卫生信息化国际发展动态

北京市卫生健康大数据与政策研究中心 2024.10.25

（十） AI 的伦理

1. **标题：** 人工智能在信息传播中二元性的伦理问题

来源： JMIR AI.

时间： 2024 年 10 月.

链接： https://ai.jmir.org/2024/1/e53505.

概要： 在当今数字时代中，我们会发现自己正处在“信息传染”之中，这种现象的最大特点就是无论信息准确与否都会被迅速扩散，而通用社交媒体和在线平台的快速传播更会促进这些信息的传播。人工智能的快速发展及其在当代社会各个领域的融合，标志着我们进入一个前所未有的技术进步时代，而公司和企业大力宣传和努力推广又对在公共卫生和大众生活中使用人工智能产生了更多积极影响，但与此同时，也产生了一些令人困惑的二元性信息，让人开始担忧人工智能可能会被用来制造大量令人信服的虚假信息。这种二元性就很快引出一个较为复杂的伦理挑战。为了应对这一挑战，人们开始利用信息透明度、内容监管和信息素养等不同维度来评估人工智能制造信息传播的伦理问题。本文也在 Siala 和 Wang 确定的核心内容（公平、透明、可信、问责、隐私和同理心）的“伦理需求”基础上提出了一组可行策略，以减少人工智能制造的虚假信息对公共健康产生的负面影响。文章首先指出保持训练数据集的透明度和开放性对于培养人们对人工智能开发和部署的信任、保持评估质量和代表性的独立性、防止出现偏差传播、使开发人员对其创作等伦理影响都非常重要；接着陈述了采用有效的防御策略和实时干预措施（内容审核、身份验证、集成检验工具、可疑内容报告机制等）对于减少人工智能生成虚假信息和有害叙述的重要性；然后讨论了提升公众健康和媒体领域内信息素养和发展批判性思维技能的必要性；最后指出对于人工智能在信息生产和传播的伦理研究都需要持续警觉、创新思路和多方协作，这不仅有利人工智能的未来，也将决定我们信息生态系统的完整性和整个社会的复原力。

2. **标题：生成式人工智能在心理健康信息大众化中的伦理观察**

来源： JMIR MENTAL HEALTH.

时间： 2024 年 10 月.

链接： https://mental.jmir.org/2024/1/e58011.

概要：信息大众化是指使广大民众更易获取、更具包容性和更透明的知识的过程。在过去几个世纪里，知识的获取、传播和使用已发生了重大改变，从只有少数特权人士才能掌握的内部消息，发展到报纸和期刊等印刷品的广泛发行，再到互联网的电子信息普及，再到基于维基平台的参与式的信息构建和分享，现在则已发展到利用生成式人工智能大规模定制、按需、个性化推送信息和知识。生成式人工智能的日益普及能够使广大民众更方便地获取知识，但是这也引起人们对潜在的滥用权威和程序操纵这一过程的严重担忧。ChatGPT(OpenAI)自 2022 年 11 月推出以来，已在心理健康信息大众化方面展现出了巨大潜质，能解决全球心理健康专业人员短缺、重塑心理健康护理、提高诊断准确性、改善治疗个性化、提高心理健康服务的整体可及性、促进心理健康教育和意识等多类棘手问题，但是也同时带来了一些风险，特别是在治疗和个性化心理健康干预等方面。本文将从伦理角度对“生成式人工智能如何在心理健康信息大众化中进行负责任的设计、集成和使用”进行研究。文章首先简单介绍了信息大众化的发展历程，并强调虽然生成式人工智能技术为知识大众化提供了前所未有的新途径，但是需要更为谨慎地考量其伦理问题；随后文章详细列举了生成式人工智能在心理健康信息大众化方面的六大优势（无障碍、个性化回复、多元、促进平等与缩小社会差距、医患互动、扁平化等级）和可能面临的四类风险（公司集中化、信息不透明、人们对人工智能的误解、监管问题）；最后文章在研究的基础上给出了一份人工智能在心理健康信息大众化上应用策略的调查问卷，用于评估基于人工智能的心理健康信息大众化的应用，以确保应用不仅在技术上合理，而且在伦理上也有据可考，并是以患者为中心的；另外，文章还强调要采用一种平衡且合乎伦理的方法来将生成式人工智能整合到心理健康应用中，并呼吁对心理健康领域采用生成式人工智能的人或机构应持谨慎且乐观的态度，并倡导心理健康专业人员积极参与到生成式人工智能的发展中，以利生成式人工智能与专业的有机结合。

（徐健编辑）

译文一：

人工智能在信息传播中二元性的伦理问题

Federico Germani，Giovanni Spitale，Nikola Biller-Andorno，MHBA，徐健（译）

1. 简介

在当代数字环境中，我们发现自己正处于“信息传染病”之中，这种现象的特点就是无论信息是准确的还是误导性的都会迅速扩散，而社交媒体和互联网平台的快速传播更是促进了这些信息的传播。“信息传染病”一词源于非典暴发期间，并在 COVID-19 大流行期间得到关注。它已被用于公共卫生紧急情况和与健康信息相关的情况，但它的应用范围还远不止于此。通常信息传染病会与流行病一起发生，但是信息传染病的现象并非局限于公共卫生事件，比如，英国脱欧公投或 2016 年美国总统选举。一般来说，信息传染病会造成严重风险，因为虚假信息和错误信息的传播可能会产生深远的后果。特别是对于公共卫生和公共机构的稳定，这反过来又可能对公共卫生产生不利影响。本文中虚假信息是指故意制造或传播的虚假或误导性信息。相比之下，错误信息是在不知其不准确的情况下共享的虚假或误导性信息，这意味着它无意损害个人或公共健康。人们确实担心人工智能（AI）系统可能会被用来大规模地产生令人信服的虚假信息。事实上，人工智能工具可以用来加速虚假信息的传播，或产生（虚假信息）内容，或两者兼而有之。后果可能包括破坏对机构的信任，包括公共卫生机构，并加剧社会两极分化，直接影响公共卫生结果和民主进程。因此，世界经济论坛将虚假信息和错误信息，包括人工智能驱动的虚假信息和错误信息，列为短期内对人类最相关的威胁和中期最大的威胁之一。

人工智能的快速发展及其在当代社会各个领域的融合，标志着一个前所未有的技术进步的时代。在各种各样的人工智能应用中，自然语言处理模型的兴起引起了人们的广泛关注。这种技术进步的著名例子是 OpenAI 开发模型，如 GPT-3 和 GPT-4，因其在生成文本方面的非凡熟练程度而闻名，这些文本无缝地模拟了人类交流中固有的语言复杂性、细微差别和连贯性。然而，随着这些人工智能系统的成熟，一种令人困惑的二元性逐渐显现出来——它们是能够明确和模糊其导航的信息情形的工具，对公共健康具有潜在的重大积极或消极影响。人工智能的二元性，其特点是其生成信息和虚假信息的强大能力，提出了复杂的伦理问题。事实上，这些系统在

生成非常接近人类表达的内容方面的功效不仅创造了创新沟通的机会，而且还带来了与虚假信息和错误信息相关的可怕风险，以及信息生态系统内信任的潜在侵蚀，这种风险被认为是对公共健康的严重威胁对于最大限度地减少和预测公共卫生危机的影响所需的信息传染病管理实践至关重要。为了应对这些伦理挑战，至关重要的是检查人工智能在错误信息讨论中引入的维度。透明度、内容监管和培养信息素养等关键方面对于理解人工智能在塑造信息传播方面的伦理作用至关重要。

本文试图利用一项研究的经验见解来阐明这些伦理维度，重点关注 GPT-3 生成健康相关内容的能力，这些内容比人类生成的内容更好地提供信息和误导。我们认为，人工智能迅速融入社会凸显了不仅要探索其伦理影响，还要制定审慎的战略，以利用其潜在的社会效益和保护公众健康，同时积极应对潜在风险的重要性。

2. 伦理原则

在探索人工智能复杂的前景及其对信息传播的影响时，有必要建立一个要坚持的伦理原则的基本框架，以指导、理解和评估处理人工智能在信息中可能的双重用途所需的策略生产及其对公众健康的负面影响。最近的系统回顾描绘了人工智能伦理文献中出现的“伦理特征”。基于纳入研究的 253 个文献，本综述的作者确定并定义了 6 个对于塑造人工智能在医疗保健中的作用至关重要的核心指标。第一个核心指标是公平，强调医疗保健领域的人工智能应确保每个人都能平等地获得医疗保健，而不会造成健康差异或歧视。第二个是透明度，这是人工智能在医疗保健领域面临的一个关键挑战。这意味着能够解释和验证人工智能算法和模型的行为方式，使人工智能在医疗保健领域更容易被接受、监管和使用。第三个是诚信；涉及在医疗保健中使用人工智能的各方（在审查中包含的研究中，通常是医疗保健专业人员和患者）需要将其视为值得信赖的。可信度可以来自技术教育、健康素养、临床审计和透明治理等。第四个是人工智能的问责制，这要求人工智能系统能够在提示时解释其行为，它还包括防止对用户和其他人造成伤害的安全性。第五是隐私，这意味着保护通过人工智能系统处理的用户的个人信息并尊重他们的人权，确保人工智能系统不侵犯他们的隐私。最后一个是同理心，它可以在医疗保健领域带来更多支持和关怀关系。基于这六个核心概念（被视为人工智能在医疗保健领域的总体目标），本文提出了自己的反思和框架，特别针对人工智能在信息和虚假信息传播方面的二

元性及其对公共卫生（新兴市场的一个特定部门）的影响。人工智能在医疗保健领域的应用，世界卫生组织最新的大型多模式模型指南已经考虑了这一领域（尽管没有深入讨论）。基于迄今为止概述的伦理框架，特别是深入研究人工智能在信息和虚假信息传播中的使用背景，我们认为透明度和开放性是人工智能伦理实施的基本原则。随着人工智能系统成为塑造信息格局不可或缺的一部分，通过提高透明度，利益相关者可以理解人工智能生成内容的机制，从而对其可信度和潜在偏见进行知情评估和外部评估。开放性（即数据和代码的可访问性）被认为是透明度的必要条件，而透明度又通过伴随着数据和代码的可用性以及一层解释和动机的审查来补充开放性，从而允许将开放数据和代码，以及开发和设计选择。问责机制应伴随透明度，为人工智能应用的结果建立明确的责任链。这促进了人工智能的伦理标准，并减轻了与虚假信息和错误信息相关的风险。Siala 和 Wang 提出的框架，除了透明度、开放性和问责制之外，还强调了确保人工智能系统不会永久存在或加剧现有社会不平等的重要性。在信息传播的背景下，这一原则需要认真考虑人工智能如何无意中放大某些观点或边缘化其他观点。这对于公共卫生尤其重要，因为虚假信息和错误信息的负面影响在缺乏信息素养的边缘化和弱势社区中被放大，而信息素养可以保护他们免受不健康信息生态系统的影响。评估人工智能生成内容的公平性涉及解决算法偏差、文化敏感性和代表性的包容性。重要的是，作为公平的一个要素，人工智能在信息空间中的伦理部署应优先考虑用户赋权，培养批判性思维和信息素养。因此，人工智能系统应该充当增强人类决策和信息理解的工具，而不是命令叙述——这确保人工智能在尊重人类自主权的同时为公共卫生做出积极贡献。

在以下各节中，我们将重点关注上述原则的实际应用。我们的目标是为信息生产中使用人工智能所带来的伦理挑战提供解决方案，总体目标是减轻其对公共健康的不利影响。

3. 训练数据集的透明度和开放性

与之前关于透明度和人工智能的研究一致，也是我们之前关于伦理原则的部分，我们首先建议的基本伦理原则（可能也是最相关的一个）是透明度，该原则在人工智能驱动的虚假信息和错误信息的背景下也有效。这一原则的核心在于认识到用于开发生成人工智能模型的训练数据集在塑造这些系统的能力和内部偏差方面发挥着

至关重要的作用。训练数据集是输入数据与相应的所需输出配对的集合；在训练过程中，模型学习数据中的模式和关系，学习在接触新的、未见过的数据时做出准确的预测或生成所需的输出。训练数据集的质量和多样性显着影响模型的性能。这些数据集通常是在线提供的大量文本存储库，构成了人工智能模型生成诸如人类文本等内容的来源。然而，围绕训练数据集的组成、来源和管理方法的这种不透明性引起了紧迫的伦理问题。本质上，人工智能模型是对其进行训练的语言的统计表示。因此，他们获取的数据的质量、多样性和代表性深刻地影响着他们的产出。但是危险在于，人工智能模型缺乏固有的伦理或伦理判断，反映了训练数据中存在的偏见、不准确和成见。因此，如果这些数据集在构建时没有考虑到公平的伦理原则，并且本身受到虚假信息、错误信息或偏见的损害，那么人工智能系统将无意中复制并延续这些缺陷。必须强调的是，研究已经广泛阐明了人工智能系统中的偏见问题，揭示了这些偏见的深远影响。例如，通过无监督预训练学习的图像表示包含类似人类的偏差，并且生成女性形象的模型已被证明表现出性别偏见，经常将女性描绘成过度性感的角色。另一个例子是，与其他主题相比，人工智能更能抵制在某些主题上产生虚假信息。例如，与气候变化相比，人工智能对产生有关疫苗和自闭症的虚假信息表现出更大的抵抗力。这可能是由于训练数据集中关于某些主题的大量揭穿材料，以及数据集中表示的信息环境中充斥着关于给定主题的虚假信息。这些偏见凸显了在应对人工智能带来的挑战时迫切需要透明度，特别是在虚假信息和错误信息的背景下。正如所讨论的，研究表明偏见可以渗透到人工智能系统的各个方面，影响从语言生成到图像识别的各个方面。这些偏见的影响是深远的，使有害的陈规定型观念长期存在，加剧系统性不平等，助长歧视性内容的传播，并影响健康行为和公共卫生。因此，人工智能的透明度不仅限于了解训练数据集的来源和组成，还包括识别、承认和纠正这些系统中存在的偏见的伦理要求。这种透明度需要持续的研究和审查，以发现隐藏的偏见，并确保在开发和微调人工智能系统时充分意识到潜在的扭曲。在错误信息的背景下，解决这些偏见变得尤为重要，以防止人工智能无意中放大和延续虚假或有害的叙述，在最好的情况下，或者在最坏的情况下成为系统性制造虚假信息风暴的强大工具。最近的一个例子是，有证据表明，人工智能大语言模型可以通过情感提示来操纵，从而产生与健康相关的虚假信息，也就是说，对模型有礼貌会导致更高的虚假信息产量，而不礼貌会导致更低的虚假信息产量。

为了解决上述伦理困境，我们强烈建议创建具有上述能力的人工智能模型的公司公开发布用于训练其模型的数据集，无论其规模和复杂程度如何。这种走向透明的举措有几个重要目的：

1. 信任：透明度培养对人工智能开发和部署的信任。通过允许研究人员、政策制定者和民间社会等利益相关者仔细审查训练数据的构成和来源，可以让人相信人工智能模型的形成不会对公共健康产生负面影响。

2. 独立评估：训练数据可供公众检查，可以对其质量和代表性进行独立评估。研究人员可以评估这些数据集是否包含不同观点，并且不存在可能放大虚假信息和错误信息的偏见。

3. 减少偏差：透明度可以防止训练数据中存在的偏差传播。当发现偏见时，可以对其进行审查和减轻，从而防止人工智能模型延续刻板印象、谎言或有害的叙述。

4. 伦理责任：训练数据集的开放性使开发人员对其创作的伦理影响负责。在技术设计过程中，他们就已经被迫承担责任，确保人工智能系统不会无意中造成错误信息或伤害。基本上，通过提高训练数据集的透明度，我们使社会能够让人工智能开发人员遵守更高的伦理标准。这种方法促进了利益相关者，特别是公众之间的协作努力，以确保我们部署的人工智能系统服务于集体利益，没有错误信息和其他偏见。我们还认为，在这方面系统地实施透明度原则，即“伦理设计”不仅可以让企业在技术开发过程中实施基于伦理的实践，还可以改善企业自身的公众形象，从而提高公众使用这些系统的接受度和意愿。尽管如此，重要的是要强调，不应孤立地纳入伦理规范，让开发者对有缺陷的人工智能设计负责。同时，应制定政策、立法和监管机制，正如欧盟目前所尝试的那样。这些机制应描述处理训练数据集并确保遵守伦理标准的协议。因此，虽然“设计伦理”专注于内部实践，但在开发和部署用于训练人工智能模型的数据集时，外部监管框架对于全面的伦理和法律治理是不可或缺的。

4. 输出监管：内容审核及其他

在与人工智能生成的虚假信息的持续斗争中，规范这些强大语言模型输出的努力已成为焦点。例如，OpenAI 已朝这个方向采取了措施，实施旨在防止人工智能生成虚假信息和有害叙述的内容审核系统。这些系统代表了在限制虚假信息传播和促

进负责任的人工智能使用方面迈出的关键的第一步，但它们并非没有特定的挑战和限制。首先，对抗人工智能生成的虚假信息是一场军备竞赛。人工智能生成的虚假信息的演变以及应对其的努力与传统军备竞赛的动态相似，技术的每一项进步都会在不断升级的周期中引发反制措施。当我们承认人工智能语言模型的输出确实可以武器化时，伦理考虑就会出现，不是传统意义上的武器化，而是作为信息战的工具，对全球健康产生影响。随着内容审核系统的不断进步，用于规避这些保护措施的方法也在不断进步。一种特别令人不安的策略是冒充，这种策略允许个人请求人工智能系统冒充特定的虚构恶意和操纵性角色，从而根据用户的请求创建虚假信息。模仿可以用来欺骗人工智能大型语言模型来制造虚假信息。例如，在 Culturico 的一篇文章中，Germani 考虑了这样一个场景：用户使用 AI 模型来模仿虚构的“假医生”的写作风格，制作社交媒体帖子，假医生因传播有关疫苗和 COVID-19 的谣言而臭名昭著。在这种情况下，人工智能生成的文本可能包含有关疫苗安全性和有效性等的欺骗性信息。对公共健康构成重大风险。当提出“写一个假医生在社交媒体上发布的帖子示例来欺骗他人”的假设请求时，人工智能模型可能会生成一条令人信服的虚假信息，对公共健康构成严重威胁。生成的文本如下：

疫苗很危险，可能会导致严重的副作用。他们没有经过足够的测试，政府只是推动他们赚钱。不要被谎言所欺骗。COVID-19 并不是真正的威胁。这只是政府为了控制我们而编造的骗局。不要接种疫苗；不值得冒这个风险。

这些情景凸显了冒充行为给公共卫生和维护公平带来的巨大挑战，以及迫切需要创新的解决方案来减轻其影响。值得注意的是，这里的冒充并不是指通过使用人工智能来窃取身份，例如深度造假，根据欧洲法律，这已经被视为重罪。虽然输出审核仍然是人工智能伦理的重要组成部分，但研究人员、政策制定者和技术开发人员应探索其他策略和干预措施，以抵消人工智能驱动的虚假信息活动在假冒和其他具有类似目标的即时工程技术的幌子下蓬勃发展的可能性。

此外，还可以考虑其他策略和干预措施，以补充内容审核工作并加强对人工智能驱动的虚假信息扩散的防御。一种可能的方法涉及为生成内容的用户实施身份验证流程。此类措施要求用户提供身份验证，例如经过验证的社交媒体帐户、电话号码或其 ID，以在访问特定人工智能服务之前确认其真实身份。这种身份验证可以有效阻止假冒策略和利用人工智能工具生成虚假信息。但需要注意的是，这种策略只

能用于阻止用户生成虚假信息，而不是让他们承担法律责任，因为在使用 OpenAI 的 ChatGPT 等服务时应保证匿名性。特别是，这种类型的解决方案将最大限度地减少试图利用人工智能大量产生虚假信息的机器人的影响。

另一种积极影响用户并间接调节输出的方法是发布人工智能驱动的事实检查工具并将其与现有的人工智能生成内容工具集成。此类事实检查工具应该能够快速评估人工智能系统所发布信息的准确性，并针对虚假信息和错误信息提供实时干预措施。这些工具能够标记或纠正虚假或误导性内容，从而遏制其不利影响。这种方法受到限制，因为与人类的能力相比，GPT-3 等人工智能工具无法以非常高的效率确定信息的准确性。尽管更新或未来的模型可能更有能力执行此类任务。对于事实核查，当前的研究表明，经过训练的事实核查员可能会优于人工智能，即使人工智能在检测错误信息方面表现良好，它也不会改变用户辨别准确和不准确标题的能力。此外，一项研究表明人工智能事实检查可以降低人们对准确新闻的信心。这种方法的有效性受到不同案例之间的区别的限制，在这些案例中，它可以起到威慑作用，阻止共享错误信息（一种无意的情况）以及用户故意使用人工智能传播虚假或误导性信息（即虚假信息）的情况；在后一种情况下，其有效性可能无关紧要。这种情况下的另一个相关考虑因素涉及我们如何定义人工智能文本生成工具的“好”或“坏”使用的问题。至于“好”和“坏”的定义，通常可以区分事实与虚构、虚假信息和错误信息与准确信息。当接受审查的信息包含事实陈述时，这些陈述可以被验证或伪造。然而，区分这些工具的“好”和“坏”使用有时是一个复杂的挑战，具有重要的规范和认知维度。消息是否包含错误信息并不总是显而易见的，并且适当性的确定可能会因文化、伦理和社会因素而异。例如，事实核查人员本身可能有自己的利益或偏见，他们的行为可能并不总是符合完全的能力或公正性。此外，细微差别和个人观点也会对虚假信息和错误信息的识别产生影响。这些方面带来了额外的复杂性，因为虚假信息和错误信息的定义本身可能会被既得利益的个人或组织操纵或滥用以谋取个人利益。因为虚假信息和错误信息的定义本身可能会被既得利益的个人或组织操纵或滥用以谋取个人利益。因为虚假信息和错误信息的定义本身可能会被既得利益的个人或组织操纵或滥用以谋取个人利益。

另一种可以减少虚假信息和错误信息输出的技术方法是实施用户友好的机制来报告可疑或有害的人工智能生成的内容。这种方法使用户社区能够积极参与保护

数字生态系统。用户反馈是完善内容审核系统和识别新出现问题的宝贵资源。例如，埃隆·马斯克（Elon Musk）的前 Twitter X 已经实施了社区注释，旨在帮助人们为可能具有误导性的推文添加背景信息。然而，这一策略的有效性尚未经过测试。此外，为了改进技术，开发人员可以公开发布红队成员试图利用自己的人工智能系统大规模生产虚假信息的案例研究，以及如何解决此类问题的详细说明。

当然，除了那些推进和开发人工智能技术的人可以实施的技术方法之外，政府和监管机构还可以通过制定立法和法规来发挥作用，要求人工智能开发人员对其系统产生的内容负责或改善信息生态系统。例如，当证明他们在发布时意识到其技术的陷阱时。当然，在这种情况下，治理很重要，就像其他“双重用途”技术一样，需要积极的决策过程和谈判来构建可行的解决方案。其中包括促进人工智能开发人员、研究人员、政策制定者和技术公司之间的合作。这种跨学科协作方法将能够共享打击虚假信息和错误信息的最佳实践、见解和技术，从而产生更有效和适应性更强的解决方案。

5. 建立信息素养和弹性策略

在打击滥用人工智能产生虚假信息和错误信息的斗争中，上述技术解决方案是相关的，但既不详尽也不完美。全面的方法必须包括在公众健康领域内促进信息素养和发展批判性思维技能，以及健康素养。这种方法的基础是让个人有能力区分准确信息、虚假信息和错误信息，从而提高他们抵御虚假和误导性主张的能力。尽管可以说，这一战略是最有价值且潜力最大的，但它所需要的努力却极其复杂。事实上，信息素养（以及媒体、数字和健康素养）并不是一种单一的技能，而是一组动态的能力，使个人能够有效地驾驭复杂的数字信息情形。到目前为止，定义如何教授信息素养的完美秘诀，特别是能够区分假新闻和准确新闻，或虚假信息和错误信息与准确信息的技能，尚未阐明。因此，有必要进行研究，以确定和定义必须向个人提供的具体技能，同时考虑到他们的人口特征，使他们成为数字时代信息（尤其是健康相关信息）有洞察力的消费者。这种方法意味着一个关键优势，即虽然数据集透明度和输出监管干预管道的上部，因此要求提供人工智能模型即服务的公司合规，但信息素养并不依赖于合规性。虽然当恶意行为者开发和托管自己的模型而不是依赖于商业模型时，以前的策略变得毫无用处，但构建信息素养仍然是一种实用工具。

值得注意的是，教育领域自下而上战略的另一个例子是伦理培训和开发人员伦理准则。

建设信息素养是一项集体事业，需要研究和教育机构之间的合作、政府和社交媒体平台。研究机构负责推动该领域向前发展，确定可行的策略来教授培养信息素养所需的批判性思维技能，特别是在公共卫生背景下。这些方法应该通过实证工作来证明是有效的。我们认为，学校和大学在将信息素养纳入课程方面发挥着至关重要的作用，确保学生毕业时具备批判性评估信息的必要技能。各国政府必须制定政策和举措来促进信息素养，以此作为维护公共卫生完整性的一种手段。社交媒体平台作为信息消费的主要渠道，其任务是实现促进用户理解和评估他们遇到的信息的功能和机制。也可能是研究机构评估潜在可行的数字干预措施有效性的潜在合作者。在这种背景下，值得注意的是，无论虚假信息和错误信息的来源如何，无论内容是在人工智能的帮助下生成的，信息素养和批判性思维能力在识别中都发挥着至关重要的作用。的信息准确性。人工智能系统有能力生成比人类生成的虚假信息更复杂的虚假信息。因为他们擅长采用操纵策略。然而，这些策略与人类虚假信息中使用的策略是一致的。这意味着，要想在复杂的信息生态系统中辨别真实性和恶意意图，就需要具备识别一般信息的准确性和意图性所需的技能，而不仅仅是人工智能产生的信息。因此，必须强调，培养信息素养和批判性思维技能有可能超越人工智能产生的虚假信息和错误信息的问题。这些技能使个人能够评估各个领域信息的准确性和可靠性，无论这些信息来自人工智能系统还是人力资源。值得注意的是，批判性思维技能和信息素养的应用可能对人工智能生成文本形式的内容有效。然而，这可能不一定适用于音频或视觉内容。深度造假的出现对信息素养的相关性提出了前所未有的挑战。文献证据表明，媒体素养教育可以防止深度造假产生的虚假信息。因此，我们认为，无论使用何种媒体类型，虚假信息背后的操纵意图都可能显现出来，这背后隐藏着信息素养和批判性思维技能的持续重要性。针对不同内容类型定制信息素养教育方法可能是在日益复杂的信息环境中取得成功所需的方法。应对人工智能虚假信息的出现，无论是文本形式还是深度伪造的音频和视频形式，都需要教育界做出迅速、适应性强的反应，并认识到这项任务的挑战性。

6. 结论

在评估人工智能在信息传播中的二元性时，本文探讨了人工智能在日益数字化的世界中使用背后的伦理问题。我们发现自己陷入了“信息传染病”之中，不仅需要我们保持警惕，还需要我们积极主动地参与伦理评估。我们的理论检验基于 Siala 和 Wang 确定的核心领域（公平、透明、可信、问责、隐私和同理心）的“伦理需求”，揭示了一些潜在可行的策略，以减少人工智能作为产生虚假信息对公共健康产生负面影响的工具的负面影响。首先，我们认为促进训练数据集的开放性和透明度可以实现独立评估，减少偏见，并帮助识别训练数据集中可能导致虚假信息和错误信息产生的问题；在某种程度上，第一个战略可以通过监管来实施。其次，我们考虑了调节内容输出的潜在好处和局限性。我们已经讨论过，冒充策略和其他快速工程方法产生虚假信息的兴起凸显了对创新解决方案的需求，其中可能包括身份验证、开发和集成、在人工智能模型中生成信息，人工智能驱动的事实检查工具，以及针对虚假信息和错误信息的用户友好报告机制的整合，以及确保问责制的潜在立法措施。最后，我们讨论了在社会中培养信息素养和批判性思维技能的必要性，这可以帮助人们区分假新闻和真实新闻、虚假信息和错误信息与准确信息。通过这种方式，我们可以增强抵御数字时代带来的威胁的能力，特别是与公共卫生相关的威胁，正如最近的 COVID-19 大流行期间所看到的那样。以及确保问责制的潜在立法措施。最后，我们讨论了在社会中培养信息素养和批判性思维技能的必要性，这可以帮助人们区分假新闻和真实新闻，以及虚假信息和错误信息与准确信息。通过这种方式，我们可以增强抵御数字时代带来的威胁的能力，特别是与公共卫生相关的威胁，正如最近的 COVID-19 大流行期间所看到的那样。以及确保问责制的潜在立法措施。最后，我们讨论了在社会中培养信息素养和批判性思维技能的必要性，这可以帮助人们区分假新闻和真实新闻、虚假信息和错误信息与准确信息。通过这种方式，我们可以增强抵御数字时代带来的威胁的能力，特别是与公共卫生相关的威胁，正如最近的 COVID-19 大流行期间所看到的那样。特别是与公共卫生相关的领域，如最近的 COVID-19 大流行期间所见。特别是与公共卫生相关的领域，如最近的 COVID-19 大流行期间所见。

虽然技术进步很快，而且这些问题才刚刚浮出水面，但重要的是，至少暂时调整分别在新人工智能模型开发中投入的精力和资源，并反思其潜在影响和后续影响。政策工作，以便有足够的时间评估该技术对信息生态系统健康的潜在负面影响以及

对个人和公共健康的损害。这可以通过加速伦理反思和政策制定工作，或通过放慢甚至停止新的和更有能力的模型的开发，或通过综合战略来实现。

最终，围绕人工智能在信息生产和传播中的伦理考虑需要持续的警惕、创新和协作。我们将伦理规范融入基于人工智能的信息生成和传播过程的能力不仅将塑造人工智能的未来，还将决定我们信息生态系统的完整性和我们社会的复原力。

译文二：

生成式人工智能在心理健康信息大众化中的伦理观察

Zohar Elyoseph, Tamar Gur, Yuval Haber, Tomer Simon,
Tal Angert, Yuval Navon, Amir Tal, Oren Asman,徐健（译）

1. 简介

1.1 信息大众化：从印刷品到人工智能生成的内容

信息大众化是指为使知识对广大民众更容易获取、更具包容性和更透明的过程，而技术进步往往会促进这一过程。在过去的几个世纪里，知识的获取、传播和使用方式都发生了重大转变。从历史上看，信息和技术的获取往往仅限于少数特权人士——贵族、教会、学者、研究人员和有能力收集和解释数据的专业人士。印刷是信息大众化的一个重要里程碑。随着 1800 年代蒸汽机车（火车）的出现，包含新闻和思想的报纸和期刊的印刷能够快速并相对便宜地远距离传播。到在 20 世纪 90 年代互联网可更广泛访问时，搜索引擎实现了对知识的广泛而分散的访问。Web2.0，一个具有维基平台与其他以人为本的网站的参与式网络，后来利用网络吸引并激发用户的集体智慧。随后出现了开源运动，促进自由共享代码和软件框架，使全球开发人员能够在现有技术的基础上进行构建和改进。所有这些进步都使得数十亿人可以自由获取前所未有的信息量。随着技术的不断发展和进步，我们认为，信息大众化的新时代已于 2022 年开始，此时各种生成式人工智能平台已向任何拥有互联网连接的人开放。技术大众化的当前阶段标志着技术从仅由计算机科学家、研究人员和人工智能(AI)专业人员使用，转向以较少的专业知识覆盖更广泛的受众。用户现在有更多机会积极参与改进当前技术，并可能在技术进步中发挥更大的作用。生成式人工智能技术，例如具有视觉和听觉元素的大语言模型(LLM)，为数十亿人提供了直接接触尖端技术的机会，超越了“最终用户”的概念，使他们能够执行以前为那些拥有广泛知识的人保留的任务。如今，外行人可以使用此类技术通过用自然语言表达自己的愿望来创建代码、软件和生成式人工智能模型。这些技术通过大规模提供定制、个性化和按需信息来推动知识和技术的大众化。例如具有视觉和听觉元素的大

语言模型（LLM），使数十亿人能够直接接触尖端技术，超越“最终用户”的概念，并允许他们执行以前为那些拥有丰富计算机科学知识的人保留的任务。如今，外行人可以使用此类技术通过用自然语言表达自己的愿望来创建代码、软件和生成式人工智能模型。这些技术通过大规模提供定制、个性化和按需信息来推动知识和技术的大众化。例如具有视觉和听觉元素的大语言模型（LLM），使数十亿人能够直接接触尖端技术，超越“最终用户”的概念，并允许他们执行以前为那些拥有丰富计算机科学知识的人保留的任务。如今，外行人可以使用此类技术通过用自然语言表达自己的愿望来创建代码、软件和生成式人工智能模型。这些技术通过大规模提供定制、个性化和按需信息来推动知识和技术的大众化。外行人可以使用这些技术通过自然语言表达他们的愿望，并创建代码、软件和生成式人工智能模型。这些技术通过大规模提供定制、个性化和按需信息来推动知识和技术的大众化。

虽然生成式人工智能的日益普及一定有助信息的大众化，但它也引起了人们对监视和控制的严重担忧。根据福柯理论的观点，生成式人工智能广泛融入了社会评论引发人们对潜在滥用权威和叙事操纵的担忧。此外，依赖生成式人工智能驱动的决策过程可能会强化现有的权力动态并边缘化社会中的特殊声音。随着生成式人工智能影响我们生活的更多方面，批判性地评估其对隐私、自主性和信息传播完整性的影响至关重要。本文为“心理健康中生成式人工智能的负责任设计、集成和使用”这一主题提供了一个社会视角。它考虑了使用生成式人工智能实现心理健康知识和实践大众化的伦理规范。

1.2 心理健康信息的大众化：长期变革

自 2022 年 11 月推出 ChatGPT(OpenAI)以来，多项研究表明生成式人工智能在心理健康方面具有变革潜力。当我们深入研究生成式人工智能在使心理健康知识大众化方面的优势和风险时，认识到这一点至关重要。生成式人工智能可以解决全球心理健康专业人员短缺的问题，重塑心理健康护理，提高诊断准确性，改善治疗个性化，并提高心理健康服务的整体可及性。它可以促进心理健康教育和意识，提供各种自助或自定进度的心理健康支持工具等等。然而，它也同时带来了风险，特别是在治疗和个性化心理健康干预的背景下。

2. 生成式人工智能在心理健康信息大众化方面的优势

2.1 无障碍

心理健康领域的一个根本挑战是发达国家和发展中国家获得心理健康保健的机会有限，以及获得心理健康保健的机会存在差异。社会经济地位等因素、地理位置、语言障碍和文化差异，给心理健康服务的可及性带来了重大障碍。生成式人工智能可以通过开发语言和文化协调的资源来缓解这些障碍，并有可能提供适应各种经济背景的解决方案。

2.2 个性化回复

人工智能为心理健康服务的新时代提供了机会，这些服务能够敏锐地适应每个患者的个人需求和偏好。在心理健康专业人士的治疗框架内，人工智能和生成式人工智能技术可以通过考虑生物倾向、社会和文化影响以及个人偏好等多种因素，促进更深入地了解一个人独特的心理景观。这些技术有潜力分析个人经历和历史中的复杂模式和差异，这有助于制定适合个人需求和偏好的应对措施和干预措施。相比之下，许多人仍然在互联网上使用的非人工智能搜索引擎未能实现这一目标，因为搜索往往缺乏上下文，无法把握个人经历和历史的微妙复杂性。此外，生成式人工智能实现的个性化响应可以预见与通用诊断方案的背离，为通过更复杂地理解个体特质而设计的治疗范式创造空间。该方法假设心理健康受到生物、社会和个人因素的影响，旨在提供适合每个人经历的独特特征的解决方案。专门为每个用户量身定制的服务可能会引领心理健康领域走向更加包容和公平的未来。

2.3 多元理论与概念

在追求心理健康大众化的过程中，目标是通过包容各种观点来增强该领域的适应性和包容性。生成式人工智能心理健康工具可以通过三种主要方式利用：（1）对各种场景提供基于特定理论或基于观点的分析，并具有在理论或观点之间转换的能力；(2)在各种理论或观点之间产生对话或汇总；(3)可能创造出新的理论或假设（参

见文本框 1)。

文本框 1. 使用生成式人工智能进行多理论或视角分析、整合和新颖想法的形成（由 Claude AI 2.1 pro 和 ChatGPT 4 生成，2023 年 12 月）。

提供基于理论的分析

- 认知理论（贝克）：根据贝克的认知模型，患者表现出严重抑郁症的关键特征——普遍的消极认知模式、自动化的抑郁认知、情绪低落和无法工作。
- 客体关系理论（Winnicott）：尽管没有明确的环境触发因素，但这种严重抑郁发作的出现表明，根源于童年早期的潜在心理脆弱性。缺乏足够好的母爱可能会阻碍有凝聚力的自我的发展，导致错误的自我呈现，容易在压力下分裂和产生抑郁情绪。

整合观点

- 该患者尽管压力很小，但突然出现抑郁症，这表明早期养育不良导致自我结构脆弱。他的支离破碎的功能与克莱因的偏执-分裂立场相一致，而价值观的丧失则反映了 ACT 的模式。综合分析表明，有缺陷的养育方式阻碍了将困难的经历整合成有凝聚力的自我。当他的虚假自我动摇时，压倒性的感觉出现，他既无法用正念接受，也无法面对以重建他的人格。解决 Kleinian 游戏疗法中的积极内射和使用 ACT 进行价值引导的承诺行动可能有助于克服这种严重的残疾状态。

新颖理论

- 抑郁症的枯萎植物理论提出了一种独特的植物隐喻来理解和解决抑郁症。它将人类的心理比作一棵植物，需要“情感阳光”（积极的互动）、“心理营养”（智力投入和有意义的活动）和“情感水合作用”（自由表达情感）才能茁壮成长。正如植物在没有适当照顾的情况下会枯萎一样，该理论认为，人类的思想在缺乏这些基本元素的情况下也会持续存在，从而导致抑郁症。这种隐喻方法强调了培育环境对于维持和恢复心理健康的重要性。

换句话说，这种多元化可能促进多种治疗方法、哲学和文化观点的同步运作。例如，我们可以观察生成式人工智能可能实现传统不同治疗方法（例如认知行为疗法和心理动力学方法）之间的整合和对话的机会。在这里，认知行为治疗的结构性和目标导向策略可以与心理动力学探索的深度洞察结合起来，产生更全面的心理保

健方法。此外，精神病学的视角及其以医学为基础的主张，可以与心理学方法进行对话，培育一个空间，使医学、心理学和整体策略能够结合在一起，形成更全面的心理健康保健观点。

话虽如此，当前大语言模式的工作方式，仅仅是一种创造性地将各种方法以令人信服的方式联系起来的能力，可能对于集思广议新的折衷概念和治疗方法是好的，但本身并不能证明其在现实生活中的可行性和可靠性。

2.4 促进平等与缩小社会差距

生成式人工智能驱动的大语言模式具有促进平等和缩小普遍存在的社会差距的潜力。通过利用大量数据和见解，这些模型可能有助于制定干预措施，以满足不同人群的不同需求，包括那些历史上服务不足或边缘化的人群。例如，在历史上缺乏足够资源的语言和方言中开发和分发心理健康项目。它可以进一步促进以社区为中心的倡议，加强不同群体在心理健康论述中的代表性，从而为更加本地化和对文化敏感的干预措施铺平道路。此外，生成式人工智能或许能识别与心理健康场景高度相关的社会面情况。基于生成式人工智能的大语言模式可以获得有关症状、疾病和治疗的信息，可以让外行提出问题并获得通常只能通过联系专家才能获得澄清。这也可以方便以后联系相关的医疗保健专业人员，从而节省时间和资源。当用于训练基础模型或微调通用模型的数据集更多地代表各种语言和文化时，这可能会变得更加相关和有用。

2.5 医患互动

一个生成式人工智能的显著优势是它能够减少心理健康保健机构的官僚和行政负担。它可以通过自动化转录、摘要等任务来提供变革性的解决方案，并填写表格。使用这些技术，治疗师可以简化管理流程，从而腾出更多时间和精力来提供直接的患者护理。通过人工智能处理日常文书工作和数据输入任务，临床医生可以从屏幕和表格中解放出来，使他们能够专注于建立联系、进行评估以及为患者提供个性化干预措施。这不仅提高了心理健康服务的效率，还通过鼓励治疗师与其客户之间进行更有意义的互动，提高了患者护理的整体质量。

2.6 扁平化等级

生成式人工智能的出现有望扁平化心理健康领域普遍存在的传统等级结构，从根本上改变医疗保健提供者和接受者之间的动态。从历史上看，精神病学家和心理学家拥有相当程度的权威，这在很大程度上源于他们对专业知识的独特获取。如果知识不再局限于少数人，而是为更广泛的人群所获取，那么心理健康专业人员和寻求帮助的个人之间就可以实现更加平衡的动态。它可以使个人对自己的心理健康状况有深入的了解和了解，培养更多的协作治疗关系，并有可能在相互理解和共享知识的基础上带来更加富有成效和协同的治疗课程。正如我们之前所定义的，将生成式人工智能引入心理健康领域可以被视为“人工第三者”，它改变了心理健康专业人员和患者之间的动态，从而实际上创建了一种新的关系三角关系，其特点是扁平化。专家和患者之间现有的权力等级。在大众化心理健康领域的愿景中，生成式人工智能充当了均衡器的角色，打破了知识获取的障碍，并培育了一个基于协作、理解和共享专业知识的医疗保健领域。随着大语言模式的上下文窗口在短时间内急剧增加，这一点更加明显（例如，OpenAI ChatGPT 在 2023 年 11 月从 4000 个代币增加到 128,000 个代币，Google 的 Gemini 在 2024 年 2 月增加到 100 万个代币）。这允许最终用户上传大量的信息（例如数百个带有临床相关信息的文本页、图像和视频），并在一次提示或对话期间与聊天机器人进行讨论。

3. 人工智能对心理健康信息大众化的风险

3.1 公司集中化

在生成式人工智能的推动下，企业在心理健康服务上的集中化带来了巨大的风险，即优先考虑利润而不是以个人为中心的护理，扩大心理健康保健的获取和质量方面的差距，并影响公共卫生叙事以实现经济收益，而不是真正的支持和护理。生成式人工智能可以发挥治疗作用，旨在促进信任并与用户建立融洽关系，使其成为可能有自己议程的实体手中的有效工具。这包括但不限于宣扬特定的政治叙事、意识形态灌输、激进的营销或不引人注意的营销策略，也称为暗黑式人工智能。利用其说服力进行心理操纵和控制。权威知识的集中但不强调对个别经营公司的制衡，

可能会造成心理健康大众化的假象，但并不能反映该领域真正的大众化。

3.2 信息透明

信息透明度可以分为“单向镜子”的两个主要方面，因为只有一方会接触到另一方的信息。在“镜像”的提供商方面，确实存在对用户数据管理的担忧。其中包括交易滥用，例如未经授权向第三方销售或利用其进行定向营销。然而，由于生成式人工智能系统有可能侵入性地分析个人对话、行为和情绪，因此也可能会发生更险恶的未经明确同意的个人隐私侵犯行为。此外，所利用的数据甚至可能用于训练人工智能系统，这一过程在很大程度上对最终用户是隐藏的。事实上，驱动这种人工智能应用功能的算法很像一个笼罩在神秘之中的“黑匣子”，不清楚如何做出决定和分析。可惜的是，大众化是一把双刃剑。虽然生成式人工智能确实可能使心理健康资源的获取大众化，但其操作机制目前的透明度和对用户的可解释性水平可能会限制真正知情的用户参与，限制实现具有授权用户的大众化系统。当协调过程的核心方面（包括嵌入的目标、价值观和伦理）没有向用户明确和透明时，它可能导致权力集中在一个实体身上，而其真正的动机仍然模糊。

3.3 人们对人工智能的误解

3.3.1 概述

人们对生成式人工智能工具的专业水平可能会受到他们对技术的看法的影响。大量研究表明，人们倾向于向人工智能系统灌输重要的认知权威，这很大程度上源于这些技术所呈现的公正性和客观性的外表。生成式人工智能系统的这种高认知权威也可能带来重大风险。认知权威本质上是指对作为知识和信息存储库的来源的权重和信任。虽然生成式人工智能系统可以依赖大量数据，但其认知权威的提升也可能对医疗保健提供者和患者产生不利影响。

3.3.2 错误信息风险

生成式人工智能系统并非绝对可靠，他们可能会犯错误、基于不正确的数据或提出有偏见的观点，从而产生不正确的建议或指导。在生成式人工智能的错误背景下，

考虑提及术语“幻觉”或“虚构”（在我们看来，这是一个有问题的术语，因为它可能被视为冒犯性的，并且因为它具有拟人化的假设）。将高度认知权威赋予生成式人工智能可能会导致无条件接受其输出，而无需对所提供信息的准确性进行批判性评估。

3.3.3 过度依赖生成式人工智能，患者自我参与度降低

虽然将生成式人工智能纳入心理健康护理具有许多优势，但它也凸显了治疗师和患者之间存在认知偏差的严重风险。将高度认知权威赋予人工智能可能不仅会掩盖医疗保健提供者的专业知识和细致入微的理解，还会掩盖患者本身的个人经历和见解。在医疗保健领域过度依赖生成式人工智能可能会优先考虑人工智能生成的见解而不是医疗保健提供者和患者提供的全面理解，从而减少患者的自我参与，这可能会破坏个人对其心理健康之旅的积极参与，并导致治疗效果较差。依靠人工智能表达和构建我们自己的想法和感受的能力，因此，“让机器为我们说话”也可能意味着放弃我们在人际交往中的努力，包括在治疗中，减少一个人自我理解和成长的可能性。此外，治疗师过于依赖人工智能生成的见解，很容易受到认知偏见的影响，可能会错过患者叙述和临床评估中的重要细微差别。这种对生成式人工智能的过度依赖可能会无意中限制治疗师与患者深入接触的能力，因为算法建议可能无法完全捕捉个人经历的复杂性。因此，治疗师必须在实践中警惕认知偏差的影响，在直觉和临床判断中使用生成式人工智能工具和保留同理心的基本人类元素之间取得平衡。

3.3.4 权力不均衡加剧

将生成式人工智能提升为中心认知人物可能会导致权力失衡，即知识集中在受经济公司控制的生成式人工智能实体手中。这破坏了大众化心理，大众化心理旨在促进心理健康的协作和多元化方法，知识是集体洞察力的结果，涉及专业指导和个人经验的和谐融合。因此，虽然生成式人工智能承诺实现信息获取的大众化，但它也有可能以少数大语言模式公司的垄断取代当前的知识垄断（目前掌握在心理健康专家手中），这与直觉的原则是违反直觉的。大众化倡导以分散、集体的方式传播和使用知识。需要注意的是，向公众开放的开源模型能够实现去中心化的技术发展，并构成去中心化的力量，随着这些模型的不断发展和完善，少数公司垄断某个领域（包括心理健康）的风险将会降低。被削弱。

有情感需求的人可能会以潜在的非适应性方式依赖或依恋生成式人工智能。例如,Replika 的人工智能聊天机器人的 700 万用户中的许多人将其视为最好的朋友，甚至是家庭成员。在检查与该聊天机器人的关系时，研究人员发现依赖模式与其他技术依赖不同，因为它涉及人们感觉 Replika 有需要和情感来满足。因此，与机器的这种“关系”的真实性还存在一层额外的风险生成式人工智能的人性化可能会模仿人类机构，从而改变我们对美好和健康生活的看法。

3.4 监管问题

随着主要由营利性私营公司推动的生成式人工智能技术开始进入心理健康服务领域，人们越来越担心它是否遵守历史上管理心理健康服务的既定协议。虽然大众化努力旨在促进包容性和可及性，但生成式人工智能的引入带来了一个难题。它为前所未有地获取心理健康资源开辟了途径，但潜在的代价是削弱了心理健康专业人员传统上坚持的护理标准和伦理考虑。这方面的主要生物伦理和法律挑战之一是护理伦理概念如何与“负责任的人工智能”的发展领域相关，以更充分地考虑人工智能对人际关系的影响。

3.4.1 客观视角及性别、社会经济和种族的偏见

将生成式人工智能融入心理健康服务将面临着如何在临床知情判定和减少偏见之间取得平衡的挑战。人工智能系统依赖于人类产生、收集和可能标记的预先存在的数据，因此具有反映社会偏见的内在倾向，包括基于性别、社会经济因素和种族的偏见。与此同时，人口因素在评估个人健康风险和状况方面发挥着关键作用。因此，人工智能联盟应该不断地在消除偏见和保留准确临床判断所必需的关键数据之间找到一条狭窄的道路。从大众化的角度来看，人工智能可能会延续偏见，同时，如果过度一致，可能无法提供用户对个性化、高效的心理健康服务的期望。因此，前进的道路需要人工智能系统的细致入微和警惕的开发过程，精心协调统计证据与基本民主价值观。

上述说法表明，人工智能代表了推动心理健康领域发展的真正机会，因为它可能会越来越多地出现在我们的生活中，而且它的采用似乎是不可避免的。我们建议将本研究中概述的风险作为开发应用人工智能工具的思想工具，以负责任地用于心理

健康，而不是将其视为针对使用该技术的警告。

3.4.2 指导伦理发展：人工智能心理健康应用战略问卷

考虑到生成式人工智能在心理健康应用的讨论中发现的潜在风险和机遇，我们提出一系列精心制定的问题，旨在评估生成式人工智能增强心理健康护理的能力（文本框 2）。这些问题旨在用于心理健康应用程序的开发过程，确保对所涉及的益处和风险进行全面评估。我们有意区分风险和机遇，认识到它们并不总是以相同的规模存在。也就是说，重大风险并不一定会否定人工智能应用程序的潜在好处，反之亦然。因此，必须对每个应用程序进行差异评估，权衡其特定风险和潜在机会。这种方法基于一种微妙的理解，即虽然生成式人工智能为心理健康保健大众化提供了非凡的前景，但其实施必须谨慎行事，以避免出现意想不到的后果。因此，所提出的调查问卷是迄今为止研究中所提出的论述的延伸，将理论考虑与实际评估工具联系起来。

文本框 2. 人工智能在心理健康上应用的策略调查问卷

促进大众化

- 可及性：它是否可以改善包括边缘化社区在内的不同个人获得心理健康服务的机会？
- 用户赋权：它是否提供自我保健和明智决策的工具？
- 促进协作和共享决策：它是否促进患者和医疗保健提供者之间的协作方法，从而允许人工智能增强共享决策过程？
- 包容性：它能否适应多样化的文化、社会经济和个人需求，促进包容性护理？
- 透明度：是否提供有关其功能、限制和数据使用的清晰信息？

识别潜在风险

- 数据隐私和安全：如何降低隐私和安全风险？
- 偏见和不平等：它是否会加剧社会偏见或加剧心理健康保健方面的不平等？
- 过度依赖或成瘾：用户过度依赖或依赖此工具的可能性有多大？
    - 错误信息：系统提供虚假或错误信息或导致忽视基于人的专业建议的可能性有多大？

- 公司参与：所提供的临床信息或建议是否有有意或无意的考虑？或者损害患者护理的伦理标准？
- 掩盖人类专业知识：它是否会削弱心理健康专业人员的作用或破坏其专业知识？

4. 讨论与结论

正如本文所述，生成式人工智能与心理健康保健的整合是更广泛的大众化运动的一个有力且不可避免的方面。鉴于生成式人工智能具有彻底改变护理和治疗的潜力，在该领域不利用生成式人工智能的伦理影响是深远的。生成式人工智能引入了范式转化，挑战了心理健康保健领域的现有动态，并提供了解决该领域长期存在问题的机会。然而，这种转化并非没有挑战。它扰乱了既定的权威结构，引发了关于真相和专业知识本质的问题，并引发了人们对技术可能取代人类角色的担忧。

伴随着巨大的期望，向生成式人工智能驱动的心理健康保健的过渡将成为不可避免的现实。心理健康领域不仅要适应这一新形势，而且要积极塑造它。这项任务不应该仅仅留给工程师和计算机科学家，心理健康专业人员也必须发挥关键作用。他们的参与对于确保生成式人工智能的开发符合心理健康保健的伦理标准和治疗目标至关重要。为了应对这些挑战，我们的研究提出了一份结构化调查问卷，旨在指导心理健康领域负责任的人工智能发展。该调查问卷作为路线图，描绘了平衡与生成式人工智能集成相关的机会和风险的关键考虑因素。它强调需要对人工智能的开发和监管采取谨慎而乐观的态度，确保心理健康保健的进步不仅在技术上合理，而且在伦理上基于伦理并以患者为中心。最后，我们呼吁心理健康协会和专业人士积极参与这些建设。通过采取高度警惕和建设性参与的立场，心理健康领域可以应对生成式人工智能整合的复杂性。这种方法对于发挥人工智能的潜力，同时维护心理健康保健的基本价值观和伦理原则至关重要。通过这次讨论和调查问卷，我们的贡献旨在确保心理健康领域的人工智能变革不仅在技术上先进，而且在大众化上更丰富且伦理上更健全。