人工智能伦理工具根据外部设定（不是由系统用户设定的）的伦理标准或规定来检测和解决人工智能伦理问题，由此可能会带来一些伦理问题。

三、人工智能伦理工具家长主义的问题和缓解建议

如前文所述，虽然人工智能伦理工具并未表现出明显的家长主义，但仍不能忽视其呈现的家长主义趋势和风险。由于目前相关的研究较少，本文主要借鉴医疗家长主义，来探讨人工智能伦理工具家长主义可能带来的问题以及如何进行缓解。

1. 人工智能伦理工具家长主义可能带来的伦理问题

有观点认为目前家长主义的主要问题是：它在很大程度上会破坏人的自主性，[34]影响平等的合作关系。[35]在医疗领域中，学者们多从个人自由和自主性的角度出发，对医疗家长主义进行讨论。[36]医疗家长主义的核心是医生从病人的利益出发代替病人决策，但可能会忽视患者的需求和意愿或是拒绝患者参与决策过程，违背了尊重整体自主权的原则，[37]影响医患之间的信任关系。人工智能伦理工具家长主义在使人工智能系统符合伦理，保护用户利益的名义下，通常也会忽视用户的自主选择和决策权，影响其自主性，引发用户对人工智能伦理工具的不信任感。

（1）影响用户的自主性。家长主义在一定程度上会影响被干预者的自主性，因为它通常为了他人的利益，干预其行为或决策，包括限制他人的选择权、代替其做决策或直接干涉其行为等。对于人工智能伦理工具家长主义而言，它往往也会以受托人的利益为由，[38]即用户的利益，影响人工智能伦理问题的决策过程，或是直接代替用户进行决策。以公平性工具Aequitas为例，它是依据公平性原则设计和开发出的一个偏见和公平的审计工具。首先基于一些敏感属性（年龄、性别和种族等），Aequitas定义了不同的群体（groups），每个群体由相同实体（数据点）构成，然后使用预测普遍性（Predicted Prevalence）和预测阳性率（Predicted Positive Rate）两个主要指标来定义分布群体度量，接着计算不同群体的假阳性率和假阴性率，最后比较不同群体的度量指标值与参考群体的差异来定义偏见，[39]比如说，如果一个群体特定属性值（如，性别）的假阳性率远高于其他群体，这就表明系统对该群体存在性别层面的偏见。在实际应用过程中，用户只需要从被审计的系统上传数据，配置偏见指标，然后该工具可以自动生成关于该系统的偏见报告。[39]

由此可以看出，第一，人工智能伦理工具的研发和目标设定主要是依据相关的人工智能伦理准则，但这在一定程度上可能会忽视用户层面的作用和影响。第二，以公平性工具为例，它预先定义了公平和偏见等概念，但公平概念具有争议性，并且不同用户对公平性的要求和理解可能也不同，[40]而公平性工具直接对公平性原则进行解读和界定，忽略了用户对公平的需求和选择权，甚至在某种程度上伦理工具可能会代替伦理学家对公平性原则的解读和应用。第三，伦理工具能够自动地对人工智能伦理问题进行检测和审计等，成为人工智能伦理问题评估和判断的标准，并认为可以帮助用户做出“最好的”的选择，从而代替用户进行评估和决策，削弱了用户的自主决策权和参与权。第四，伦理工具的研发仍是政府相关机构和企

业占主导地位，导致用户可能无法真正地理解伦理工具，只能被动地接受这些伦理工具所输出的结果和建议。

（2）引发信任危机。家长式的医疗决策忽视了患者的选择权和知情权，破坏了医患之间的信任关系，而人工智能伦理工具家长主义也可能会引发信任危机。一方面，在家长主义模式下，伦理工具根据自身的评估和判断，直接给出决策结果，例如，Aequitas直接生成偏见报告，但用户很难理解其如何得出人工智能系统是否存在偏见或解决偏见问题，或系统已实现公平性的结论，缺乏与用户进行充分的信息交互，未能尊重用户的意愿，导致用户质疑其结论的合理性，进而产生不信任感；另一方面，伦理工具限制了用户的选择，用户只能根据其提供的各种指标或算法来检测和解决人工智能伦理问题，但其实际的实施结果可能并不符合用户的实际需求和偏好。比如说，隐私保护工具在应用时，重点保护的是直接收集获取的各种数据隐私信息或用户提供的原始数据信息（如，性别、年龄、浏览记录和购买记录等），忽视了对预测隐私（基于已获取得到的各种数据信息，进一步分析和预测得到的其他隐私信息。如，根据用户的浏览和购买记录，分析得到用户的购买力和购买偏好等）的保护，[41]无法实现有效的隐私保护，难以消除用户对隐私泄露和数据不当使用问题的担忧和恐慌，从而可能进一步加剧信任危机。[42]

2. 缓解策略

为了使人工智能伦理工具更好地服务于用户，本文主要借鉴医疗家长主义的改进方式，提出以下建议，尽可能减轻人工智能伦理工具家长主义及其负面影响。

首先，遵循尊重原则，以人为中心。在医疗领域，尊重的原则最重要是提供信息并允许患者自主决策，尊重患者的自主性，主要包括关怀、同理心、尊严和关注需求等多维度因素。[43]为提升医疗家长主义的合理性，医生需要尊重和支持患者的自主性，以患者为中心，[44]但并不意味着盲目地扩大患者的自主性，而是平衡患者的自主性与适度的医疗干预。[45]医生还应与患者建立信任关系，真正了解患者的意愿、价值观和生活经历，在决策中将患者视为平等的合作伙伴。[28]因而，伦理工具也需要将用户的利益考虑在内，尊重和支持用户的自主性，以用户为中心，提供适度的干预措施，帮助用户检测和解决人工智能伦理问题，而不是由伦理工具单方面判定人工智能系统是否符合伦理，即伦理工具可以通过提供可选项、解释后果，确保用户能够理解其运行方式和支持用户的意愿和需求。

其次，构建共享决策制定（Shared Decision Making, SDM）的模型。SDM的一个核心特征是以患者为中心，[46]以共同决策的形式使患者参与其中。通过共同决策，医生可以帮助患者了解他们的价值观和偏好在做出最适合他们的决定中的重要性，不仅强调患者的参与和价值，还体现出医生专业知识的重要作用。[47]，[48]伦理工具为用户提供与人工智能伦理问题相关的技术知识，但也要了解用户的需求，同时需要参考其他专业人员对人工智能是否符合伦理标准的意见，从而既保障了用户的自主性，又保留了技术的专业性。[28]

再次，加强人工智能伦理工具与用户的交互性。用户应当积极参与到伦理工具的具体研发过程中，例如，将用户的道德和价值观念纳入到人工智能伦理工具的设置中，用户可以在公平、数据透明、隐私保护等方面调整伦理工具。

最后，加强伦理咨询和审查机制。虽然一部分伦理工具可以检测和识别人工智能伦理问题，但由于其不能涵盖所有潜在的伦理风险，因此仍需建立外部的监督和审查机制，检测和评估伦理工具的实践效果。此外，伦理工具要充分考虑到用户的利益，优化反馈机制，针对用户的反馈调整伦理工具的设计，为用户提供易于理解的信息和连续的指导与帮助，以满足用户实际的伦理需求。

总之，通过遵循尊重原则，促进和尊重用户的自主性，构建共享决策制定的模型，加强伦理工具与用户的交互性，确保用户和其他人员在决策过程中发挥积极作用，使伦理工具的

行为符合用户的利益和伦理规范，构建以用户为中心的人工智能伦理工具，以缓解人工智能伦理工具家长主义可能带来的负面影响。

结　论

人工智能的应用带来了一系列的伦理问题，为了促进人工智能可持续发展，各界开发了许多人工智能伦理工具，例如，公平性工具、可解释性工具等。然而，目前很多研究主要关注从技术层面改善伦理工具，忽视了伦理工具可能会产生家长主义的风险。虽然目前伦理工具尚未表现出明显的家长主义，但正逐渐呈现出家长主义的倾向，并与医疗家长主义相似，忽视了用户的需求和意愿，影响到用户的自主性，进而可能引发信任危机。为了进一步推动人工智能伦理工具的发展，需要遵循尊重的原则，尊重用户的自主性，以用户为中心，促进和支持用户参与决策，增强外部的伦理监督和审查机制，不断优化升级人工智能伦理工具等，以此尽可能地缓解人工智能伦理工具家长主义可能产生的问题，使之更好地服务于人类。

［参 考 文献］

[1] Morley, J., Floridi, L., Kinsey, L., et al. 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices'[J]. *Science and Engineering Ethics*, 2020, 26(4): 2141–2168.

[2] Qiu, L., Fan, Y. 'Smart Noise Jamming Suppression Method Based on Target Position Estimation'[EB/OL]. https://ieeexplore.ieee.org/document/9824130. 2022–05–20.

[3] Bonawitz, K. 'Towards Federated Learning at Scale: System Design'[J]. *Proceedings of Machine Learning and Systems*, 2019, 1: 374–388.

[4] Kaissis, G. A., Makowski, M. R., Rückert, D., et al. 'Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging'[J]. *Nature Machine Intelligence*, 2020, 2(6): 305–311.

[5] Wang, D., Shi, S., Zhu, Y., et al. 'Federated Analytics: Opportunities and Challenges'[J]. *IEEE Network*, 2021, 36(1): 151–158.

[6] Ng, J. S., Lim, W. Y. B., Xiong, Z., et al. 'A Hierarchical Incentive Design Toward Motivating Participation in Coded Federated Learning'[J]. *IEEE Journal on Selected Areas in Communications*, 2021, 40(1): 359–375.

[7] Yu, H. Q., Alaba, A., Eziefuna, E. 'Evaluation of Integrated XAI Frameworks for Explaining Disease Prediction Models in Healthcare'[A]. Qi, J., Yang, P. (Eds.) *International Workshop on Internet of Things of Big Data for Healthcare*[C], Cham: Springer Nature Switzerland, 2023, 14–28.

[8] Harari, Y. N. 'Reboot for the AI Revolution'[J]. *Nature*, 2017, 550(7676): 324–327.

[9] Huang, C., Zhang, Z., Mao, B., et al. 'An Overview of Artificial Intelligence Ethics'[J]. *IEEE Transactions on Artificial Intelligence*, 2022, 4(4): 799–819.

[10] Jobin, A., Ienca, M., Vayena, E. 'The Global Landscape of AI Ethics Guidelines'[J]. *Nature Machine Intelligence*, 2019, 1(9): 389–399.

[11] Ayling, J., Chapman, A. 'Putting AI Ethics to Work: Are the Tools Fit for Purpose? '[J]. *AI and Ethics*, 2022, 2(3): 405–429.

[12] Durmus, M. 'An Overview of Some Ethical-AI Toolkits'[EB/OL]. https://www.aisoma.de/an-overview-of-some-ethical-ai-toolkits/. 2020–12–30.

[13] Prem, E. 'From Ethical AI Frameworks to Tools: a Review of Approaches'[J]. *AI and Ethics*, 2023, 3(3): 699–716.

[14] Arias-Garzón, D., Tabares-Soto, R., Bernal-Salcedo, J., et al. 'Biases Associated with Database Structure for COVID-19 Detection in X-ay Images'[J]. *Scientific Reports*, 2023, 13(1): 3477.

[15] Saplicki, C. 'Fairness in Machine Learning: Pre-Processing Algorithms'[EB/OL]. https://medium.com/ibm-data-ai/fairness-in-machine-learning-pre-processing-algorithms-a670c031fba8. 2023–03–13.

[16] Wexler, J., Pushkarna, M., Bolukbasi, T., et al. 'The What-if Tool: Interactive Probing of Machine Learning Models'[J]. *IEEE Transactions on Visualization and Vomputer Graphics*, 2019, 26(1): 56–65.

[17] Kocmi, T., Limisiewicz, T., Stanovsky, G. 'Gender Coreference and Bias Evaluation at WMT 2020'[EB/OL]. https://arxiv.org/abs/2010.06018. 2020–10–12.

[18] Ziller, A., Trask, A., Lopardo, A., et al. 'Pysyft: A Library for Easy Federated Learning'[A], Kacprzyk, J., et al., (Eds). *Federated Learning Systems: Towards Next-Generation AI*[C], Cham: Springer Nature Switzerland, 2021, 111–139.

[19] Zliobaite, I. 'A Survey on Measuring Indirect Discrimination in Machine Learning'[EB/OL]. https://arxiv.org/abs/1511.00148. 2015–10–31.

[20]Bellamy, R. K. E., Dey, K., Hind, M., et al. 'AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias'[J]. *IBM Journal of Research and Development*, 2019, 63(4/5): 1–15.

[21]Morley, J., Elhalal, A., Garcia, F., et al. 'Ethics As a Service: A Pragmatic Operationalisation of AI Ethics'[J]. *Minds and Machines*, 2021, 31(2): 239–256.

[22]Saller, R. P. 'Pater Familias, Mater Familias, and the Gendered Semantics of the Roman Household'[J]. *Classical Philology*, 1999, 94(2): 182–197.

[23]Dworkin, G. 'Paternalism'[EB/OL]. https://plato.stanford.edu/entries/paternalism/. 2020–09–09.

[24]Khadilkar, P., Jagtap, S. 'Can Design Be Non-Paternalistic? Conceptualizing Paternalism in the Design Profession'[J]. *She Ji: The Journal of Design, Economics, and Innovation*, 2021, 7(4): 589–610.

[25]Düber, D. 'The Concept of Paternalism'[A], Schramme, T. (Ed.) *New Perspectives on Paternalism and Health Care*[C], Cham: Springer International Publishing, 2015, 31–45.

[26]Kühler, M. 'Exploring the Phenomenon and Ethical Issues of AI Paternalism in Health Apps'[J]. *Bioethics*, 2022, 36(2): 194–200.

[27]Lockton, D., Harrison, D., Stanton, N. A. 'The Design with Intent Method: A Design Tool for Influencing User Behaviour'[J]. *Applied Ergonomics*, 2010, 41(3): 382–392.

[28]Lorenzini, G., Arbelaez, O. L., Shaw, D. M., et al. 'Artificial Intelligence and the Doctor-Patient Relationship Expanding the Paradigm of Shared Decision Making'[J]. *Bioethics*, 2023, 37(5): 424–429.

[29]Bird, S., Dudík, M., Edgar, R., et al. 'Fairlearn: A Toolkit for Assessing and Improving Fairness in AI'[EB/OL]. https://www.microsoft.com/en–us/research/uploads/prod/2020/05/Fairlearn_whitepaper.pdf. 2020–05–15.

[30]Carney, T., Bigby, C., Then, S. N., et al. 'Paternalism to Empowerment: All in the Eye of the Beholder? '[J]. *Disability & Society*, 2023, 38(3): 503–523.

[31]Beck, B. 'Paternalism and Liberty/Autonomy as Dialectically Related Concepts'[J]. *Zeitschrift für Ethik und Moralphilosophie*, 2023, 6(2): 223–237.

[32]Thaler, R. H., Sunstein, C. R. 'Libertarian Paternalism'[J]. *American Economic Review*, 2003, 93(2): 175–179.

[33]Kühler, M., Mitrović, V. 'For Your Own Good? History, Concept, and Ethics of Paternalism: Part I'[J]. *Zeitschrift für Ethik und Moralphilosophie*, 2023, 6(1): 123–126.

[34]Rochi, M. 'Technology Paternalism and Smart Products: Review, Synthesis, and Research Agenda'[J]. *Technological Forecasting and Social Change*, 2023, 192: 122557.

[35]Bladon, H. 'Avoiding Paternalism'[J]. *Issues in Mental Health Nursing*, 2019, 40(7): 579–584.

[36]Chin, J. J. 'Doctor-Patient Relationship: From Medical Paternalism to Enhanced Autonomy'[J]. *Singapore Medical Journal*, 2002, 43(3): 152–155.

[37]George, A. S. H., Shahul, A., George, A. S. 'An Overview of Medical Care and the Paternalism Approach: An Evaluation of Current Ethical Theories and Principles of Bioethics in the Light of Physician-Patient Relationships'[J]. *Partners Universal International Research Journal*, 2022, 1(4): 31–39.

[38]Komrad, M. S. 'A Defence of Medical Paternalism: Maximising Patients' Autonomy'[J]. *Journal of Medical Ethics*, 1983, 9(1): 38–44.

[39]Saleiro, P., Kuester, B., Hinkson, L., et al. 'Aequitas: A Bias and Fairness Audit Toolkit'[EB/OL]. https://arxiv.org/abs/1811.05577. 2019–04–29.

[40]Sun, X., Ye, B., Xia, B. 'The Problem of Fairness in Tools for Algorithmic Fairness'[J]. *AI and Ethics*, 2024, 1–11.

[41]Mühlhoff, R. 'Predictive Privacy: Towards an Applied Ethics of Data Analytics'[J]. *Ethics and Information Technology*, 2021, 23(4): 675–690.

[42]Sun, X., Ye, B. 'Privacy Preserving or Trapping? '[J]. *AI & SOCIETY*, 2024, 39(3): 1369–1379.

[43]Dickert, N. W., Kass, N. E. 'Understanding Respect: Learning from Patients'[J]. *Journal of Medical Ethics*, 2009, 35(7): 419–423.

[44]Hirsch, A. 'Relational Autonomy and Paternalism—Why the Physician-Patient Relationship Matters'[J]. *Zeitschrift für Ethik und Moralphilosophie*, 2023, 6(2): 239–260.

[45]Song, J. 'Autonomy and Intervention in Medical Practice'[J]. *Open Journal of Philosophy*, 2018, 8(3): 294.

[46]Sandman, L,, Munthe, C. 'Shared Decision Making, Paternalism and Patient Choice'[J]. *Health Care Analysis*, 2010, 18: 60–84.

[47]Barry, M. J., Edgman-Levitan, S. 'Shared Decision Making—the Pinnacle of Patient-Centered Care'[J]. *New England Journal of Medicine*, 2012, 366(9): 780–781.

[48]Forte, D. N. 'Shared Decision-Making: Why, for Whom, and How?'[J]. *Cadernos de Saúde Pública*, 2022, 38: e00134122.

[责任编辑　李斌]